PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2018

The Pacific Symposium on Biocomputing (PSB) 2018 is an international, multidisciplinary conference for the presentation and discussion of current research in the theory and application of computational methods in problems of biological significance. Presentations are rigorously peer reviewed and are published in an archival proceedings volume. PSB 2018 will be held on January 3 – 7, 2018 in Kohala Coast, Hawaii. Tutorials and workshops will be offered prior to the start of the conference.

PSB 2018 will bring together top researchers from the US, the Asian Pacific nations, and around the world to exchange research results and address open issues in all aspects of computational biology. It is a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, and other computational methods, as applied to biological problems, with emphasis on applications in data-rich areas of molecular biology.

The PSB has been designed to be responsive to the need for critical mass in sub-disciplines within biocomputing. For that reason, it is the only meeting whose sessions are defined dynamically each year in response to specific proposals. PSB sessions are organized by leaders of research in biocomputing's "hot topics." In this way, the meeting provides an early forum for serious examination of emerging methods and approaches in this rapidly changing field.

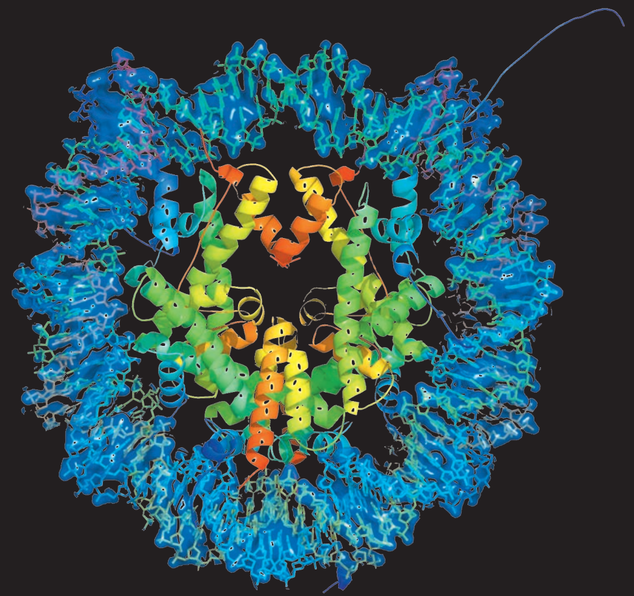
World Scientific
www.worldscientific.com
10864 eb

ISBN 978-981-3235-53-3 (eb)

R. B. Altman
A. K. Dunker
L. Hunter
M. D. Ritchie
T. Murray
T. E. Klein

PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2018

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018



Edited by

Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Marylyn D. Ritchie, Tiffany Murray & Teri E. Klein

Cover image:
This image depicts a molecular model of the Nucleosome (PDB ID: 1aoi, Luger et al. (1997) Nature 389, 251–260) — The nucleosome is the organising principle behind higher ordered chromatin structure. The histone core of the nucleosome exemplifies the many molecular mechanisms that have evolved to regulate access to the DNA in chromatin.

Image by D. Rey Banatao,
Pacific Symposium on Biocomputing.

DEMOCRATIZING HEALTH DATA FOR TRANSLATIONAL RESEARCH

IMAGING GENOMICS

PRECISION MEDICINE: FROM DIPLOTYPES TO DISPARITIES TOWARDS

IMPROVED HEALTH AND THERAPIES

READING BETWEEN THE GENES: COMPUTATIONAL MODELS TO DISCOVER FUNCTION FROM NONCODING DNA

ADVANCES IN TEXT MINING AND VISUALIZATION FOR PRECISION MEDICINE

WORKSHOPS

ERRATUM

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018

2018 marks the 23rd Pacific Symposium on Biocomputing (PSB). The original founders of PSB (Hunter & Klein) chose the term "Biocomputing" in 1996 in order to be as broad as possible—a wise decision as bioinformatics, clinical informatics, medical informatics, biomedical informatics, biomedical data science, and other terms (computational intelligence, artificial intelligence, data mining, machine learning, for example) have also gone in and out of style. Of course, each of these has it own connotation and there is no problem with a proliferation of useful and descriptive words. The anthropologist Franz Boas studied the inuit eskimos and commented on the large number of words they had for snow (New Scientist, 1/14/2013, "There really are 50 Eskimo words for 'snow')—it was because snow was important to them and they needed to describe many different subtle variations. We in Biocomputing also like to (and sometimes need to) make distinctions between different approaches and technologies to analyzing biological and medical data, so no word is perfect. In fact, "Biocomputing" is not a particularly preferred word in current parlance and a google search for it produces results that might be called anemic. In fact, biocomputing engenders some confusion—does it mean computing *with* biological matter or computing *about* biology? Who knows--PSB 2118 may be dominated by biological computers. In any case, the term Biocomputing has served the conference well—it is just ambiguous enough to allow us to include whatever makes sense. The session organizers critically evaluate trends in the field and bring exciting and emerging new branches of Biocomputing in their proposals, and the organizers get to decide which ones to feature. This year is no exception, with a variety of new and old topics from Biocomputing.

The mission of PSB is to provide a forum for the best *emerging* science in Biocomputing, providing both formal and informal mechanisms for scientific communication. PSB depends on the community to define emerging areas in biomedical computation. Its sessions are usually conceived at the previous PSB meeting as people discuss trends and opportunities for new science. The typical program includes sessions that evolve over two to three years as well as entirely new sessions. This year we revisit topics such as precision medicine, pattern matching and text-mining, while nurturing emerging interest in genomic drug discovery, open science, imaging genomics, and the interpretation of noncoding DNA

In addition to being published by World Scientific and indexed in PubMED, the proceedings from all PSB meetings are available online at http://psb.stanford.edu/psb-online/. PSB has 1066 papers listed in PubMED (as of today). These papers are routinely cited in archival journal articles and often represent important early contributions in new subfields—many times before there is an established literature in more traditional journals; for this reason, many papers have garnered hundreds of citations. The Twitter handle PSB 2018 is @PacSymBiocomp and the hashtag this year will be #psb18.

The efforts of a dedicated group of session organizers have produced an outstanding program. The sessions of PSB 2018 and their hard- working organizers are as follows:

Applications of Genetics, Genomics and Bioinformatics in Drug Discovery
Richard Bourgon, Rick Dewey, Zhengyan Kan, and Dan Li

Challenges of Pattern Recognition in Biomedical Data
Anurag Verma, Anna Basile, Marta Byrska-Bishop, Christian Darabos, and Shefali Setia Verma

Democratizing Health Data for Translational Research
Philip Payne, Nigam Shah, Jessie Tenenbaum, and Lara Mangravite

Imaging Genomics
Heng Huang, Junzhou Huang, Kun Huang, Li Shen, Paul M. Thompson, and Lin Yang

Precision Medicine: from diplotypes to disparities towards improved health and therapies
Bruce Aronow, Steven E. Brenner, Dana C. Crawford, Joshua C. Denny, and Alexander A. Morgan

Reading Between the Genes: Computational Models to Discover Function and/or Clinical Utility from Noncoding DNA
Yves Lussier, Maricel Kann, Jason Moore, Kenneth Ramos, Joanne Berghout, and Francesca Vitali

Text Mining and Visualization for Precision Medicine
Graciela Gonzalez, Casey Greene, Hongfang Liu, and Abeed Sarker

We are also pleased to present four workshops in which investigators with a common interest come together to exchange results and new ideas in a format that is more informal than the peer-reviewed sessions. For this year, the workshops and their organizers are:

Diversity and Disparity in Biomedical Informatics
Philip R.O. Payne, William M. Southerland, S. Joshua Swamidass, Laura Wiley, and ClarLynda Williams-DeVane

Integrating Community-level Data Resources for Precision Medicine Research
Dana C. Crawford and William S. Bush

Machine Learning and Deep Analytics for Biocomputing: Call for Better Explainability
Dragutin Petkovic, Lester Kobzik, Christopher Re

Methods for Examining Data Quality in Healthcare Integrated Data Repositories
Vojtech Huser, Michael Kahn, and Jeffrey Brown

We thank our keynote speakers Carlos Bustamante (Science keynote) and Jennifer Wagner (Ethical, Legal and Social Implications keynote).

Tiffany Murray has managed the peer review process and assembly of the proceedings since 2003, and also plays a key role in many aspects of the meeting. We are grateful for the support of the Institute for Computational Biology, a collaborative effort of Case Western Reserve University, the Cleveland Clinic Foundation, and University Hospitals; the Institute *for* Informatics (I^2), Washington University in St. Louis, School of Medicine; and Pfizer for their support of PSB 2018. We also thank the National Institutes of Health[1] and the International Society for Computational Biology (ISCB) for travel grant support. The research parasite and symbiont awards benefit by support from: GigaScience, Scientific Data, Springer Nature, Jeff Stibel, Mr. and Mrs. Stephen Canon, and Drs. Casey and Anna Greene.

We are particularly grateful to the onsite PSB staff Al Conde, Ryan Whaley, BJ Morrison-McKay, Cynthia Paulazzo, Jackson Miller, Kasey Miller, and Paul Murray for their assistance. We also acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

[1] Funding for this conference was made possible (in part) by Grant # 5 R13 LM006766 – 20 from the National Library of Medicine. The views expressed in written conference materials or publications, and by speakers and moderators, does not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

We look forward to a great meeting once again. Aloha!

Pacific Symposium on Biocomputing Co-Chairs,
October 15, 2017

Russ B. Altman
Departments of Bioengineering, Genetics, Biomedical Data Science & Medicine, Stanford University

A. Keith Dunker
Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

Lawrence Hunter
Department of Pharmacology, University of Colorado Health Sciences Center

Marylyn D. Ritchie
Department of Biomedical and Translational Informatics, Geisinger Health System

Teri E. Klein
Department of Biomedical Data Science & Medicine, Stanford University

Thanks to the reviewers…

Finally, we wish to thank the scores of reviewers. PSB aims for every paper in this volume to be reviewed by three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Vida Abedi
Melinda Aldrich
Gil Alterovitz
Vinayagam Arunachalam
Chloe-Agathe Azencott
Christopher Bauer
Brett Beaulieu-Jones
Andrew Beck
Matt Breitenstein
William Bush
Mariusz Butkiewicz
Weidong Cai
Colin Campbell
Hang Chang
Ronghua Chen
Yin Hoon Chew
Keith Ching
Brian Cole
Jessica Cooke Bailey
Mark Craven
Dana Crawford
David Crosslin
Ying Ding
Lei Du
Jian Fang
Ruogu Fang
Di Feng
Julio Fernandez
Marc Fink
Jason Fries
Lana Garmire
Olivier Gevaert
Mario Giacobini
Anthony Gitter
Benjamin Glicksberg
Rachel Goldfeder
Jean-Philippe Gourdine
Casey Greene
Jacob Hall
Yangyang Hao
Steve Hart
Jon Hill
Joshua Hoffman
Brittany Hollister
Ting Hu
Junzhou Huang
Jake Hughey
Wendy Ingram
Janina Jeff
Suiwang Ji
Kenneth Jung
Anne Justice
Zhengyan Kan
Sarvnaz Karimi
Jonathan Karr
Ramakanth Kavuluru
Dokyoon Kim
Ravikumar Komandur
Robert Kueffner
William La Cava
Samir Lal
Nicholas Larson
Robert Leaman
Dingcheng Li
Haiquan Li
Hongyang Li
Ruowang Li
Hai Lin
Hongfang Liu
Jingyu Liu
Sijia Liu
Tianming Liu
Gang Luo
Yves Lussier
Meng Ma
Shannon McWeeney
Jason Miller
Tejaswini Mishra
Diego Molla
Sean Mooney
Jason Moore
Xia Ning
Matt Oetjens
Randy Olson
Patryck Orzechowski
Casey Overby
Shraddha Pai
Chetanya Pandya
Michael Paul
Philip Payne
Stephen Pfohl
Hoifung Poon
Nicole Restrepo
Dingyao Ruo
Satya Sahoo
Abeed Sarker
Matthew Scotch
Nigam Shah
Li Shen
Marina Sirota
Gregory Sliwoski
Sandra Smieszek
Yang Song
Ryan Sullivan
Hanna Suominen
Jeff Sutherland
Jessica Tenebaum
Nathan Tintle
Manabu Torii
Anna Tyler
Jacob Ulirsch
Ryan Urbanowicz
Fabio Vandin
Rohit Vashisht
Olivia Veatch
Sudha Veturi
Francesca Vitali
Chen Wang
Junwen Wang
Xiaoqian Wang
Yalin Wang
Jeremy Warner
Vivian West
Scott Williams
Chunlei Wu
Song Wu
Tao Xie
Jun Xu
Rong Xu
Jingwen Yan
Lin Yang
Xiaohui Yao
Meliha Yetisgen
Kun Yu
Yong Yue
Baohong Zhang
Chi Zhang
Daoqiang Zhang
Shaoting Zhang
Yan Zhang
Yanfei Zhang
Dajiang Zhu
Xiaotong Zhu
Daniel Ziemek
Pascal Zille

APPLICATIONS OF GENETICS, GENOMICS AND BIOINFORMATICS IN DRUG DISCOVERY

RICHARD BOURGON
Genentech Inc.
South San Francisco, CA 94080
Email: bourgon.richard@gene.com

FREDERICK E. DEWEY
Regeneron Genetics Center
Tarrytown, NY 10591
Email: frederick.dewey@regeneron.com

ZHENGYAN KAN
Pfizer Inc.
San Diego, CA 92121
Email: Zhengyan.Kan@pfizer.com

SHUYU D. LI
Sema4, a Mount Sinai venture
Stamford, CT 06902
Icahn School of Medicine at Mount Sinai
New York, NY 10029
Email: shuyu.li@sema4genomics.com

As the impact of genetics, genomics, and bioinformatics on drug discovery has been increasingly recognized, this session of the 2018 Pacific Symposium on Biocomputing (PSB) aims to facilitate scientific discussions between academia and pharmaceutical industry on how to best apply genetics, genomics and bioinformatics to enable drug discovery. The selected papers focus on developing and applying computational approaches to understand drug mechanisms of action and develop drug combination strategies, to enable *in silico* drug screening, and to further delineate disease pathways for target identification and validation.

1. Introduction

Drug discovery and development continues to face the challenges of rising cost and declining productivity. While the estimated average cost to bring a new molecular entity to market has exceeded US$ 1.5 billion, R&D return on investment fell considerably from 10.1% in 2010 to 3.7% in 2016 [1]. Recent advances in genetic and genomic research has not only accelerated our studies of disease mechanisms, but also enabled drug discovery in many areas. For example, the power of human genetics in therapeutic target validation has been underscored by a retrospective analysis that selecting targets with supportive human genetics evidence doubled the success rate in

clinical development [2]. A recent report on the clinical impact of loss-of-function (LoF) genetic variants in 50,726 exomes confirmed previously known associations between genes such as PCSK9 and cardiovascular disease-related phenotypic traits, and identified novel associations with therapeutic implications [3]. Genomics and genetics also play an increasingly important role in other areas in drug discovery such as biomarker identification for drug efficacy [4] and safety [5], understanding drug mechanisms of action [6], and selecting disease relevant experimental models [7]. To facilitate the application of genomics in drug discovery, data quality and reproducibility have been systematically assessed [8] to increase our confidence on findings from pharmacogenomic studies. Furthermore, new methods and tools have been developed for integrative genomic data analysis [9].

Although the impact of genetics, genomics and bioinformatics in drug discovery has been recognized by both academia and pharmaceutical industry, the coverage of the topic in scientific conferences is very limited. The main objective of this session "Applications of genetics, genomics and bioinformatics in drug discovery" in the 2018 PSB is to cover recent advances in developing and applying computational approaches to enable drug discovery in the above-described areas. Furthermore, the session is also intended to promote more interactions and collaborations between academic and industry experts. We believe such a session dedicated to bioinformatics in the context of drug discovery could significantly benefit academic preclinical drug discovery activities, as a large number of academic drug discovery centers have been established in recent years [10]. We define the following topics and problems are within the scope of this session.

- Target identification and validation: integrative analysis of molecular data at scale, coupling genetic, epigenetic, gene expression profiling, proteomic, metabolomic, phenotypic trait measurements to disease diagnosis and clinical outcome data to generate hypotheses on molecular etiology of diseases in service of identification or validation of novel therapeutic targets.
- Biomarker discovery: utilizing genetic and genomic data derived from cell lines, animal models, human disease tissues and PBMC to develop preclinical or clinical biomarkers for target engagement, pharmacodynamics, drug response, prognosis, and patient stratification; applying genomic profiling in clinical trials to identify early response markers to predict clinical end points.
- Pharmacogenomics: identify associations between germline SNPs, somatic mutations, gene expression and other molecular alterations and drug responses.
- Toxicogenomics: integrative analysis of genomic, histopathology, and clinical chemistry data to develop predictive toxicology biomarkers in preclinical 4-day, 14-day and 30-day studies and clinical studies.
- Understanding drug mechanisms of action (MoA): applying genomic profiling to de-convolute targets and delineate MoA of non-selective drugs or drugs from phenotypic screening.
- Characterization of mechanisms of acquired resistance: analysis of genetic and genomic data derived from preclinical isogenic models or clinical patient samples to study the mechanisms of acquired resistance.

- Selection of disease-relevant experimental models: comparative analysis of genetic and genomic data to assess and select cell line and animal models in drug discovery that best represent the disease indications.
- Developing drug combination strategies: analysis of genetic and genomic data to identify synthetic lethality genes as drug combination targets; computational analysis to understand gene regulatory networks to develop combination strategies that target parallel pathways or reverse drug resistance.
- Drug repurposing: applying *in silico* approaches to identify new disease indications for existing drugs.
- Novel methods and tools for multi-omics data integration, analyses, and visualization.

2. Session Contributions

A total of eight papers were selected from the submissions. We categorized the eight papers into the following three groups. The papers in the first group focus on drug mechanisms of action and drug combinations. The second group includes studies that can be applied to enable computational drug screening. The papers in the third group apply various computational approaches on genetic and genomic data to further understand diseases.

2.1. *Drug mechanisms of action and drug combinations*

Understanding drug mechanisms of action is critical in clinical development and precision medicine, particularly in identifying early response markers as surrogates for clinical end-point as well as biomarkers for patient stratification. In addition, a better knowledge of MoA may allow us to reposition the existing drugs for new indications. In the study by *Luo et al.* [11], the authors developed a novel method, referred as Mania, for scalable data integration incorporating chemical structure, drug sensitivity and gene expression changes in response to drug treatment. Drug similarity networks were first constructed based on each of these data sources, followed by integration through Mania into a low-dimensional vector representation of each drug. It was shown that integration of various data sources improves quantification of drug-drug similarities, and achieves more accurate prediction of drug targets and MoA. Functionally enriched "drug communities", as referred by the study, was also identified using the low-dimensional vector representation matrix. Finally, the authors illustrated potential utilities of their new method by analyzing the most significantly mutated genes across 21 tumor types in the cancer genome atlas (TCGA) and presented examples of drugs that are predicted to target some of the significantly mutated cancer genes.

Gene expression profiling in cell lines in response to drug perturbation has provided a valuable tool to study drug MoA. Although a large number of drugs have been profiled in many cancer cell lines of various tissue origins, there are still substantial missing drug-cell line combinations in these data sources. *Hodos et al.* [12] attempted to fill the gaps by predicting cell specific drug perturbation expression profiles. The authors developed a computational framework to first arrange existing gene expression profiles into a three-dimensional array (or tensor) indexed by drugs, genes, and cell types, and then use either local (nearest-neighbors) or global (tensor

completion) information to predict unmeasured profiles. The prediction accuracy was thoroughly evaluated and it was found that the two methods (local vs. global) have complementary performance, each superior in different regions in the drug-cell space. Finally, the authors demonstrated that the predicted profiles add value for downstream prediction of drug targets and therapeutic classes. For example, it was shown the classifiers trained on the complete dataset are of higher quality than those trained only on the measured dataset, with particularly significant impact on those cell types with fewer measured profiles available.

Drug treatment may induce alternative splicing as a key response event with functional consequences. However, limitation of short-read sequencing poses a barrier to accurately detect different splicing isoforms. *Chen et al.* [13] described characterization of the transcriptional splicing landscape in a prostate cancer cell line treated with a previously identified synergistic drug combination, by using a combination of third generation long-read RNA sequencing technology and short-read RNA-seq to create a high-fidelity map of expressed isoforms and fusions to quantify splicing events triggered by treatment. The authors found strong evidence for drug-induced, coherent splicing changes that disrupt the function of oncogenic proteins, and detected novel transcripts arising from previously unreported fusion events. The study demonstrated the benefit of long-read technology in identifying highly homologous isoforms routinely and with high fidelity.

Most patients with advanced cancers ultimately develop drug resistance to chemotherapy or targeted therapy due to reactivation of the same pathway or compensatory pathways. Combination therapy targets multiple pathways, therefore may improve efficacy and also overcome drug resistance in some cases. *Xu et al.* [14] presented a novel computational approach to predict combinations through assessing the potential impact of inhibiting a drug target on disease signaling network. Using melanoma as an example to apply the approach, the authors first constructed a disease network by integrating gene expression profiling and protein-protein interaction data. A drug-disease "impact matrix" was computed using network diffusion distance from drug targets to signaling network elements. The drugs were then clustered into "communities" that are supposed to share similar mechanisms of action. Finally, drug combinations maximally impacting signaling sub-networks are ranked and proposed as potential combination strategies for melanoma.

2.2. *Drug metabolism and in silico drug screening*

Human gut bacteria have the ability to activate, deactivate, and reactivate drugs with huge implications in drug efficacy and toxicity at individual patient level. Understanding the complete space of drug metabolism by the human gut microbiome is critical for predicting bacteria-drug relationships and their effects on drug response. To address the challenge that there are limited computational tools for predicting drug metabolism by the gut microbiome, *Mallory et al.* [15] developed a pipeline for comparing and characterizing chemical transformations using continuous vector representations of molecular structure based on unsupervised learning, and characterized the utility of vector representations for chemical reaction transformations. After clustering molecular and reaction vectors, enriched enzyme names, Gene Ontology terms, and Enzyme

Consortium (EC) classes were detected within the reaction clusters. Finally the authors queried reactions against drug-metabolite transformations known to be metabolized by the human gut microbiome, and showed the top results for these known drug transformations contained similar substructure modifications to the original drug pair. The method described in this study could be potentially applied in high throughput screening of drugs and their resulting metabolites against chemical reactions common to gut bacteria.

The study by *Greenside et al.* [16] addresses a critical component in drug discovery, identification of small molecule ligands that bind to the target proteins as a first step in drug screening. While the currently available computational tools for predicting protein-ligand binding largely rely on 3D protein structure, this study described an interpretable confidence-rated boosting algorithm to predict protein-ligand interactions with high accuracy from ligand chemical substructures and protein 1D sequence motifs, without relying on 3D protein structures. The authors showed that their models can be generalized to unseen proteins and ligands, demonstrating the possibility to predict protein-ligand interactions using only motif-based features and that interpretation of these features can reveal new insights into the molecular mechanics underlying each interaction.

2.3. *Disease genes and pathways*

Novel computational approaches have been continuously developed and applied to analyze genetic and genomic data. Recently, deep learning has emerged as a novel class of machine leaning methods. While deep learning has been applied in many domains such as speech recognition, image recognition, natural language processing, its application in analyzing genomic data is very limited. *Way et al.* [17] applied variational autoencoders (VAEs), an unsupervised deep neural network approach to analyze TCGA gene expression profiling data. Specifically, the extent to which a VAE can be trained to model cancer gene expression, and whether or not such a VAE would capture biologically relevant features were evaluated. The paper introduced a VAE trained on TCGA pan-cancer RNA-seq data, identified specific patterns in the VAE encoded features, and discussed potential merits of the approach. To illustrate the utility of VAEs in further delineating cancers, the authors described examples from their analyses on significant pathways separating primary and metastatic melanoma, and on pathways over-represented in different subtypes of high-grade serous ovarian cancer.

Human genetic data based on genome-wide sequencing or genotyping, coupled with hospital electronic medical records (EMRs) have provided a powerful tool to study the genetic basis of human diseases. *Smith et al.* [18] described integrative analysis of genetic data derived from DNA samples in a biobank and the accompanying clinical diagnosis information in EMRs to identify several neuroplasticity genes associated with neurodevelopmental diseases. The authors first developed a neuroplasticity gene signature from two independent gene expression profiling datasets. Subsequently, carriers of loss-of-function (LoF) genetic variants in the neuroplasticity genes were identified in the biobank cohort. The authors then performed an association analysis to discover significant associations between LoF in neuroplasticity genes and neurodevelopmental

diseases. Finally, a thorough literature review was described to demonstrate the validity of the results.

3. Acknowledgments

We thank our respective organizations for supporting our involvement in organizing the session. We also thank the following reviewers for providing expert reviews of the submitted manuscripts: Vinayagam Arunachalam, Kristin Ayer, Ronghua Chen, Keith Ching, Ying Ding, Di Feng, Julio Fernandez, Marc Fink, Rajarshi Guha, Yangyang Hao, Jon Hill, Kipp Johnson, Robert Kueffner, Samir Lal, Hai Lin, Meng Ma, Gianni Panagiotou, Chetanya Pandya, Kiran Patil, Jeff Sutherland, Alex Tropsha, Song Wu, Tao Xie, Kun Yu, Yong Yue, Baohong Zhang, Chi Zhang, Yan Zhang, Xiaotong Zhu, Daniel Ziemek.

References

1. Mullard, A. R&D returns continue to fall. *Nature reviews. Drug discovery* **16**, 9 (2016).
2. Nelson, M.R. et al. The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860 (2015).
3. Dewey, F.E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science (New York, N.Y.)* **354** (2016).
4. Kelloff, G.J. & Sigman, C.C. Cancer biomarkers: selecting the right drug for the right patient. *Nature reviews. Drug discovery* **11**, 201-214 (2012).
5. Khan, S.R., Baghdasarian, A., Fahlman, R.P., Michail, K. & Siraki, A.G. Current status and future prospects of toxicogenomics in drug discovery. *Drug discovery today* **19**, 562-578 (2014).
6. Nijman, S.M. Functional genomics to uncover drug mechanism of action. *Nature chemical biology* **11**, 942-948 (2015).
7. Horvath, P. et al. Screening out irrelevant cell-based models of disease. *Nature reviews. Drug discovery* **15**, 751-769 (2016).
8. Haverty, P.M. et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333-337 (2016).
9. Fernandez-Banet, J. et al. OASIS: web-based platform for exploring cancer multi-omics data. *Nature methods* **13**, 9-10 (2016).
10. Dahlin, J.L., Inglese, J. & Walters, M.A. Mitigating risk in academic preclinical drug discovery. *Nature reviews. Drug discovery* **14**, 279-294 (2015).
11. Luo, Y., Wang, S., Xiao, J. & Peng, J. Large-Scale Integration of Heterogeneous Pharmacogenomic Data for Identifying Drug Mechanism of Action. *Pacific Symposium on Biocomputing* **23** (2017).
12. Hodos, R. et al. Cell-specific prediction and application of drug-induced gene expression profiles. *Pacific Symposium on Biocomputing* **23** (2017).
13. Chen, X. et al. Characterization of drug-induced splicing complexity in prostate cancer cell line using long read technology. *Pacific Symposium on Biocomputing* **23** (2017).
14. Xu, J. et al. Diffusion Mapping of Drug Targets on Disease Signaling Network Elements Reveals Drug Combination Strategies. *Pacific Symposium on Biocomputing* **23** (2017).

15. Mallory, E.K., Acharya, A., Rensi, S.E., Bright, R.A. & Altman, R.B. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. *Pacific Symposium on Biocomputing* **23** (2017).
16. Greenside, P., Hillenmeyer, M. & Kundaje, A. Prediction of protein-ligand interactions from paired protein sequence motifs and ligand substructures. *Pacific Symposium on Biocomputing* **23** (2017).
17. Way, G.P. & Greene, C.S. Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes with Variational Autoencoders. *Pacific Symposium on Biocomputing* **23** (2017).
18. Smith, M.R. et al. Loss-of-function of Neuroplasticity-related genes confers risk for human neurodevelopmental disorders. *Pacific Symposium on Biocomputing* **23** (2017).

Characterization of drug-induced splicing complexity in prostate cancer cell line using long read technology

Xintong Chen[1], Sander Houten[1], Kimaada Allette[1], Robert P.Sebra[1], Gustavo Stolovitzky*[1,2] and Bojan Losic*[1]

1. *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, NY 10029, USA*
2. *IBM Translational Systems Biology and Nanobiotechnology Research, Yorktown Heights, NY 10598, USA*

**Corresponding Authors: Email: bojan.losic@mssm.edu, gustavo@us.ibm.com*

Abstract

We characterize the transcriptional splicing landscape of a prostate cancer cell line treated with a previously identified synergistic drug combination. We use a combination of third generation long-read RNA sequencing technology and short-read RNAseq to create a high-fidelity map of expressed isoforms and fusions to quantify splicing events triggered by treatment. We find strong evidence for drug-induced, coherent splicing changes which disrupt the function of oncogenic proteins, and detect novel transcripts arising from previously unreported fusion events.

Keywords: Combination treatment; Long read sequencing; Alternative splicing; Cancer.

Introduction

Background

Prostate cancer is the second-most common cancer and has the third leading cancer mortality among men in the USA.[1] Major clinical interventions for prostate cancer include surgical procedure, radiation, androgen depletion treatment (ADT) and chemotherapy. As with other cancers, prognosis of prostate cancer varies largely depending on its molecular characteristics [2]. A number of large-scale collaborative efforts and crowd-sourcing initiatives have recently been used to profile genomic data on cancer systems perturbed by thousands of compounds to infer agents with curing potential. The Library of Integrated Network-Based Cellular Signatures (LINCS) Program, for example, provides a rich public source of gene expression data collected from cell lines with exposure to various compounds. These data allows researchers to gain mechanistic insight into the biological processes that are altered by different drugs in a given cellular context (cell line) [http://www.lincsproject.org/]. On the analytical side, the NCI-DREAM Drug Sensitivity Prediction Challenge, run in 2012, encouraged the development of algorithms to predict the sensitivity of cancer cell lines to a panel of drugs based on multi-omics

data, including gene expression, copy number variation, mutation and proteomics data [3]. The AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge, launched in 2015, is a similar international competition seeking for algorithms that accurately predict synergistic combination treatment based on gene expression data and multiple cancer cell lines [4]. Despite the importance of these and other efforts in probing the drug treated omics expression landscape, splicing modulation in treatment has remained largely unexplored.

Alternative splicing (AS) events are key generators of proteomic diversity. Yet the functional impact of alternative splicing is only now beginning to be systematically quantified, thanks, in part, to new technological advances that overcome the difficulties related to the read length of sequencing assays to detect isoforms. Indeed, the short-read length of second generation sequencing technology (usually 50bp or 100bp) directly leads to the key difficulty of unambiguously phasing isoforms and mapping highly repetitive sequence. Third generation sequencing platforms such as PacBio and Oxford Nanopore utilize long read sequences to help address this issue. Isoform sequencing (IsoSeq) is a recently developed PacBio assay which can directly sequence full-length transcript sequences. With the help of IsoSeq, an astonishing diversity of splicing events in various systems has started to emerge even in well-studied cancer cell-line systems such as MCF7[5]. A number of other analyses in other cellular contexts have also utilized the IsoSeq assay[6-11]. Given that alternative splicing is known to be tissue and condition dependent, we hypothesized that drug administration should also alter the splicing landscape of cells as part of multifaceted cellular response to stimuli which cannot be completely captured using standard RNAseq analysis.

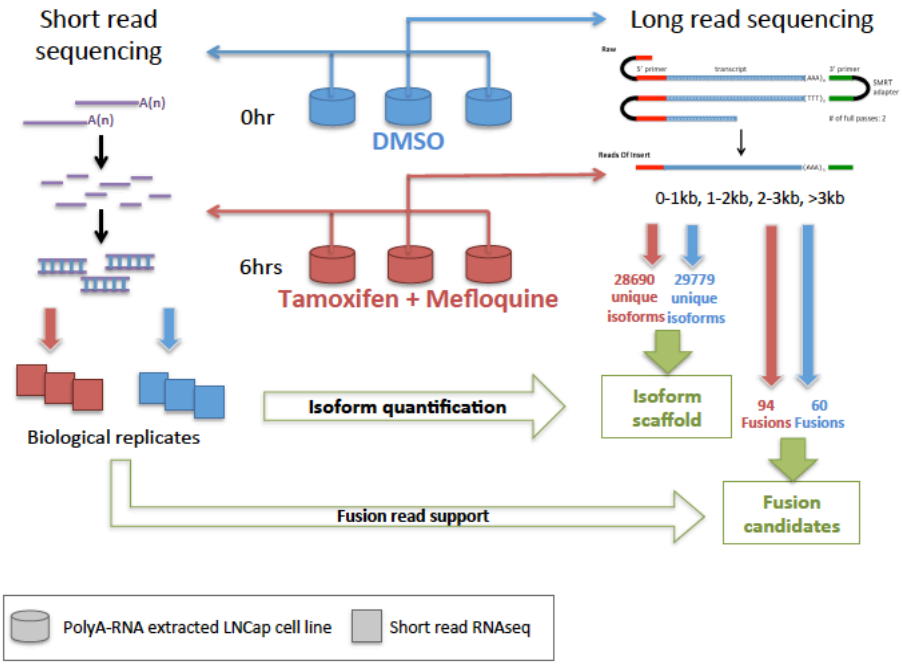


Figure1. A schematic of study design. Three biological replicates of short read RNAseq were generated for untreated (DMSO) and treated (TM) and used for subsequent analysis; long read sequencing were performed for untreated and treated samples. Both technologies were combined for isoform quantification and identification of fusion events.

Experimental design

We used LNCap prostate cell lines as our cancer model. After treatment with a previously identified synergistic combination treatment of Tamoxifen and Mefloquine(TM)[4, 12], we measured the LNCap cell viability as a function of time relative to treatment with DMSO and found that cell viability decreases to ~30% at times as early as 6h. As shown in Figure 1, we generated three biological replicates from a baseline condition with LNCap cells cultured in DMSO at 0hr; and a treated condition where cells were cultured with Tamoxifen and Mefloquine at 6hrs. We collected polyA enriched RNA from these six samples and performed short read RNAseq experiments. We also randomly selected one of the replicates from each condition to perform long read PacBio isoform sequencing (IsoSeq) (Figure1). To avoid loading bias against short transcripts, we performed size-selection for multiple bins (0-1kb, 1-2kb, 2-3kb and >3kb) using SageELF device for both samples. Purified SMRTbell libraries were sequenced on the PacBio RSII machine using P6-C4 chemistry with 8 SMRT cells.

Results

Bioinformatics pipeline for IsoSeq

We wrapped up a pipeline for IsoSeq analysis (Figure S1). In brief, by searching for sequencing adapters, reads of insert were classified as Circular Consensus Sequences (CCS) and non-CCS, and further classified as full length or not through searching of 5' and 3' primers and polyA signals. Next, we performed isoform-level-clustering (ICE) and Quiver polish on non-artifacts full length (FL) reads to improve error corrected consensus accuracy via SMRT Portal, yielding 30669 and 31095 FL consensus isoforms for baseline and treated conditions respectively, with expected accuracy greater than 0.99. All high quality FL transcripts were aligned to the hg19 genome using GMAP[13] with default parameters and collapsed to remove redundant sequences via the pbtranscript-TOFU package [https://github.com/PacificBiosciences/cDNA_primer/wiki] with minimum alignment accuracy of 0.99 and minimum coverage of 0.85. Transcripts sharing the exact same exons except those with an extended 5' end were collapsed into a single transcript. The unique transcripts were then aligned back using STAR[14] and compared with the gencode.v19 database[15] by MatchAnno [https://github.com/TomSkelly/MatchAnnot]. Junction modes were predicted for each transcript using Astalavista[16] through classifying splice sites of alternative splicing external(ASE), alternative splicing internal (ASI), splice sites that extend transcript structures(DSP) and any other differences in transcript structures not involving splice sites(VST). From there we computed open reading frames (ORFs) with at least 100 amino acids from all high quality FL unique transcripts and screened for homology to known proteins by

aligning to UniProtKB/Swiss-Prot protein database (BLASTP, e-value<1e-5) and applying hmmer (3.1b1)[17] to scan for protein domains (Pfam,E<10). Finally, TransDecoder (3.3.0) was used to leverage blast hit and detected domains to look for coding regions.

	#Unique isoform	Coding potential	Known isoforms	Novel isoforms			Others (unknown gene/not aligned)
				Exons match except size	Some exons match	Novel exons	
TM6hrs	28690	74.1% (21257)	31% (8894)	5%	46%	14%	4%
DMSO 0hr	29779	73.2% (21808)	36% (10720)	5%	40%	15%	4%

Table1. General statistics of PacBio FL isoforms detection.

Isoform complexity in LNCap

We detected 29779 and 28690 unique full-length (FL) isoforms from IsoSeq data of un-treated and treated LNCap cell line respectively (see methods), as shown in Table 1. Transcripts mapped to unknown gene locus or not aligned to the reference genome are marked as others and not included in the following comparison. For comparison, following the same procedure in the publicly available high quality deep sequenced (with 119 SMRT Cells) MCF7 IsoSeq data (2013 version) we detected 40447 unique FL isoforms. For all three datasets, the majority of the detected isoforms are novel: 60%, 65% and 71% for un-treated LNCap, treated LNCap and MCF7 respectively. We characterized the novel isoforms into three categories: 1) Novel isoforms with exons matching to exons one-for-one, but sizes of the internal exons may disagree; 2) Novel isoforms with only some of the exons matching annotated exons, and 3) Novel isoform overlap with annotated genes, but no exon matches annotation. The majority of detected novel isoforms fall into category (1) and (2), as shown in Table1. These "partially novel" isoforms are likely due to undocumented splicing patterns in current annotations, while only a small proportion of FL isoforms are completely novel (7%~15%).

These results imply that there is non-trivial information in the fine splitting of the short-read expression spectrum (over our long-read scaffold) across some genomic loci into distinct transcripts. In fact, the entropy S_i at a given gene locus i **for a given sample** is computed as the empirical Shannon entropy of the normalized, variance stabilized expression frequencies across all **informative** transcripts T_{ij} at that locus, namely the average log

$$S_i = -< \ln [P(T_{ij})] >_{P(T_{ij})} \quad (1)$$

such that $P(T_{ij})$ is the probability (normalized frequency) of j-th informative transcript T to appear at genomic locus i. The frequency bins are estimated using Sturge's rule [18]. A transcript is defined as **informative** if it is on average well expressed after being penalized by fitting a

global mean variance trend [19] of the short-read expression dataset. Indeed, this penalty is proportional to the **inverse variance** of the transcript expression, which increases as expression decreases due to amplification noise. We thus demand that $P(T_{ij})$ in equation (1) only includes transcripts that satisfy

$$E'_{ij} > \mu(E_{ij}\,\omega ij) \quad (2)$$

such that E_{ij} is the variance stabilized expression values, ωij is a quality factor proportional to the inverse count variance, and μ is mean of the distribution of $E_{ij}\,\omega ij$ at each locus i. Note that genomic loci with zero or only one informative transcript are naturally assigned an entropy of zero and $E_{ij}\,\omega ij$ tends to zero with low expression much faster than E_{ij}. Length-bias and sample variations are treated by dividing each S_i by the **number of transcripts** at the ith locus and then taking their **median across all samples**, i.e.

$$S' = median\,(\frac{S_i}{N}) \quad (3)$$

Figure 2A depicts the distribution of S' and its dependence on overall expression, from which it is clear that fine-splitting entropy per transcript is not a trivial artifact dominated by noisy, lowly expressed transcripts in outlier samples. It is also clear that while many genomic loci are well approximated by collapsing all transcripts to a single genic locus -- and thus have few informative distinct transcripts which corresponds to a low entropy (the 'zero mode') -- many others do not and likely imply nontrivial biological regulation.

To better characterize detected IsoSeq isoforms, we surveyed their coding potential. Through scanning of ORFs (>100 aa) in protein databases, protein homologies were then leveraged to maximize coding regions prediction sensitivity. We found 73.2% and 74.1% of detected isoforms were predicted to have coding potential which strongly suggests functional consequences of splicing events in the system. Taken together, novel splicing events leading to isoforms with coding potential are observed in LNCap cell lines with comparable statistics in external MCF7 IsoSeq data.

Treatment induced nontrivial splicing signals in LNCap cell line.

Only 10417 of the detected isoforms overlapped between treated (19362 treated-unique) and untreated (18273 untreated-unique) conditions. To analyze the functional impact of these isoforms, we perform functional annotation through classification of protein families and domains (see methods). In brief, we found 603 and 569 domains in DMSO and TM isoforms respectively, with 553 domains in common and recurrent (>2) condition-unique domains highlighted in text (Figure 2.B). We observed key oncogenic protein domains (families) frequently found in DMSO but not in treatment, such as histone deacetylase interacting, PI3-kinase, p85- and Ras binding

domains and DNA polymerase A/B families. Notably frizzled protein, which activates the Wnt pathway is ranked as the top DMSO-unique proteins[20]. A full list of condition specific domain annotation is shown in Table S1. We also observed a similar pattern through a different protein family prediction algorithm as shown in Figure S3. Thus we hypothesize that the decrease in survival under drug treatment may be implemented by breaking the oncogenic domains in LNCap by induction of targeted splicing events. For example, histone deacetylation (HDAC) has been found to play major role in prostate cancer progression making HDAC inhibitor a potential anti-tumor therapeutic target.[21, 22] In our IsoSeq data, HDAC interacting domain is only detected in isoforms in LNCap+DMSO but not the TM treated condition, while overall gene expression of *HDAC1* (encoding histone deacetylase family as a component of the histone deacetylase complex) is up-regulated in treatment of TM (fold change>1.68, FDR< 3.4e-07), indicating treatment induced splicing at the locus may shut down histone deacetylation regardless of increased gene expression.

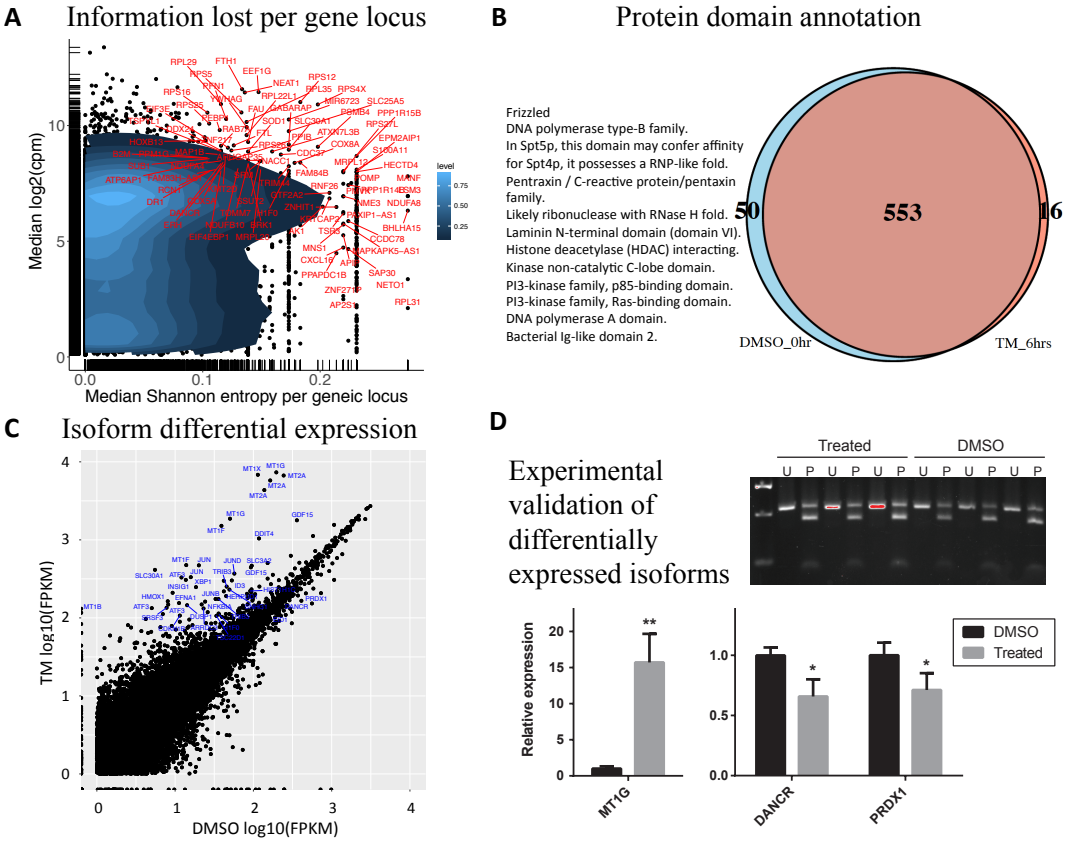


Figure2. A) Median normalized fine splitting entropy for each genomic locus is on x-axis with median expression on y-axis, outlier genes are labeled in red. B) Venn diagram of predicted protein domains in DMSO and TM. Frequently detected domains (frequency>=3) are highlighted in text. C) Plot of FPKM in log10 scale of DMSO (x-axis) VS FPKM in log10 scale of TM (y-axis), well expressed significant

differentially expressed isoforms (FDR<0.05, FPKM >1) are gene labeled in blue. D) Upper panel: MT1G PCR fragment (325bp) uncut (U) or digested with PstI (P). The intensity of the digested fragment is higher indicating that the NM_001301267 transcript is more prominent; lower panel: Quantitative PCR for MT1G (both isoforms), DANCR (NR_024031.2) and PRDX1 (NM_181697.2). Relative expression is calculated as 2 to the power of the Ct of RPLP0 - Ct of the tested gene. Average of the DMSO control group is set to 1.

We also find that the relative prevalence of key splicing modalities (exon skipping, intron retention, alt donor/receptor, etc.) among the detected isoforms shows no significant bias (Figure S2) although we do note that generally the baseline level of intron retention exceeds that found in the reference annotation [23]. We further summarize intron-retention events for known and novel isoforms (Table S2), partial novel isoforms has similar distribution as the known ones while novel isoforms with novel exons appear to be much fewer. Despite this lack of specificity of splicing modality, we conclude that treatment related isoforms are gaining and losing key functional protein domains previously identified to be important for cancer progression[21, 22, 24, 25].

Quantification and validation of IsoSeq FL isoforms using short read data

Although long read technology creates an integrated transcriptome scaffold, it is suboptimal for expression quantification essentially due to a relatively sparse and non-uniform coverage profile. Following previous work[5, 6], we used short read replicates to quantify IsoSeq isoforms (Figure 1). In brief, we generate an assembled GTF file summarizing all IsoSeq FL transcripts and their genomic information (hg19) and assign short reads from the 6 RNAseq samples to its features. Replicates per condition were grouped together to improve abundance estimates for isoform differential expression analysis between treatments and untreated. We detected 38 highly expressed isoforms up-regulated and 3 isoforms down-regulated in the treatment group relative to DMSO. The parent genes of these isoforms are known to be involved in cancer related pathways, such as *JUN*, *DDIT4*, *CDKN1B* and transcription factor *ATF3*. We also found a splice variant of a cell differentiation regulating long non-coding RNA: *DANCR* being down-regulated in treatment (Figure 2.C). We selected 3 differentially and well expressed isoforms (fpkm>100) for further experimental validation. The genes and variants of interest were first amplified by PCR and product sizes were estimated by agarose gel electrophoresis followed by Sanger sequencing to confirm correct amplification. The *MT1G* mRNA had two variants; NM_005950 and NM_001301267. The latter variant uses an alternate in-frame splice site in the 5' coding region and has a 3bp insert that introduces a PstI restriction site. A restriction digest revealed that most of the PCR product is digested by PstI indicating that NM_001301267 is the most prominent transcript (Figure 2.D), which was further confirmed by Sanger sequencing. Of the two tested

DANCR isoforms, we detected only NR_024031.2. For *PRDX1*, two variants were detected, but NM_181697.2 was most prominent. We performed quantitative PCR for *MT1G* (both isoforms), *DANCR* (NR_024031.2) and *PRDX1* (NM_181697.2). Treatment of cells increased *MT1G* mRNA expression, but decreased *DANCR* and *PRDX1* confirming the short read and IsoSeq data.

Fusion landscape revealed in treatment

Gene fusions have been found to play a major role in prostate tumorigenesis. We leveraged the long read lengths of the IsoSeq assay to create a detailed map to track the expression of fusion transcripts. Using HQ consensus reads we infer a set of fusion transcripts and further filter these candidates with stringent threshold of short read support (see Method), we detected 94 fusions (83 inter-chromosomal and 11 intra-chromosomal) and 60 fusions (55 inter-chromosomal and 5 intra-chromosomal) in treated and untreated, respectively (Figure S5). The number of fusion candidates found in treated condition more than in untreated with very few in common, suggest treatment may reveal the altered genome structure. We collected a list of 14 known LNCap fusions from literature[26-28](Table S4), and found 6 of them present in DMSO and 2 present in treated, in the set of IsoSeq fusion transcripts candidates. We also compare the long read detected fusions with short read fusion calls inferred from chimeric junctions. To our surprise, short read detects only 5 fusion candidates in untreated and only 3 fusions in treated. All short read called fusions in untreated cells are found in IsoSeq fusion transcripts before being filtered by short read support. Since any short read fusion caller must effectively exclude low-complexity regions in the genome to constrain false positives [https://github.com/LosicLab/starchip] this dramatically reduces our power in exploring full set of fusions since over 49% of human genome is composed of repetitive sequences[29, 30]. We find that our final set of well supported IsoSeq fusions preferentially originates from repeat enriched region (76 out of 94, 52 out of 60 coming from repetitive region for TM and DMSO respectively, $p < .01$) further reinforcing a key advantage of long read technology to complement current short read technology.

Discussion

The role of alternative splicing in the transcriptomic landscape in general and more so as a post-treatment cellular response is just starting to be explored. Our understanding of the splicing complexity is benefitting from the advances in long-read technology, which is allowing for the identification of highly homologous isoforms routinely and with high fidelity. Our results strongly suggest that there is ample biological and clinical relevance for transcript sensitive modulation in prostate cancer treatment and, we speculate, in many other disease systems in cancer and beyond.

To our knowledge, our work is the first effort to characterize the significance of treatment-induced alternative splicing in a cancer model using a de-novo assembled isoform reference via long read sequencing. Our results suggest treatments induce a large number of varied alternative splicing events that alter known oncogenic proteins. Crucially, although overall gene expression is *higher* in TM compared to DMSO, our functional analysis shows that the splice variants in fact lead to *fewer* functional protein products in treatment. Indeed, we observed more intron-retention in untreated cells and that treated cells preferentially splice in/out oncogenic domains. For example, we find that ER stress marker *JUN* is down-regulated in treatment (log2 fold change = -3.5; FDR < 4e-8), however we also find that treatment specifically activates certain domains within JUN such as leucine zipper domain. This key domain facilitates DNA binding and participation in dimerization [21,22], and we suspect forms a central element in cellular response to the treatment. Finally, our fusion analysis uncovers a number of novel fusion candidates, including treatment specific examples, which are currently undergoing validation. We hope to generalize this method to more cell lines/treatments to investigate if the observed splicing signal observed is a global mechanism for anti-cancer treatment and shed light on drug resistance study. In summary, it seems plausible that an entirely new layer of transcriptomic regulation in cancer-drug interactions is finally becoming amenable to systematic study and enabling novel pathway and target discovery.

Methods

Materials

Our LNCap cell line is from clone FGC (ATCC CRL-1740). The cells were plated at a density of 8,000 cells per well in a 96well plate (Greiner Cat. No. 655083) and placed in an incubator. After 24 hours, the plates were removed from the incubator and treated with drugs using the HP D300 Digital Dispenser. The cells were then collected at the targeted time point (0hr for control; 6hrs for after TM combination treatment) by removing the media and pipetting 150uL of Qiagen Buffer RLT into each well. The plates were then frozen and stored at -80C.

RNA Preparation and Illumina RNA sequencing

The Qiagen RNeasy 96 kit (Cat. No. 74181) was used to extract RNA, with the Hamilton ML STAR liquid handling machine equipped with a Vacuubrand 96 well plate vacuum manifold. A Sorvall HT six floors centrifuge was used to follow the vacuum/spin version of the RNeasy 96 kit protocol. The samples were treated with DNAse (Rnase-Free Dnase Set Qiagen Cat. No.79254) during RNA isolation. The RNA samples were then tested for yield and quality with the Bioanalyzer and the Agilent RNA 6000 Pico Kit. The TruSeq Stranded mRNA Library Prep Kit

(RS-122-2101/RS-122-2102) was then used to prepare the samples for 30 million reads of single end sequencing (100bp) with the Illumina HiSeq2500.

Quantitative PCR validation

We used 15ng of poly(A)+ RNA for cDNA synthesis with the SuperScript IV first strand synthesis system (Thermo Fisher Scientific) and random hexamers as primers. Quantitative PCR was performed using the ABI Prism 7900HT with Bio-rad iQ SYBR Green Mastermix. The primers used for these studies are designed using primer3 and listed in table S3.

Short read data analysis

Raw reads from Illumina sequencing were aligned to a hg19 genome with features from GTF file (hg19) downloaded from UCSC genome browser using STAR[14] with chimSegmentMin=15; chimJunctionOverhangMin=15, outSAMmapqUnique=60 and other parameters as default. The output chimeric junctions are used for short read fusion caller STARCHIP with >10 junction reads support and repeat region penalty as 0.5 [https://github.com/LosicLab/starchip].

Short read integration for isoform quantification

Raw reads from Illumina sequencing were aligned to a hg19 genome but with a scaffold generated from both IsoSeq samples using STAR (2.5.2b)[14] with parameters described in table S5. IsoSeq isoforms were quantified using Cufflinks (2.2.1)[31]. Aligned short reads of each sample are assigned to features of IsoSeq scaffold to estimate transcript abundance. Cuffdiff [31] were called to compare expression between treated and untreated condition with default parameters, three replicates are fed to Cuffdiff to increase statistical power.

Fusion transcript detection from IsoSeq

Search criteria for fusion transcripts included mapping to at least two distinct genomic loci and each mapped locus has to cover >10% of the transcript, with all mapped loci combined covering >99% of the entire transcript. The mapped loci must be at least 100kbp apart from each other. We further filter the fusion transcripts based on short read support. For each candidate transcript, we aligned short reads against it using bwa mem (0.7.15)[32] and require >90% of the transcript to have at least 40 read supports at each base position for all three replicate samples since the number of fusions start to saturate from 40 read support as shown in Figure S4.

Functional annotation of IsoSeq transcripts

InterProScan(5.15-54.0)[33] was used to infer isoform functions. These includes two steps: 1) Screening for protein domains/ functional motifs through PRINT [34] and SMART[35] with default setting for all unique FL isoforms. 2) Obtained domains were then mapped to GO terms

through InterPro[36]. We compared and visualized GO annotations for baseline and treated specific isoforms using WEGO[37].

Estimates of information volume of splicing events

To estimate the cumulative information loss of averaging all distinct transcripts (i.e. ignoring splicing) in a given experiment, it is straightforward to remove the correction due to locus length in Eq. (3) and sum up the sample-median loci entropies across loci to obtain a naive upper limit estimate of order **kilobit**. Simply summing Eq. (3) across all genomic loci instead reduces this by an order of magnitude to $O(100)\ bits$, and any co-splicing will also obviously reduce this estimate by reducing the number of independent splicing events. In the present work we see strong evidence of coherent splicing patterns but cannot estimate co-splicing with this number of samples. By comparison, however, gene *expression* states are often well approximated by a three-state system ($S \sim 1.09\ bits$) and we verify that this is the case of our data as well across samples. Thus, a naive upper limit estimate is such that $S_{splicing} \sim 10^2\, S_{expression}$. In the absence of a significant co-splicing effect this implies that splicing dynamics contain significantly more information than expression dynamics from the point of view of bulk RNA-seq.

Supplementary

All supplementary materials are hosted on https://chenx08.u.hpc.mssm.edu/PSB_CHEN/

Acknowledgements

We thank Ronald B. Realubit and Charles Karan from Columbia Genome Center for providing LNCap cells and performing Illumina sequencing.

References

1. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.
2. Cancer Genome Atlas Research, N., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell, 2015. **163**(4): p. 1011-25.
3. Costello, J.C., et al., *A community effort to assess and improve drug sensitivity prediction algorithms.* Nat Biotechnol, 2014. **32**(12): p. 1202-12.
4. Bansal, M., et al., *A community computational challenge to predict the activity of pairs of compounds.* Nat Biotechnol, 2014. **32**(12): p. 1213-22.
5. Weirather, J.L., et al., *Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing.* Nucleic Acids Res, 2015. **43**(18): p. e116.
6. Au, K.F., et al., *Characterization of the human ESC transcriptome by hybrid sequencing.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. E4821-30.
7. Treutlein, B., et al., *Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing.* Proc Natl Acad Sci U S A, 2014. **111**(13): p. E1291-9.
8. Wang, B., et al., *Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing.* Nat Commun, 2016. **7**: p. 11708.
9. Abdel-Ghany, S.E., et al., *A survey of the sorghum transcriptome using single-molecule long reads.* Nat Commun, 2016. **7**: p. 11706.

10. Singh, N., et al., *IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform.* Meta Gene, 2016. **7**: p. 70-5.
11. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nat Biotechnol, 2013. **31**(11): p. 1009-14.
12. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
13. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics, 2005. **21**(9): p. 1859-75.
14. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
15. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
16. Foissac, S. and M. Sammeth, *ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W297-9.
17. Eddy, S.R., *A new generation of homology search tools based on probabilistic inference.* Genome Inform, 2009. **23**(1): p. 205-11.
18. Sturges, H.A., *The choice of a class interval Case I Computations involving a single Series.* Journal of the American Statistical Association, 1926. **21**: p. 65-66.
19. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
20. Ueno, K., et al., *Frizzled homolog proteins, microRNAs and Wnt signaling in cancer.* Int J Cancer, 2013. **132**(8): p. 1731-40.
21. Abbas, A. and S. Gupta, *The role of histone deacetylases in prostate cancer.* Epigenetics, 2008. **3**(6): p. 300-9.
22. Halkidou, K., et al., *Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer.* Prostate, 2004. **59**(2): p. 177-89.
23. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.
24. Ransone, L.J., et al., *Fos-Jun interaction: mutational analysis of the leucine zipper domain of both proteins.* Genes Dev, 1989. **3**(6): p. 770-81.
25. Kouzarides, T. and E. Ziff, *The role of the leucine zipper in the fos-jun interaction.* Nature, 1988. **336**(6200): p. 646-51.
26. McPherson, A., et al., *Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data.* Bioinformatics, 2011. **27**(11): p. 1481-8.
27. Maher, C.A., et al., *Chimeric transcript discovery by paired-end transcriptome sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(30): p. 12353-8.
28. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.
29. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
30. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
31. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
32. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
33. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.
34. Attwood, T.K., et al., *The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012.* Database (Oxford), 2012. **2012**: p. bas019.
35. Schultz, J., et al., *SMART: a web-based tool for the study of genetically mobile domains.* Nucleic Acids Res, 2000. **28**(1): p. 231-4.
36. Finn, R.D., et al., *InterPro in 2017-beyond protein family and domain annotations.* Nucleic Acids Res, 2017. **45**(D1): p. D190-D199.
37. Ye, J., et al., *WEGO: a web tool for plotting GO annotations.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W293-7.

Prediction of protein-ligand interactions from paired protein sequence motifs and ligand substructures

Peyton Greenside
Program in Biomedical Informatics, Stanford University
Stanford, CA 94305
Email: pgreens@stanford.edu

Maureen Hillenmeyer
Stanford Genome Technology Center, Stanford University
Palo Alto, CA 94304
Email: maureenh@stanford.edu

Anshul Kundaje
Departments of Genetics and Computer Science, Stanford University
Stanford, CA 94305
Email: akundaje@stanford.edu

Identification of small molecule ligands that bind to proteins is a critical step in drug discovery. Computational methods have been developed to accelerate the prediction of protein-ligand binding, but often depend on 3D protein structures. As only a limited number of protein 3D structures have been resolved, the ability to predict protein-ligand interactions without relying on a 3D representation would be highly valuable. We use an interpretable confidence-rated boosting algorithm to predict protein-ligand interactions with high accuracy from ligand chemical substructures and protein 1D sequence motifs, without relying on 3D protein structures. We compare several protein motif definitions, assess generalization of our model's predictions to unseen proteins and ligands, demonstrate recovery of well established interactions and identify globally predictive protein-ligand motif pairs. By bridging biological and chemical perspectives, we demonstrate that it is possible to predict protein-ligand interactions using only motif-based features and that interpretation of these features can reveal new insights into the molecular mechanics underlying each interaction. Our work also lays a foundation to explore more predictive feature sets and sophisticated machine learning approaches as well as other applications, such as predicting unintended interactions or the effects of mutations.

Keywords: ligand, protein, interactions, motifs, drug discovery, quantitative structure-activity relationships, QSAR

1. Introduction

1.1. *Decreasing returns in drug discovery pipelines*

Return on investment in drug discovery efforts continues to decline, an observation that has been called Eroom's law, or Moore's law spelled backward [Scannell 2012]. Despite the use of high-throughput drug screens and increasingly sophisticated experimental and computational methods, many compounds that initially appear promising fail in later stages of testing after substantial investment has already been made. Predicting drug efficacy and toxicity as early as possible is of great advantage in an increasingly difficult drug discovery pipeline.

1.2. *Existing methods for prediction of protein-ligand interactions*

The most cost-effective procedure for screening compounds at any early stage would enable computational methods before more expensive, time-consuming experiments in vitro or in animals. There have been numerous efforts to computationally predict protein-ligand interactions (often referred to as quantitative structure-activity relationship or QSAR models). However, many of these methods rely on solved 3D structures to locate the binding pocket or other key features of the protein [Wang et al. 2014, Ragoz et al 2016, Ewing et al. 2001, Leach et al. 2006, Liang et al. 2003]. However only a small proportion of proteins, mapping to around 5,000 unique genes, with complete amino acid sequences have resolved 3D crystal structures [Berman et al. 2000, Hicks et al. 2017]. Thus, as an alternative to secondary or tertiary protein structure, we explore the use of features derived directly from known 1D sequences of proteins to describe the protein in the absence of structural information. This approach allows us to train models spanning a much larger collection of proteins.

Unlike methods relying on 3D structure, we featurize 1D protein sequences using protein motifs (amino acid-based) and ligand motifs (substructure-based) and learn combinations of these motif pairs that underlie each protein-ligand interaction. It is known that protein-ligand binding largely occurs at a single broad site (i.e. the binding pocket) in the protein and a specific set of atoms in the ligand. If we can successfully learn the properties of these molecular interfaces in the form of motif interactions, we could potentially screen for interactions for proteins that lack well annotated crystal structures.

Several methods have attempted to predict protein-ligand interactions without 3D structural data. One class of methods [Jacob and Vert 2008] uses the therapeutic class of a protein in conjunction with the chemical structure of the ligand. Other classes of methods [Campillos et al. 2008] follow a "guilt by association" principle by trying to determine similarity of an uncharacterized compound to other compounds with known targets. Both of these methods rely on existing annotated knowledge that limit extension to classes of uncharacterized proteins. We show that we are able to use sequence-based features of proteins and structure-based features of ligands to predict protein-ligand interactions without depending on existing targets or similar annotations.

The primary advantage of our approach is that it relies only on protein and ligand motifs without needing crystallized structures or direct comparisons to well understood protein-ligand interactions. In addition, we are able to look at which protein-ligand motif pairs combine together in a given interaction, thereby localizing the interaction to key parts of the protein and ligand. This work also lays a foundation to use the same motif-based approach to predict unintended secondary interactions of a compound, which is useful as many drugs also fail for such off-target effects. Further extensions enabled by reducing interactions to key motifs are assessing the impact of mutations through their effect on essential motifs underlying the interaction and predicting which protein variants are most susceptible to altered drug binding.

2. Methods

2.1. *Data set*

With the goal of predicting protein-ligand interactions from only protein sequence and ligand structure features, we first aggregated several sources of known protein-ligand interactions for training data. From the PubChem database of 3D structures with ligands from the Protein Data Bank (PDB) [Berman et al. 2000] we identified 69,959 proteins with a ligand bound. From the DrugBank [Wishart et al. 2006] we extracted 10,888 interactions of FDA-approved and experimental drugs with their known protein targets. From BindingDB [Chen et al. 2001] we used 30,925 interactions of drug target proteins and small molecules.

In order to eliminate identical or nearly-identical interactions, we compared all protein sequences to each other using pairwise BLAST (blastp), and grouped proteins having >90% identity. This resulted in 16,357 proteins and 25,118 ligands with a total of 62,561 positive interactions. We sampled protein-ligand pairs for training and test sets using different cross-validation strategies (see Section 2.5) to have a 1:100 imbalance of positives to negatives. The data set as a whole contains ~0.1% positives, but we found improved training and feasible construction of cross-validation folds with a 1:100 imbalance.

2.2. *Protein Featurization*

We used a bag-of-features representation for protein sequences, which does not preserve sequential order of features. We explored several different types of features to represent protein sequences. We used conserved signatures, defined by Prosite motifs [Hulo et al. 2008] and Pfam domains [Finn et al. 2006]. Prosite motifs include biologically meaningful residues, including but not limited to binding domains, post-translational modification sites and other active sites. Pfam domains are conserved protein domains based on multiple alignments and hidden Markov model profiles. We used two types of Prosite domains: “all Prosite” (all original motifs) and “short Prosite” motifs that exclude any motifs longer than 50 amino acids in order to avoid motifs that cover a majority of the protein. We trained models on all three motif types. Our featurizations resulted in 1,472 unique Prosite motifs, 1,071 short Prosite motifs and 3,324 Pfam motifs after limiting only to motifs that appeared in our data set. Each protein has an average of 5.8 all Prosite motifs, 1.5 short Prosite motifs and 1.5 Pfam motifs.

2.3. *Ligand Featurization*

We also used a bag-of-features representation for ligands. To featurize ligands we used structural signatures, also known as chemical substructures. PubChem [Kim et al. 2005] makes freely available a set of 554 substructures in the SMARTS format [Weininger 1988] that can be scanned in the ligand. 299 of these substructures were present in at least one compound in our full set of protein-ligand interactions. Ligands had an average of 22.8 substructure motif features.

2.4. *Boosting Model*

Our classification task was set up as the prediction of a binary matrix of protein-ligand interactions from paired feature sets of proteins and ligands, one for each dimension of the interaction matrix. Proteins are represented by the presence/absence of a set of protein motifs. Ligands are represented by the presence/absence of a set of ligand motifs. This formulation of the learning problem thus involves three binary matrices: matrix ***I*** for interactions of size (# proteins, # ligands) that we try to predict from matrix ***P*** for protein features of size (# proteins, # protein motifs) and matrix ***L*** for ligand features of size (# ligands, # ligand motifs). **Figure 1** illustrates the data set up.

To learn the prediction function, we use an algorithm known as MEDUSA [Kundaje et al. 2008] based on confidence-rated boosting algorithms [Schapire et al. 1999] and specifically optimized for learning from factorized paired interacting feature spaces. We implemented the algorithm as an efficient, parallelized software package called PFBoost **[**Greenside et al. 2017**]**. The algorithm iteratively minimizes the exponential loss to learn the structure and composition of a model known as an Alternating Decision Tree (ADT), which is a margin-based generalization of decision trees [**Figure 1**]. The algorithm begins with an empty ADT. All training examples are initially assigned an equal weight. Iteratively, (i) the algorithm learns a rule (called a splitter node) which is a simple binary predictor based on the presence of a protein motif-ligand motif pair (shown as rectangle boxes in **Figure 1, Right**), with an associated score (shown as ovals prediction nodes in **Figure 1, Right**), that minimizes the exponential loss across all training examples; and (ii) learns the optimal position for the rule to be added to the current structure of the ADT, either to the root prediction node or conditionally following another prediction node elsewhere in the ADT. After each iteration, the training examples are re-weighted according to the error in the current predictions to prioritize finding rules for incorrectly predicted examples in the subsequent iterations.

A path in the ADT is defined as any subset of contiguous, connected nodes from the root node down to a terminal node. Each path captures conditional dependencies in the rules. A training/test example can satisfy a rule (presence/absence of a motif-pair) encoded in a node in a path only if the example satisfies all the rules in the nodes that precede it in the path, i.e. the example has all the motif-pair rules in the preceding nodes. The final ADT model is thus an ensemble of conditional rules that allow a quantitative ‘prediction score’ on a protein-ligand example as the sum of the scores of all rules in valid paths that the example satisfies. The sign of the prediction score indicates the predicted binary output (interaction or no interaction) and the magnitude indicates the confidence of prediction.

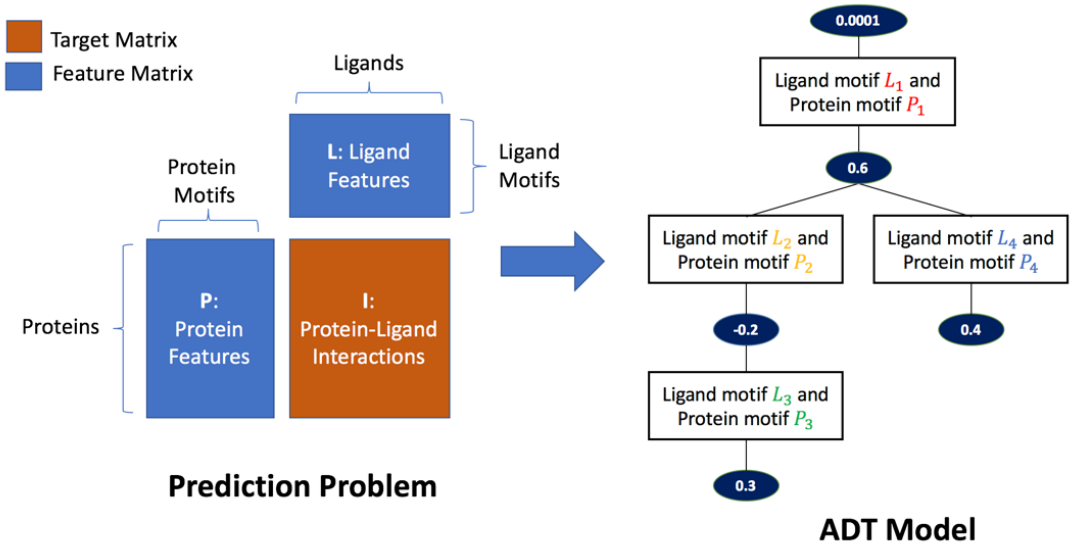


Figure 1 Left: Prediction of protein-ligand interactions from two feature matrices. Right: The resulting Alternating Decision Tree (ADT) model where each splitter node contains a protein motif-ligand motif pair rule and the following prediction node contains the score for the rule.

This setup lends itself naturally to the prediction of a protein-ligand interaction as a function of combinations of protein motif-ligand motif pairs that are involved in the interaction. In our efficient, parallelized implementation, we further take advantage of linear algebra tricks involving sparse matrix operations that can implicitly compute the loss of all pairs of motif features at each boosting iteration without explicitly storing the outer product of the two feature spaces [Kundaje et al. 2008]. These optimizations allow us to scale to large datasets.

2.5. *Cross Validation Approaches*

We used cross validation in order to assess our model's ability to predict held out protein-ligand interactions. In order to avoid inflated performance, we also clustered proteins based on their homology (see Section 2.1) into mutually exclusive groups and made sure that proteins belonging to the same group did not occur in the same training or testing folds. We used 10-fold cross validation, and we implemented four cross validation approaches to appropriately evaluate our ability to generalize to different categories of data:

1. Random holdout: We randomly sample entries in the target matrix to hold out. This results in seeing the same proteins and ligands in both training and test sets although not the same combinations.
2. Column holdout: We randomly sample some columns (ligands) to hold out entirely from the training set.
3. Row holdout: We randomly sample some rows (proteins) to hold out entirely from the training set.
4. Quadrant holdout: We randomly sample some columns (ligands) and rows (proteins) to hold out entirely from the training set.

3. Results

3.1. *Model Performance*

We trained our models for 4,000 iterations after observing that performance begins to plateau around that point [**Figure 2**]. Due to the significant class imbalance of the dataset we evaluated our predictive accuracy with auROC and auPRC. Boosting is well known for its inherent resistance of overfitting. We observe high performance for random holdout with minimal overfitting [**Table 1**]. However, a more interesting question is how well the model performs when entire classes of ligands and proteins are held out. As expected we observe a larger degree of overfitting when holding out groups of proteins or ligands as opposed to random entries [**Table 1**], but the model seemed to generalize quite well to held out ligands and had more difficulty with held out proteins. This result could be partly explained by sparse featurization of proteins as well as the grouping of homologous proteins into common cross-validation folds.

Table 1 Performance for all types of cross validation and for the three protein feature sets.

	All Prosite	**Pfam**	**Short Prosite**
auROC, random holdout (Train/Test)	0.96/0.95	0.95/0.93	0.94/0.93
auPRC, random holdout (Train/Test)	0.45/0.42	0.42/0.41	0.31/0.30
auROC, ligand holdout (Train/Test)	0.96/0.93	0.92/0.82	0.97/0.93
auPRC, ligand holdout (Train/Test)	0.46/0.32	0.38/0.19	0.39/0.31
auROC, protein holdout (Train/Test)	0.98/0.84	0.93/0.84	0.97/0.85
auPRC, protein holdout (Train/Test)	0.54/0.23	0.40/0.20	0.38/0.27
auROC, quadrant holdout (Train/Test)	0.96/0.95	0.91/0.86	0.97/0.95
auPRC, quadrant holdout (Train/Test)	0.45/0.34	0.36/0.29	0.38/0.34

We compared different types of motif feature sets for proteins - all Prosite, Pfam and short Prosite motifs - and we found that highest predictive power is achieved with all Prosite motifs by only a small amount. This is likely at least partly due to the larger number of Prosite motifs per protein, which give a richer featurization. As a result, we performed all downstream analysis on the model trained with all Prosite motifs and random holdout.

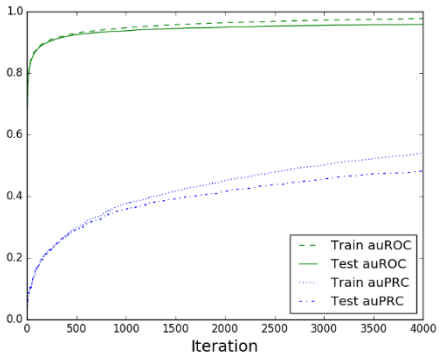


Figure 2 Learning curve over 4,000 iterations for prediction with all Prosite motifs and random holdout.

3.2. *Most predictive motif features*

The ADT model has all the advantages of a boosting-based ensemble method but also retains interpretability since it is a single generalized tree structure in contrast

to the more standard boosted ensemble of decision trees. We analyzed the ADT to identify predictive protein motifs, ligand motifs and interactions of protein motifs and ligand motifs. We found that 595 of 1,472 Prosite protein motifs and 237 of 299 ligand substructure motifs were selected by the algorithm at some node in the model. We can directly assess how much each feature (protein or ligand motif) or feature pair (protein motif-ligand motif pair) in the ADT contributed to the overall margin of prediction (true label*prediction score) for each example by computing the difference in the margin score before and after nullifying the contribution of all nodes from the ADT containing the feature or feature pair. This essentially deletes these nodes from the ADT and re-computes the prediction as though they did not exist. We rank each protein motif-ligand motif pair from most to least "important" by its total effect on the prediction margin [**Table 2**]. We also separately compute the margin scores for each protein motif and ligand motif, separating out those that contributed to a positive or negative prediction.

Table 2 Top 10 nodes in the ADT for contribution to the margin of prediction for positive interactions.

	Prosite Name	Prosite ID	SMART string	SMART name
1	PROTEIN_KINASE_DOM	PS50011	C(-C)(-N)	Ethylamine
2	PKC_PHOSPHO_SITE	PS00005	C(-C)(-O)	Ethanol
3	CK2_PHOSPHO_SITE	PS00006	C(-N)(=O)	Formamide
4	PROTEIN_KINASE_DOM	PS50011	C=O	Formaldehyde
5	MYRISTYL	PS00008	C(-O)(-O)	Methanediol
6	TYR_PHOSPHO_SITE	PS00007	C(-C)(-O)	Ethanol
7	CK2_PHOSPHO_SITE	PS00006	C(-C)(-O)	Ethanol
8	PKC_PHOSPHO_SITE	PS00005	O-C-C-C-O	1,3-Propanediol
9	CK2_PHOSPHO_SITE	PS00006	C(-C)(-N)	Ethylamine
10	CAMP_PHOSPHO_SITE	PS00004	C(-C)(-N)	Ethylamine

To determine whether the margin scores for each feature were significant, we computed empirical *p*-values by shuffling the target matrix 100 times and calculating the number of times the feature's margin score was greater than or equal to the set of margin scores for all features over all permuted matrices. This resulted in 2149 significant nodes or protein-ligand motifs pairs that significantly contributed to a positive protein-ligand interaction.

3.3. *Known positive examples*

Our margin-ranked features and feature pairs identified many compelling protein-ligand motif pairs that explain known protein-ligand interactions.

3.3.1. *Uricase - Uric acid*

One of the significant nodes in our model shows PS00366 binds ligand motif O=C-N-C-N. PS00366 is uricase, which binds uric acid. The ligand motif O=C-N-C-N looks just like part of the

uricase structure [**Figure 3**]. Further, we could confirm in PDB that there are many instances where PS00366 is within 5 angstroms of O=C-N-C-N [Berman et al. 2000].

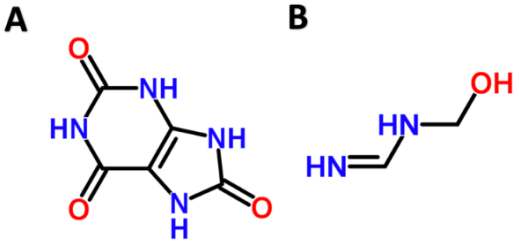


Figure 3 Uric acid (A) is bound by Uricase. The ligand motif O=C-N-C-N (B) was paired with Uricase in our model.

3.3.2. *Chloramphenicol O-acetyltransferase – Chloramphenicol*

In another example, we found that PS00100, which is Chloramphenicol O-acetyltransferase, binds C(-Cl)(-Cl). C(-Cl)(-Cl) is a substructure of Chloramphenicol [**Figure 4**] and Chloramphenicol is the target of Chloramphenicol O-acetyltransferase.

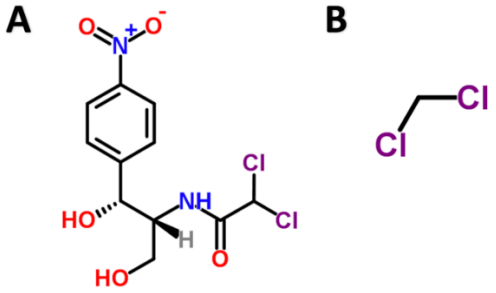


Figure 4 Chloramphenicol (A) and substructure C(-Cl)(-Cl) (B), which was paired with Chloramphenicol O-acetyltransferase in our model.

3.3.3. *Transthyretin –T4*

We found PS00768 paired to substructure motif Oc1c(Br)cccc1. PS00768 is transthyretin, a thyroid hormone-binding protein that is known to transport thyroxine (T4) from the bloodstream to the brain [Sigrist et al. 2012]. The substructure Oc1c(Br)cccc1 is 2-bromophenol, which looks just like one of the key substructures in thyroxine with one halogen replaced for another [**Figure 5**].

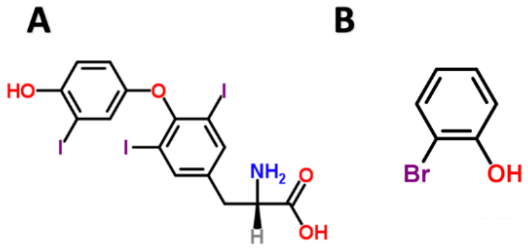


Figure 5 Thyroxine (A) is a hormone transported by transthyretin. 2-bromophenol (B) is a substructure motif paired with transthyretin in our model.

3.4. *Interpreting ADT Paths*

3.4.1. *Path lengths*

While it is encouraging to see many well known interactions explained by specific protein-ligand motifs, it is even more interesting to understand how these motif pairs combine with each other. Each path of size *n* in the ADT represents a combination of *n* protein motifs paired with *n* ligand motifs that additively give a final prediction of an interaction. When there is little interaction between features we often see stumps or short paths, but many paths in our model showed predictive combinations of up to 9 motif pairs [**Figure 6**].

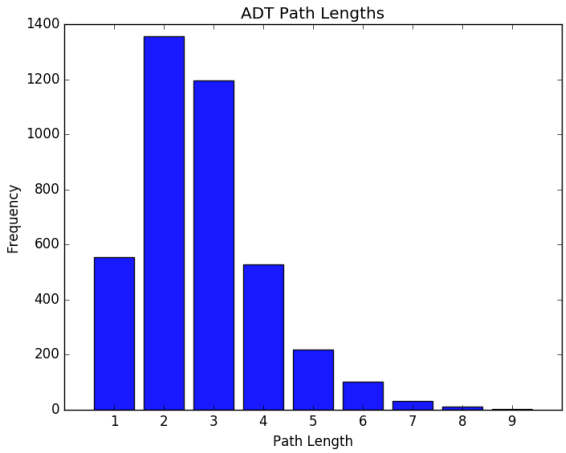


Figure 6 Distribution of path lengths in the ADT model suggests many interactions between protein-ligand motif pairs.

3.4.2. *Protein kinase C – Phosphatidylserine*

One such compelling example is a path of length 6 consisting of 4 ligand substructures (2 repeated in different nodes) all paired with PS00005, which is Protein kinase C phosphorylation site. Protein kinase C binds phosphatidylserine [Newton 1995]. The four unique ligand substructures were P(-O)(=O), C(~C)(~O), P(~O)(~O), and O=C-C-N, all of which resemble substructures of phosphatidylserine [**Figure 7**].

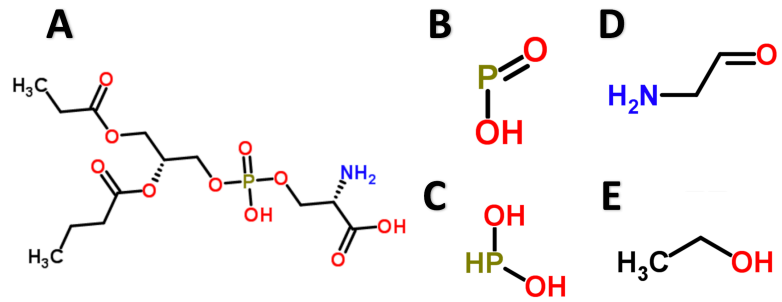


Figure 7 Phosphatidylserine (A) is known to bind to Protein kinase C. Four substructure motifs (B, C, D, E) in the same path all paired with Protein kinase C and resemble substructures in (A).

4. Discussion

We have presented a method to predict protein-ligand interactions from only protein sequence motifs and ligand substructure motifs, without relying on 3D structure of proteins or known targets. This is one of the few methods that bridges chemical knowledge of ligand structures with biological knowledge of protein sequence in order to predict interactions between the two with high accuracy. In addition to demonstrating that it is possible to predict protein-ligand interactions using only motif-based featurizations, we further demonstrate that it is possible to extend our predictions to entirely held out ligands and also, although to a lesser extent, held out proteins.

We then demonstrate that we are able to interpret protein and ligand motif features from the model. There is limited existing knowledge on how individual motifs in ligands and proteins interact with one another in a given protein-ligand interaction without a 3D structure. We show that some of the pairs that we recover as most important in the model are actually known pairings from crystallized structures and thus our predictions may extend to non-crystallized structures. We also show how we can rapidly propose hypotheses about essential motif pairs with predictive pairings from our boosting model. In future work, we would like to extend the same modeling effort to predict interactions that may be off-target effects, although it is much harder to obtain enough confident training labels in that application.

Based on the success of using known protein motifs and ligand substructures, we envision generating novel feature sets that may provide even greater resolution. Protein motifs annotated in Prosite and PFAM are annotated with ligand-binding domains, which we have successfully recapitulated; the next step is to extend to unannotated features of sequences. It may be useful to extend the method to more general motifs such as protein *k*-mers or electrostatic groups of those *k*-mers in place of known protein motifs. We see an opportunity to apply more sophisticated methods such as multi-modal deep neural networks, which could learn powerful de-novo features from raw protein sequences and ligands, and functional embeddings. While there are numerous opportunities to extend this work, we have shown that even simple motif-based approaches can achieve competitive accuracy in protein-ligand predictions and can provide useful interpretation.

Acknowledgments

We are grateful for support from the Burroughs Wellcome Career Award at the Scientific Interface (MH) and BioX Stanford Interdisciplinary Graduate Fellowship (PG).

References

1. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. Nucleic Acids Res, 28(l):235-42.
2. Campillos, M., Kuhn, M., Gavin, A.-C, Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. Science, 321(5886):263-6.
3. Chen, X., Lin, Y., and Gilson, M. (2001). The binding database: overview and user's guide. Biopolymers, 61(2):127-41. Journal Article United States.
4. Dassault Systèmes BIOVIA, Pipeline Pilot, San Diego: Dassault Systèmes, 2008.
5. Ewing, T., Makino, S., Skillman, A., and Kuntz, I. (2001). Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. / Comput Aided Mol Des, 15(5):411-28.
6. Finn, R., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S., Sonnhammer, E., and Bateman, A. (2006). Pfam: clans, web tools and services. Nucleic Acids Res, 34(Database issue):D247-51.
7. Greenside P, Hussami N, Chang J, Kundaje A. PyBoost: A parallelized Python implementation of 2D boosting with hierarchies. bioRxiv 170803; doi: https://doi.org/10.1101/170803
8. Hicks M, Bartha I, Iulio J, Abagyan R, Venter C, Telenti A. Functional characterization of 3D-protein structures
9. Informed by human genetic diversity. bioRxiv 182287; doi: https://doi.org/10.1101/182287.
10. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B., de Castro, E., Lachaize, C , Langendijk- Genevaux, R, and Sigrist, C. (2008). The 20 years of prosite. Nucleic Acids Res, 36(Database issue):D245-9.
11. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4; 44(D1):D1202-13. Epub 2015 Sep 22 [PubMed PMID: 26400175] doi: 10.1093/nar/gkv951.
12. Kotz, Treichel, and Weaver (2008). Chemistry and Chemical Reactivity, Enhanced Review Edition.
13. Kundaje, A., Xin, X., Lan, C, Lianoglou, S., Zhou, M., Zhang, L., and Leslie, C. (2008). A predictive model of the oxygen and heme regulatory network in yeast. PLoS Comput Biol, 4(ll):el000224.
14. Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006). Prediction of protein-ligand interactions, docking and scoring: successes and gaps. J Med Chem, 49(20):5851-5.
15. Liang, M. P., Banatao, D. R., Klein, T. E., Brutlag, D. L., and Altman, R. B. (2003). WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucleic Acids Res, 31(13):3324-7.
16. Newton, A. C. (1995). Protein kinase C: structure, function, and regulation. *Journal of Biological Chemistry*, *270*(48), 28495-28498.
17. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2016). Protein-Ligand Scoring with Convolutional Neural Networks. *arXiv preprint arXiv:1612.02751*.
18. Royal Society of Chemistry (2015). ChemSpider. http://www.chemspider.com/. Accessed July 2017.
19. Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery*, *11*(3), 191-200.

20. Schapire, RE., and Singer Y. "Improved boosting algorithms using confidence-rated predictions." *Machine learning* 37.3 (1999): 297-336.
21. Sigrist C.J.A., de Castro E., Cerutti L., Cuche B.A., Hulo N., Bridge A., Bougueleret L., Xenarios I. New and continuing developments at PROSITE. Nucleic Acids Res. 2012; doi: 10.1093/nar/gks1067.
22. Wang, C., Liu, J., Luo, F., Tan, Y., Deng, Z., & Hu, Q. N. (2014). Pairwise input neural network for target-ligand interaction prediction. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference* (67-70).
23. Weininger D (1988) SMILES 1. Introduction and encoding rules. J Chem Inf Comput Sci 28: 31-36. http://www.daylight.com.
24. Wishart, D., Knox, C, Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res, 34(Database issue):D668-72.

Cell-specific prediction and application of drug-induced gene expression profiles

Rachel Hodos[1-3], Ping Zhang[4], Hao-Chih Lee[1,2], Qiaonan Duan[5-7], Zichen Wang[5-7], Neil R. Clark[5-7], Avi Ma'ayan[5-7], Fei Wang[4,8], Brian Kidd[1,2,9], Jianying Hu[4], David Sontag[10], Joel Dudley[1,2,9*]

[1]Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai (ISMMS), New York, NY, 10065; [2]Department of Genetics and Genomic Sciences, ISMMS, New York, NY, 10029; [3]Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012; [4]IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598; [5]Dept. of Pharmacological Sciences, [6]BD2K-LINCS Data Coordination and Integration Center, and [7]Mount Sinai Center for Bioinformatics, ISMMS, New York, NY, 10029; [8]Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY, 10065; [9]Harris Center for Precision Wellness, ISMMS, New York, NY 10065; [10]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, 02139

Gene expression profiling of *in vitro* drug perturbations is useful for many biomedical discovery applications including drug repurposing and elucidation of drug mechanisms. However, limited data availability across cell types has hindered our capacity to leverage or explore the cell-specificity of these perturbations. While recent efforts have generated a large number of drug perturbation profiles across a variety of human cell types, many gaps remain in this combinatorial drug-cell space. Hence, we asked whether it is possible to fill these gaps by predicting cell-specific drug perturbation profiles using available expression data from related conditions--i.e. from other drugs and cell types. We developed a computational framework that first arranges existing profiles into a three-dimensional array (or tensor) indexed by drugs, genes, and cell types, and then uses either local (nearest-neighbors) or global (tensor completion) information to predict unmeasured profiles. We evaluate prediction accuracy using a variety of metrics, and find that the two methods have complementary performance, each superior in different regions in the drug-cell space. Predictions achieve correlations of 0.68 with true values, and maintain accurate differentially expressed genes (AUC 0.81). Finally, we demonstrate that the predicted profiles add value for making downstream associations with drug targets and therapeutic classes.

Keywords: Drug discovery, chemogenomics, tensor completion, gene expression, drug repurposing

1. Introduction

Genome-wide expression profiling of *in vitro* drug perturbations has proven to be useful for many aspects of drug discovery and development[1]. Applications include elucidation of drug mechanisms[2], lead identification[3], and drug repurposing[4, 5]. Despite this success, the capacity to leverage cell-specific responses has been hindered by limited data availability across cell types[6, 7]. To address this limitation, the Library of Integrated Cellular Signatures (LINCS) program[8, 9] has greatly expanded the publicly available data to nearly one million profiles characterizing thousands of drugs exposed to dozens of cell types. However, this combinatorial space of drugs and cell types is vast, and many gaps remain in this space (see white space in Figure 3B). These gaps present difficulties both for large-scale analysis as well as for making cell-matched comparisons, e.g. between two drugs or between drug and disease. Therefore, we asked whether it is possible to leverage existing expression profiles to predict the remaining unmeasured profiles.

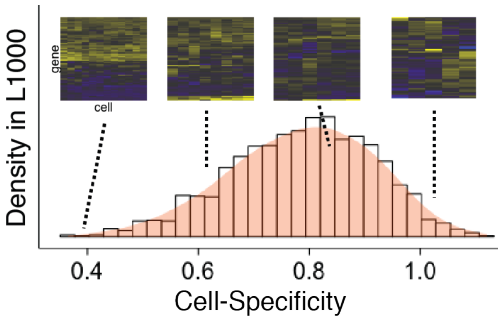


Fig. 1: Distribution of cell-specificity of 2,130 drugs in the L1000 dataset. The cell-specificity is quantified per drug as the mean pairwise cosine distance between all of its cell-specific profiles, with a range of 0 (all cells identical) to 2 (perfect anti-correlation). Four examples are shown (L to R: homoharringtonine, terfenadine, dexamethasone, and JNJ-38877605). While some drugs induce very similar expression across cell types, the majority have higher cell-specificity corresponding to distinctive patterns in different cell types.

Expression responses to drug exposure are often highly cell-specific, e.g. due to differences in expression of drug targets. Indeed, we observe a high degree of cell-specificity for many drugs in the LINCS data (see Figure 1). The utility of such cell-specific gene expression has previously been demonstrated for a variety of applications. For example, a recent analysis[10] found that LINCS expression profiles were more predictive of anti-cancer drug efficacy when using cell lines sharing a common lineage with the queried cancer type. Similarly, another study[11] showed that using transcriptional similarity to predict drug-target interactions is more accurate when comparing drug profiles in the same cell line.

Prior studies have described methods to predict expression profiles using outside information. For example, Gamazon, et al. (*12*) predict tissue-specific expression profiles from genetic variants, but are limited to heritable variation in expression. Conversely, Lagunin, et al. (*13*) predict drug-induced expression responses from a drug's chemical structure, but are agnostic to cell type. There are also many techniques to impute missing entries of a gene expression matrix, generally using either local (e.g. nearest neighbors) or global (e.g. low-rank matrix approximation) information[14, 15]. However, most of these methods are not directly applicable to our setting, as they rely on having at least some measurements available in the target experimental setting.

Here, we draw inspiration from this prior work to solve a new problem: predicting *entire* expression profiles for cell-specific drug perturbations that have not yet been measured. Our two approaches are complementary in their use of local vs. global information. The local algorithm, *Drug Neighbor Profile Prediction* (DNPP) is inspired by K-nearest neighbors but adapted to this *de novo* prediction setting. The global algorithm, *Fast Low-Rank Tensor Completion* (FaLRTC)[16] fills in the missing entries of a tensor using the observed entries. The underlying assumption here is that the data are low-rank, i.e. some small set of underlying factors (e.g. drug targets) explain most of the variation in the data.

We evaluate our methods along with two baselines using several approaches. We use cross-validation (CV) to measure correlation of true and predicted expression, as well as accuracy of differentially expressed genes (DEGs). We also study the dependence of accuracy on the amount of input data and explore the cell-specificity of our predictions. Finally, we demonstrate that the completed dataset adds value for downstream prediction of therapeutic classes and drug targets.

2. Methods

2.1. *Notation and terminology*

T refers to a tensor, with $T_{d,g,c}$ for drug d, gene g, and cell c. A colon subscript refers to all elements of that index. C_d and D_c respectively refer to the cell lines measured for drug d, and the drugs measured in cell c. Error bars in figures and text refer to ± one standard deviation. All correlations are Pearson's correlations, denoted by r or $cor(\cdot,\cdot)$. 'Drug' refers to compounds represented in the data, including approved drugs, drug-like compounds, and tool compounds.

2.2. *Data processing*

The LINCS drug expression data (herein, the "L1000 data") is measured on a targeted expression profiling platform called L1000[17]. The platform measures the expression of 978 "landmark" genes (roughly 1000, hence the name), selected to be maximally predictive of the other genes while being widely expressed across many cell and tissue types.

Differential expression computed from the level 3 L1000 data were downloaded from amp.pharm.mssm.edu/public/L1000CDS_download. The dataset was generated using the Characteristic Direction (CD) method[18] and is validated and described more fully in[19]. Briefly, a CD was calculated for each replicate using linear discriminate analysis, to find the direction in gene space that best separates cases and controls. Replicates were averaged and normalized to unit length. Average cosine distance (ACD), i.e. the mean pairwise cosine distances between an experiment's CD replicates, was used to estimate significance. The null distribution of the ACDs was calculated per batch using random sampling (n = 10,000) of replicates in the same batch. A p-value for each profile (ACD p-value) was computed by comparing its ACD to the null.

Tensor construction: The 201,484 CD profiles (20,413 drugs, 72 cell types) were filtered to 34,716 profiles (6,928 drugs, 72 cell types) with ACD $p \leq 0.1$ in order to remove the most unreliable data. Drugs and cell types with < 3 remaining experiments were removed, as well as duplicate drug id's corresponding to the same drug, for a final count of 25,672 profiles (2,130 drugs, 71 cells, 12.7% of all CDs). Profiles were averaged across all available concentration and time points, renormalized to have unit norm, and then arranged into a tensor (see Figure 2A). Of the 151,230 possible drug-cell pairs, the tensor contains 15,855, corresponding to 10.5% observation density. A smaller, more dense subset of this tensor was also used for some of the experiments, using the top 300 drugs and 15 cell lines, reaching an observation density of 71.4%. The tensor element $T_{d,g,c}$ is the g^{th} coordinate of the CD vector for drug d in cell c. All values lie in the range [-1,1] after normalization, where a positive [negative] value corresponds to up- [down-] regulation. The 10 cell lines with the most data are listed in Table 1, along with the corresponding tissue of origin and number of profiles (i.e. drugs) present. Most of the 71 cell lines are cancer cell lines, and represent a range of human tissues including skin, lung, brain, kidney, and prostate.

cell line	MCF7	VCAP	PC3	A375	A549	HA1E	HT29	HCC515	HEPG2	NPC
tissue	breast	prostate	prostate	skin	lung	kidney	colon	lung	liver	brain
# profiles	1505	1368	1340	1168	1139	1127	1022	934	798	441

Table 1. The top ten cell types in the data tensor, along with tissue of origin and number of drug profiles available.

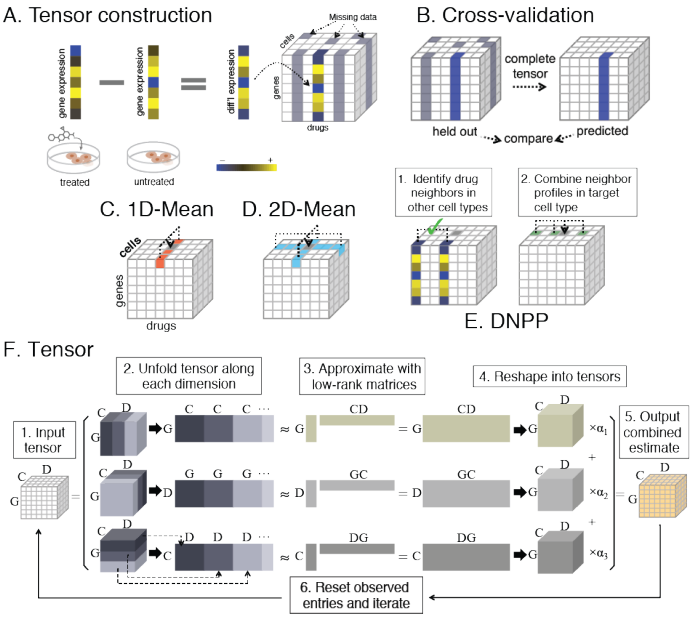


Fig. 2: Schematic overview. A. Expression profiles are compiled into a tensor of 978 genes x 2,130 drugs x 71 cell types. Profiles are either completely missing, in grey, or fully observed, denoted by both white and multicolor columns. B. CV setup, where entire profiles are held out. C-D Averaging baselines; target value is in grey and the averaged entries are colored. E. DNPP algorithm. Target value is in grey. Drug neighbors are identified by comparing profiles in other cell lines, then neighbor profiles in the target cell line are combined to form the prediction. F. FaLRTC algorithm. The data tensor is input on the left, and then unfolded in step 2 to form three matrices with dimensions G x CD (top), D x GC (middle), and C x DG (bottom). Each matrix is approximated using a spectral method and then reshaped into a tensor. The three tensors are then combined into one. Observed entries are reset to their initial values, and the process is iterated to minimize the matrix trace norms.

2.3. *The Drug Neighbor Profile Prediction algorithm*

The DNPP algorithm (Figure 2E) is an adaptation of K-nearest neighbors (KNN) to the *de novo* prediction setting. In other words, KNN normally requires at least some data present in the target condition in order to identify neighbors. To overcome this limitation, DNPP defines similarity between drugs[a] instead of profiles. The similarity (S) between two drugs d and d' is defined based on average correlation between the two drugs' profiles as measured in other cell types:

$$S(d,d') = \frac{1}{\left|C_d \cap C_{d'}\right|} \sum_{c' \in C_d \cap C_{d'}} cor(T_{d,:,c'}, T_{d',:,c'}). \quad (1)$$

DNPP then estimates the profile for drug d and cell c as a weighted average of (up to) K profiles from cell type c corresponding to neighboring drugs. To generate a prediction for (d, c), drug neighbors of d are chosen only amongst drugs that have data in cell c, and hence neighbors can differ per cell type. Finally, the weights on the K profiles are chosen proportional to $S(d,d')$, normalized to sum to 1. We use K = 10 (CV results not shown).

2.4. *The Fast, Low-Rank Tensor Completion algorithm*

Since there are many tensor completion algorithms available, we benchmarked a variety of algorithms for speed and accuracy (see supplementary for details) and subsequently selected the FaLRTC algorithm. The FaLRTC algorithm[16] is sometimes referred to herein as simply 'Tensor' or 'the tensor approach.' We briefly describe the algorithm here, in a simplified form (see Figure 2F). Like most tensor completion algorithms, FaLRTC assumes that the data has some low-rank structure. While there is a notion of rank for a tensor[20], this is in general hard to compute. Hence, FaLRTC resorts to low-rank matrix approximations instead. A three-dimensional tensor can be

[a] We also tested a similar approach defining neighbors between cell lines, but the performance was not as strong.

reshaped or 'unfolded' into matrices in three mathematically distinct ways[20], i.e. a $D\times C\times G$ tensor can be unfolded into a $D\times(CG)$, a $C\times(DG)$, and a $G\times(DC)$ matrix. The algorithm forms all three such matrices, and then performs low-rank matrix approximation via a spectral method. The prediction of missing values is based on a weighted combination of the three matrix-derived estimates, where these weights $(\alpha_i, i = 1,2,3)$ are user-defined parameters, constrained to be positive and sum to one. Observed elements are reset to their true values, and then this process is iterated using gradient descent to minimize (an upper bound on) the matrix ranks. Due to the column-structured pattern of missing entries in our tensors, gene correlation structure is less useful for predictions than correlations in the other two dimensions, and hence estimates from the matrix $(G\times(DC))$ that most strongly leverage gene correlations, were down-weighted by a factor of 100 relative to the other two (i.e. $\alpha_2 \equiv \alpha_1/100 \equiv \alpha_1/100$). This can be seen as an adaptation of the algorithm to the present setting defined by the column-structured pattern of missing entries.

2.5. *Baseline averaging schemes*

While many methods exist to impute randomly missing entries in a gene expression matrix, we are not aware of prior work predicting entire expression profiles without additional data inputs. Thus we use two simple baselines that make predictions by averaging relevant subsets of data. *1D-Mean* (Figure 2C) predicts missing expression profiles for each drug by averaging all profiles available for that drug in the tensor (i.e. across cell lines). *2D-Mean* (Figure 2D), combines the 1D-Mean average across cell lines with a similar average in the other dimension across drugs, i.e.

$$2DMean(d,g,c;\lambda) = \lambda\frac{1}{\left|D_c\right|}\sum_{d'\in D_c} T_{d',g,c} + (1-\lambda)\frac{1}{\left|C_d\right|}\sum_{c'\in C_d} T_{d,g,c'}, \quad (2)$$

We use λ = ½ based on CV experiments (results not shown).

2.6. *Cross-validation for predicting gene expression profiles*

10-fold CV experiments were performed, where *entire expression profiles* were held out and then predicted (see Figure 2B), randomly selecting 10% of the profiles per fold. All of these predictions were compiled into a tensor, $\hat{T}$, with the same dimensions and pattern of missing entries as the original tensor. Accuracy was measured as the Pearson correlation with truth (PCT). This is defined simply as $\mathrm{PCT}_\Omega = cor\left(T_\Omega, \hat{T}_\Omega\right)$, where Ω corresponds to some subset of the tensor, the correlation is taken element-wise, and missing entries are ignored. For example, Ω might correspond to an individual drug-cell profile, a CV fold, or the entire tensor.

2.7. *Predicting drug targets and ATC codes*

In order to build binary classifiers of drug-target interactions and Anatomic Therapeutic Chemical (ATC) classifications, drug profiles were compiled for all drugs represented in the data tensor, restricting to the top ten most-sampled cell lines (see Table 1). Measured profiles were used as is, and predicted profiles were generated using the DNPP method. The drug profiles and corresponding binary labels were used to train KNN, Random Forests (RF), and Regularized

Logistic Regression (LR) models via the *caret* package[21]. For each experiment (i.e. one profile type, prediction task, model, and choice of either measured or completed dataset; see Figure 5A) a grid search was performed using 10-fold CV to select model hyperparameters (see supplementary). The cross-validated predicted probabilities from the selected set of parameters were recorded and then used to compute several versions of AUC scores. In the first set of experiments, AUCs are compared between (a) classifiers trained on the completed data, versus (b) the same classifiers trained on only the measured subset of profiles. Here, AUCs are calculated on the common set of labels corresponding to the measured drug profiles only, and results were excluded from the analysis when both AUCs were < 0.5. In the second set of experiments, AUCs were computed on two complementary sets of predictions from the same model trained on the completed data, where the complementary sets are the drugs with measured profiles, vs. the set of drugs for which only predicted profiles are available. Here, experiments were again excluded if both AUCs were < 0.5, or if either drug set (for the measured or predicted profile sets) had < 3 positive examples.

3. Results

3.1. *Overall accuracy*

We start with an evaluation of the overall correlation between true and predicted values. Figure 3A shows a smoothed scatterplot of all Tensor (FaLRTC) predictions versus true values, where each point corresponds to a single, numeric entry in the tensor. The four methods achieved correlations (i.e. PCT, see Methods) of 0.53, 0.54, 0.46, and 0.40[b].

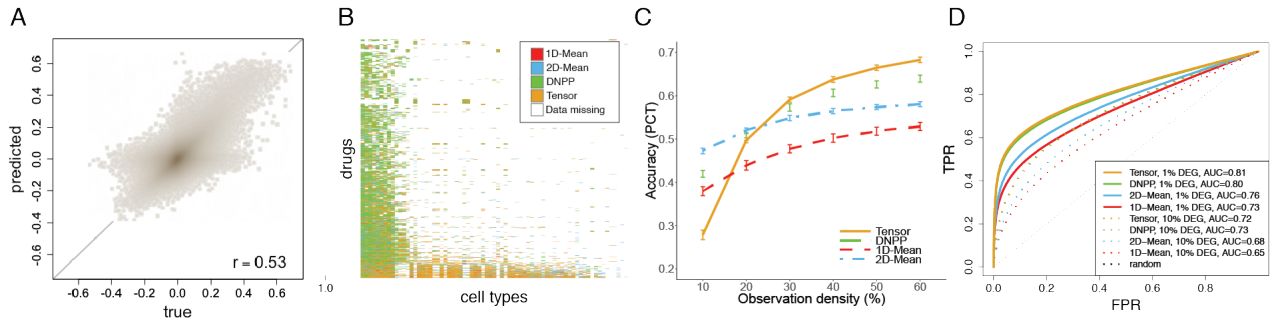


Fig. 3. Prediction accuracy. A. Scatterplot of Tensor-predicted vs. true values. B. Top-performing method per drug-cell profile in the tensor. C. Accuracy vs. observation density, where lower densities correspond to entire profiles being held out of the small tensor. D. ROC curves assessing prediction of DEGs. See text for details.

3.2. *Tradeoffs in accuracy across drug-cell space*

While DNPP and Tensor have similar overall performance, we observe a clear tradeoff in accuracy between the two methods across different regions of the space. Figure 3B shows which method was most accurate (based on PCT) for each profile in the tensor. We see that for drugs

[b] All results are reported in the following order: Tensor; DNPP; 2D-Mean; 1D-Mean.

with profiles in many cell lines (i.e. near the bottom), the tensor approach is usually the top performer, while in the region on the left where fewer cell lines but many drugs have been profiled, DNPP is generally superior.

3.3. *Effects of varying observation density*

Next, we studied the dependence of accuracy on the amount of input data by varying the percent of observed profiles in the small (and more dense) tensor. Observation density was varied by subsampling profiles in the tensor in 10% intervals from 10-60%, evaluating on a held-out set covering another 10% of the tensor. This sampling process was repeated 25 times generating the error bars in Figure 3C. At or above an observation density of 30%, Tensor had superior performance, while at lower densities, 2D-Mean was the top performer. We also observe that the tensor approach had a more dramatic improvement in performance with increasing density, reaching a mean PCT per fold of 0.68.

3.4. *Accuracy of differentially expressed genes*

We also evaluated the ability to predict DEGs in the unmeasured drug-cell experiments. To do this, we first identified DEGs in the measured profiles, and then thresholded expression values in the corresponding predicted profiles to generate ROC curves. More specifically, for each expression profile, a gene was considered a "true" DEG if its absolute expression value was at or above the p^{th} percentile relative to all genes in the profile, where p was set to either 1% or 10% (other approaches were also tried with little effect on the outcome; results not shown). ROC curves shown in Figure 3D were then generated by varying an analogous percentile threshold across the range 0-100% for the predicted profiles in the CV tensors, thereby defining a set of predicted DEGs for each profile at each possible threshold value. Each ROC curve represents aggregate results across all profiles in the tensor. The methods achieved area under the ROC curve (AUC) values of 0.81, 0.80, 0.76, and 0.73 at p = 1%. At p = 10%, a similar relationship between methods was observed (0.72, 0.73, 0.68, 0.65). For all four methods, AUC's were higher at the 1% threshold relative to the 10% threshold, and this pattern was observed more generally (results not shown), where smaller values of p correspond with higher accuracy. This is reasonable in that smaller percentile thresholds correspond to genes with stronger differential expression signals.

3.5. *Analysis of cell-specificity*

While some L1000 drugs show very similar responses across cell types, others induce highly cell-specific responses. One such example is M-3M3FBS (herein "M3"), a PLC agonist that induces a variety of effects ranging from modulation of neutrophil function to apoptosis. The tensor contains M3 profiles in 15 different cell lines, shown on the left-most panel of Figure 4A. Responses cluster into two primary groups, with one group (on the left) enriched for down-regulation of both spindle pole genes as well as valine, leucine, and isoleucine degradation, perhaps indicating a pre-apoptotic response. The mean profile of the second group (A549, AGS, RKO, and MCF7 cells) is enriched for very different types of processes including up-regulation of Akt signaling, insulin signaling, and salivary secretion, all of which have established connections to PLC[22, 23]. Figure 4A

shows that the tensor approach was able to accurately recapitulate these two classes of responses. DNPP, on the other hand, seems to "misclassify" some of the cell types into the wrong group, while 1D-Mean and 2D-Mean predictions are nearly identical across cell types.

Another example (Figure 4B) with highly cell-specific expression patterns is Carbetocin, an oxytocin analog. In contrast to the previous example, here DNPP outperforms the tensor approach. One explanation for DNPP's success with Carbetocin is that all three measured cell lines (MCF7, A549 and VCAP) are among the top five most-sampled cell lines in the tensor, and therefore have many drug neighbors from which to choose. On the other hand, M3 has data in many cell types, which is associated with better Tensor predictions. In addition to M3 and Carbetocin, two more examples are presented in the supplementary information, one (ABT-751) in which both methods do similarly well, and a second (GNF-2), where both have similarly poor performance.

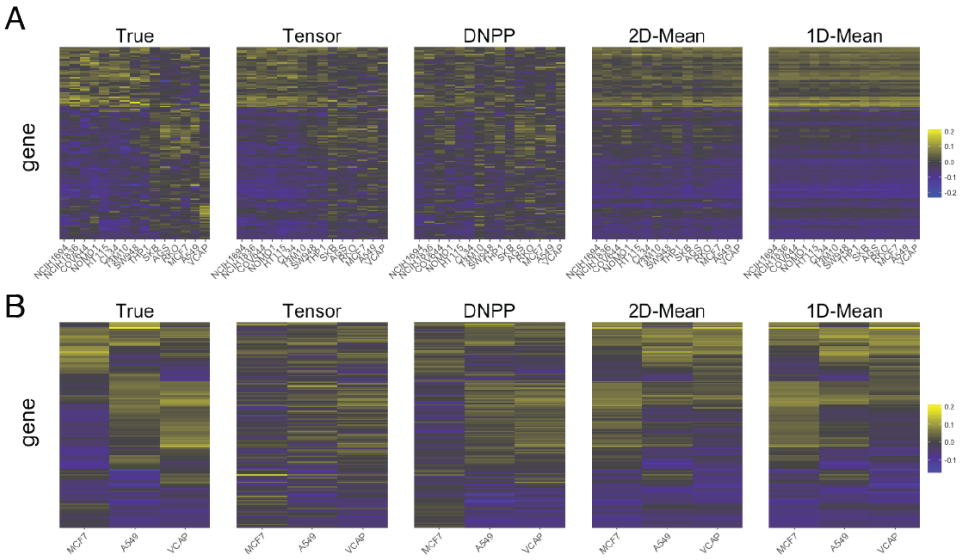


Fig. 4: Cell-specificity of predictions. A. True and predicted expression profiles for the compound M-3M3FBS (see text for details). Rows correspond to genes, and columns to cells. B. Analogous plots for Carbetocin, in the three available cell lines, MCF7, A549, and VCAP. See text for details.

3.6. *Utility of completed data for downstream prediction of drug properties*

In this final section, we aim to show that the completed data provides added value for downstream prediction of drug targets and therapeutic classes. To do this, we trained binary classifiers using the drug profiles as inputs, and designed experiments to address two questions (see Figure 5A). First, we asked whether classifiers trained on the completed data are of higher quality than those trained on only the measured subset of profiles. Second, we asked whether ATC and target predictions have comparable accuracy on measured vs. predicted profiles. Toward both of these aims, we identified the top 7 drug targets and 3 ATC classes (see Figure 5C) represented in the tensor, and trained classifiers for each of these tasks using 12 different versions of input drug profiles (cell-specific profiles from the top 10 most-represented cell lines in the tensor, as well as the mean and maximum value of each gene across these 10 cell lines). Finally, since our questions are focused on the value of the drug profiles and not about a specific algorithm, we included three different algorithms in our experiments (LR, KNN, and RF).

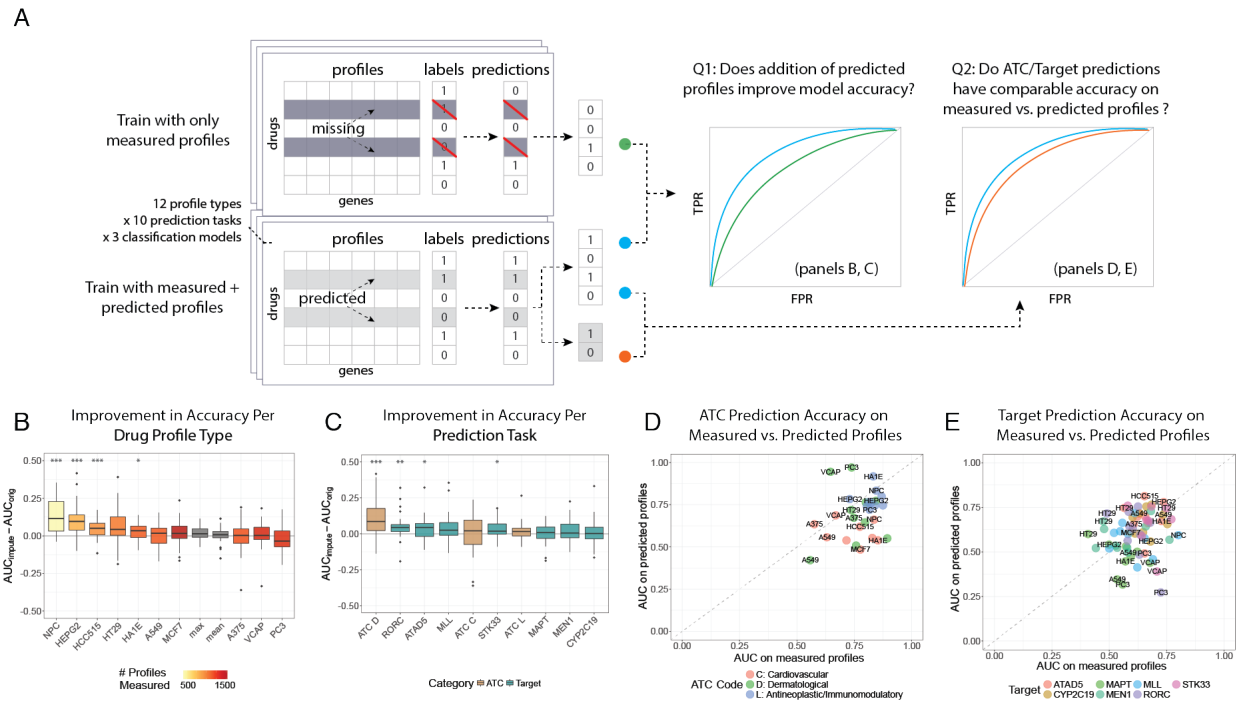


Fig. 5: Utility of completed data for downstream predictions. A. Illustration of experimental setup. Binary classification models are trained to predict drug targets and ATC codes, using either the measured subset of profiles (e.g. for a particular cell type) or the completed data, and cross-validated prediction scores are recorded. Then two types of ROC curve comparisons are made, as described in the text. B. Improvements in AUC per drug profile type, across experiments for different prediction tasks and models. C. Improvements in AUC per prediction task, across different profile types and models. D. ATC prediction accuracy on measured vs. predicted profiles, for different profile types. E. Target prediction accuracy on measured vs. predicted profiles. For both D and E, median values across models were computed to simplify the plots, but were kept distinct for all reported results.

The results addressing the first question were generally positive. More specifically, of the 360 experiments (12 profile types x 3 models x 10 prediction tasks), after removing 21 experiments where no signal could be found, 223 (65.6%) showed an increase in AUC when training on the completed data compared with only the measured subset, with a mean improvement of 0.03 ($p <$ 1e-8, paired *t*-test). Differences were also significant ($p \leq 0.01$) for each of the models individually, with mean AUC improvements of 0.05 for LR and 0.02 for RF and KNN. The improvements also varied by profile type, as shown in Figure 5B. More specifically, we observed that cell types such as NPC (neural progenitor cells) that had fewer measured profiles available saw the most gains when including the additional profiles. Overall, four profile types (NPC, HEPG2, HCC515, and HA1E cell-specific profiles) showed significant AUC improvements across models and prediction tasks (adjusted $p < 0.05$, paired *t*-test), with two additional profile types (HT29 and *max*) reaching marginal significance, and none showing significant decreases. Figure 5C shows a similar analysis per prediction task. The median AUC difference was positive for all prediction tasks, reaching statistical significance for 4 out of 10: ATC *D* code (dermatological indications), and RORC, STK33, and ATAD5 targets, with MLL reaching marginal significance.

Figures 5D and E summarize the second set of experiments addressing the question of whether accuracy is comparable on predicted vs. measured profiles. While one might expect that accuracy

would always be worse on the predicted profiles, this is not the case. We find instead that the results are mixed, and vary per feature and outcome. For example, predicting the ATC L code (antineoplastic and immunomodulating agents), had similarly high accuracy using either measured or predicted profiles, likely due to strong expression signals for this class of drugs, as well as high relevance of the cancer cell lines for observing antineoplastic effects. However, across experiments for the ATC codes, there was a mean loss of 0.08 AUC using the predicted profiles. On the other hand, in the case of target prediction, there was no significant loss of AUC across experiments. Interestingly, the predicted HT29 profiles had better accuracy than measured profiles for 19 of the 21 target prediction experiments (mean AUC improvement 0.12), perhaps indicating de-noising in the predicted profiles. Additionally, we found that for all 10 tasks, there were multiple profile types for which the AUC was higher on the predicted profiles.

4. Discussion

Expression profiles characterizing *in vitro* drug perturbations are useful for a variety of applications in drug discovery. While many thousands of such expression profiles have been measured, large gaps remain in the combinatorial space across drugs and cell types. Hence, we asked whether it is possible to leverage existing data from other drug-cell combinations to predict unmeasured profiles. We tested both local and global approaches, finding that predictions are not only accurate in an overall sense but preserve signal that is biologically and therapeutically relevant, e.g. maintaining accurate DEGs and signal to predict targets and therapeutic classes.

Both Tensor and DNPP almost uniformly outperformed the averaging baselines, with highly complementary performance between the two methods (Figure 3B). This complementarity is concordant with intuition in that, the global approach can leverage all information available and hence outperforms when a large amount of information is available per drug; whereas the local approach has better performance when many drugs neighbors are available. In addition to this complementarity, there are other tradeoffs. On the one hand, Tensor was able to "learn more" than DNPP with increasing observation density (Figure 3D). On the other hand, DNPP is conceptually simpler, uses only a single parameter, and requires less computation time.

In our experiments with ATC and target prediction, we note that the purpose is not to demonstrate state-of-the-art accuracy, but to show that the completed data adds predictive value to the LINCS L1000 drug profiles. Indeed, we observed many cases showing significantly improved accuracy, with no cases of significant decreases in accuracy. These results are likely explained by several factors. For cell-specific profiles, the completed data contains more profiles, and hence models can be trained with more labels. For the *max* and *mean* profiles, the incomplete data has heterogeneous cell-type availability per drug whereas the completed data is summarized across a uniform set of cell lines. Additionally, it is possible that the predicted profiles may, in some cases, have a stronger signal-to-noise ratio than their measured counterparts, which could explain, e.g. the high performance of the predicted HT29 profiles (Figure 5E) in multiple prediction tasks.

Our framework produces testable and usable predictions at the L1000 profile level. More specifically, each value corresponds to the differential expression (CD) value of one gene in one cell line perturbed by one drug. However, the CD values do not map directly to measurable gene-

level quantities such as fold change. Therefore, we advise that, unless one compares predictions to the result of a CD analysis, predictions should either be treated at the level of a ranked list of genes, or thresholded to define DEGs.

We noticed while processing the L1000 data that roughly 2/3 of the > 20K drugs did not have any experiments with reliable (i.e. nominal ACD $p < 0.1$) measurements between replicates. While replicate consistency may improve with advances in data processing, it is likely that many of the drugs simply do not induce a strong enough expression response to overcome biological and technical sources of noise. We believe that this should be taken into consideration for any project working with L1000 drug profiles.

One limitation of this study is the lack of established baselines. The baselines used in this study were relatively basic, but help to demonstrate that our predictions outperform alternatives that might be considered safe and intuitive. While few methods currently exist for systematic prediction of cell-type specific drug expression profiles, we expect that the methods and results presented in this study would serve as useful baselines for future work on improved methods.

Several factors may have introduced bias into the results. First, almost all of the cell lines are cancer lineages, which may result in more similarity between cell lines than otherwise expected. Second, the selection of landmark genes may have biased the results. One line of thinking is that, due to the way these genes were selected, one would expect them to be relatively independent and therefore more difficult to predict than a random set of genes. If true, this would bias the results in a more conservative direction. Third, the presence of chemically similar drugs in the tensor could potentially make the prediction problem easier than otherwise. However, our analysis indicates that this bias is quite small (< 0.02 PCT difference), and we also verified that none of the drugs highlighted in Section 3.5 have structural cognates in the tensor (i.e. all Tanimoto coefficients are less than 0.5). Fourth, our CV experiments reduced observation density by 10%, and hence results would likely be further improved by using all available data. Finally, the L1000 data has highly imbalanced sampling across the drug-cell space (see Figure 3B), and this is likely a source of positive bias. Predictions made in the less-dense regions of the drug-cell space should therefore be used with caution and would likely benefit the most from methodological improvements.

There are many directions to explore in future work, grouped into a few categories. First, the data inputs could be expanded in a variety of ways. E.g., one could use the full, imputed transcriptome as opposed to only landmark genes. Also, more inclusive data filtering could be evaluated. The Broad Institute is also continuing to generate more data across this space; however, this will likely never be comprehensive, and hence we expect that this work will continue to be relevant. The second category of extensions are methodological, including: 1) nonlinear modeling; 2) use of auxiliary similarity information[24]; 3) addition of a time dimension to the tensor; 4) modeling measurement reliability; and 5) adopting a probabilistic framework. The final category of future work relates to applications. First, our approach could readily be applied and evaluated on many other biological datasets where data span at least three categorical axes. Such datasets include CMap[25], with dimensions of drugs, genes, and cell types, and the Genotype-Tissue Expression (GTEx)[26] and Braineac datasets[27], each spanning individuals, genes, and tissues. Second, one could extend this framework to be able to prioritize the remaining experiments, e.g. using active learning, in order to optimally map out this transcriptional landscape across drugs and

cellular contexts. Finally, another exciting direction would be to make possible 'out-of-sample' predictions[28] which would be particularly useful when measurements are difficult to obtain (e.g. for human *in vivo* brain tissue expression) but where related measurements could be obtained from more accessible tissues (e.g., neuronal cell types from induced pluripotent stem cells). This would likely require an integrative approach leveraging additional datasets and metrics[24] (e.g., cell line genetic similarity as auxiliary data for tensor completion).

To the best of our knowledge, this work is the first attempt at prediction of expression profiles using only expression from related experimental conditions. Hence, we consider this work to be a compelling proof-of-concept demonstrating the feasibility and value of such predictions. It is our hope that completing the space across drugs and cell types will enable new types of analyses and predictions of cell-specific drug action that could lead to translational insights and applications.

Supplementary Information

Supplementary information including figures, code, and data can be found at goo.gl/nTy8sH.

Funding

This work was supported by the following grants from the NIH: Illuminating the Druggable Genome (IDG) sponsored by NIH Common Fund and NCI [U54CA189201]; NIDDK [R01DK098242] and NCATS [UL1TR000067] Clinical and Translational Science Award to RH, BK, HCL, and JD. Additional grants from the NIH [R01GM098316, U54HG008230 and U54CA189201] supported QD, ZW, AM, and NC, and NSF Career Award [1350965] to DS.

References

1. X. A. Qu *et al.*, *Drug discovery today* **17**, 1289-1298 (2012).
2. G. Wei *et al.*, *Cancer cell* **10**, 331-342 (2006).
3. D. C. Hassane *et al.*, *Blood* **111**, 5654-5662 (2008).
4. J. T. Dudley *et al.*, *Briefings in bioinformatics*, bbr013 (2011).
5. G. Hu *et al.*, *PloS one* **4**, e6536 (2009).
6. S. A. Khan *et al.*, *Bioinformatics* **30**, i497-i504 (2014).
7. J. A. Parkkinen *et al.*, *BMC bioinformatics* **15**, 113 (2014).
8. Q. Duan *et al.*, *Nucleic acids research*, gku476 (2014).
9. B. I. NIH. (2014).
10. B. Chen *et al.*, *Nature Communications* **8**, (2017).
11. M. Iwata *et al.*, *Scientific Reports* **7**, 40164 (2017).
12. E. R. Gamazon *et al.*, *Nature genetics* **47**, 1091-1098 (2015).
13. A. Lagunin *et al.*, *Bioinformatics*, btt322 (2013).
14. O. Troyanskaya *et al.*, *Bioinformatics* **17**, 520-525 (2001).
15. G. N. Brock *et al.*, *BMC bioinformatics* **9**, 1 (2008).
16. J. Liu *et al.*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35**, 208-220 (2013).
17. A. Subramanian *et al.*, *bioRxiv*, 136168 (2017).
18. N. R. Clark *et al.*, *BMC bioinformatics* **15**, 79 (2014).
19. Q. Duan *et al.*, *NPJ Systems Biology and Applications* **2**, 16015 (2016).
20. T. G. Kolda *et al.*, *SIAM review* **51**, 455-500 (2009).
21. M. Kuhn, *Journal of Statistical Software* **28**, 1-26 (2008).
22. A. Parrales *et al.*, *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1813**, 1758-1766 (2011).
23. J. Eichhorn *et al.*, *Endocrinology* **143**, 655-664 (2002).
24. A. Narita *et al.*, *Data Mining and Knowledge Discovery* **25**, 298-324 (2012).
25. J. Lamb *et al.*, *Science* **313**, 1929-1935 (2006).
26. M. Melé *et al.*, *Science* **348**, 660-665 (2015).
27. D. Glass *et al.*, *Genome Biol* **14**, R75 (2013).
28. J. Wang *et al.*, *The American Journal of Human Genetics* **98**, 697-708 (2016).

Large-scale integration of heterogeneous pharmacogenomic data for identifying drug mechanism of action

Yunan Luo, Sheng Wang, Jinfeng Xiao and Jian Peng*

Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, Illinois 61801, USA
**Corresponding author: jianpeng@illinois.edu*

A variety of large-scale pharmacogenomic data, such as perturbation experiments and sensitivity profiles, enable the systematical identification of drug mechanism of actions (MoAs), which is a crucial task in the era of precision medicine. However, integrating these complementary pharmacogenomic datasets is inherently challenging due to the wild heterogeneity, high-dimensionality and noisy nature of these datasets. In this work, we develop Mania, a novel method for the scalable integration of large-scale pharmacogenomic data. Mania first constructs a drug-drug similarity network through integrating multiple heterogeneous data sources, including drug sensitivity, drug chemical structure, and perturbation assays. It then learns a compact vector representation for each drug to simultaneously encode its structural and pharmacogenomic properties. Extensive experiments demonstrate that Mania achieves substantially improved performance in both MoAs and targets prediction, compared to predictions based on individual data sources as well as a state-of-the-art integrative method. Moreover, Mania identifies drugs that target frequently mutated cancer genes, which provides novel insights into drug repurposing.

Keywords: data integration, drug mechanisms of action, drug target, drug similarity network, dimensionality reduction

1. Introduction

Accurate identification drug mechanism of actions (MoAs) and drug targets is of great importance for developing new drug as well as repurposing existing drugs. During the past decades, many computational approaches have been developed to identify drug MoAs and targets according to molecular docking analysis,[1] annotated target profiles,[2] adverse drug reactions,[3] and scientific literature.[4] However, these methods were limited to the prediction for drugs that are well-studied either in literature or existing biological experiment assays. Consequently, computational approaches that can be generalized to all drugs are a pressing need in the field.

Fortunately, with the recent advances in sequencing technology, large-scale pharmacogenomic data offers us exciting opportunities to systematically identify drug MoAs and targets. For example, chemical structure has been used to predict drug-target interaction.[5,6] The motivation behind this is that drugs that are structurally similar tend to interact with similar genes, thus sharing similar MoAs. Another notable dataset, drug perturbation data has also been widely used to identify MoAs.[7] Drug perturbation data, such as Connectivity Map (CMap) Library[8] and the L1000 dataset from the Library of Integrated Network-based Cellular Signatures (LINCS),[9] reveals drug-induced transcriptional profiles. It measures the gene expression

change in the presence of a drug and these gene signatures enable the comparison between drugs. Moreover, high-throughput *in vitro* drug screening over large panels of tumor cell lines have been shown to be useful in identifying clinically relevant drugs. For example, the recent developed Cancer Therapeutics Response Portal (CTRP) project[10] contains the drug sensitivity profiles of 481 small-molecule compounds across 860 cancer cell lines, which provides additional insights into the MoA of small-molecule compounds and novel therapeutic hypotheses. Since drugs with the same MoAs tend to exhibit similar transcriptional and cellular responses, these more accessible pharmacogenomic collections can be used to systematically infer drug MoAs and targets.[7]

Intuitively, integrating these datasets can further improve the identification of drug MoAs and targets. However, the sheer amount and heterogeneity of these multi-omics data pose great challenges in the integration process: (i) the mixed formats, scales, and metrics, (ii) the complementary but high-dimensional information, and (iii) the incomplete and noisy nature of these datasets. As far as we know, Drug Network Fusion (DNF) [11] was the only previous attempt to simultaneously integrate the drug structure, perturbation and sensitivity data. Notably, DNF used a similarity network fusion approach,[12] in which a similarity network is constructed for each input data sources, and these similarity networks are then iteratively fused together until convergence to obtain a single similarity network. The major drawback of this approach is that the context-specific similarity measures were mixed together in the collapsed single network, where the context-specific information may be lost or obscured.

In this work, we introduce Mania (prediction of mechanism of action by network integration), a novel method for characterizing drug-drug relationships and predicting drug mechanism of actions (MoAs) and drug targets through integrating multiple large-scale pharmacogenomic data, including drug structure, sensitivity, and perturbation data. Mania takes full advantage of the fine-grained inherent structure in the individual data source and integrates heterogeneous information by learning low-dimensional vector representations for drugs, which best explain the relationships among drug across all pharmacogenomic data. We demonstrate that, unlike DNF which directly produces a drug-drug similarity matrix, Mania is a versatile method in that the low-dimensional vector representations of drugs not only capture more accurate similarity measure with any type of distance metric, but can also be used as plug-in feature vectors of many off-the-shelf machine learning algorithms for the prediction of drug MoAs and targets. Experiment results suggested that Mania outperforms DNF, the state-of-the-art method, with substantial improvements in MoAs/targets prediction. In addition, based on the low-dimensional vector representations of drugs, Mania consistently identified functionally-enriched drug clusters, in which drugs within the same cluster are interacting with same targets. Moreover, we show that Mania found new drugs that may target significantly mutated cancer genes, which provides potential insights into drug repurposing. Overall, our experiment results suggested the superior ability of Mania in integrating multiple pharmacogenomic data for drug MoAs and targets prediction, and also demonstrated its potential as a practical tool to support network pharmacology.

2. Materials and Methods

We first provide an overview of Mania (Fig. 1). Taking one or more types of drug-related data for the same set of drugs as input, Mania first constructs a similarity network for each type of data source separately. It then integrates these heterogeneous similarity networks by combining a network diffusion algorithm and a dimensionality reduction scheme to learn a low-dimensional vector representation for each drug. These vector representations of drugs simultaneously capture the complementary information from different data sources. Intuitively, the vector representations of two drugs will be co-localized in the low-dimensional space if they are structurally similar or functionally correlated, e.g., share common chemical structure features, have similar sensitivity profiles, and/or perturb the same set of genes. Finally, Mania constructs the integrated drug-drug similarity network and infers MoAs and targets based on the low-dimensional vector representations.

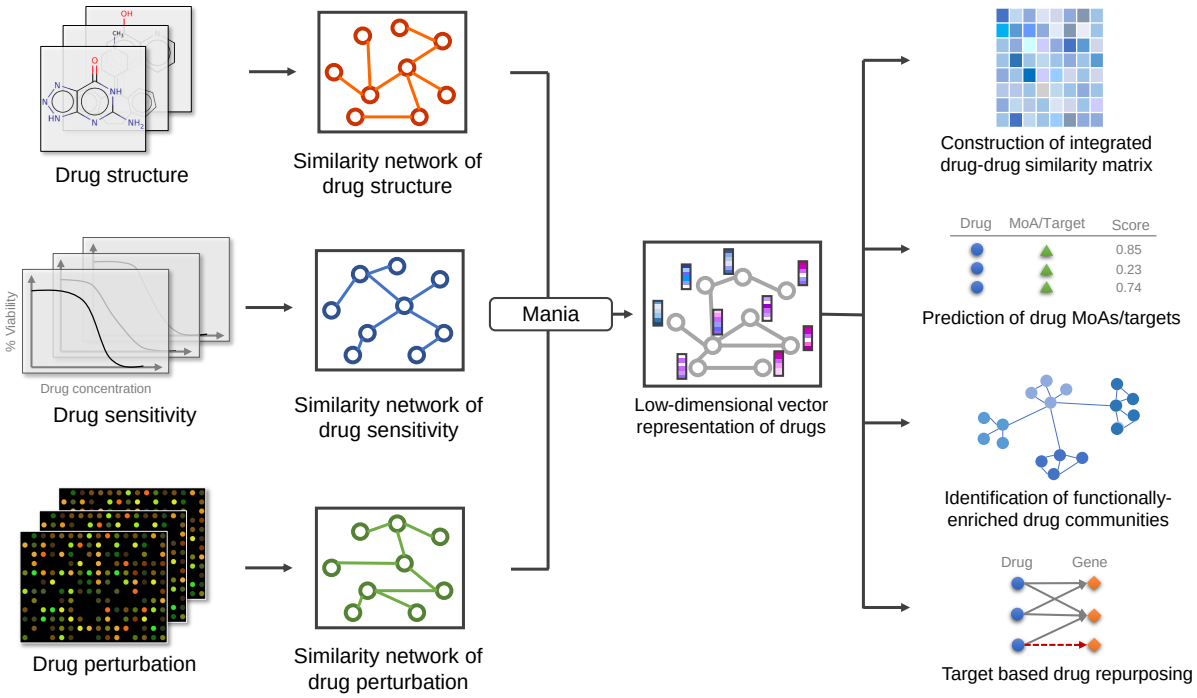


Fig. 1. **Schematic illustration of Mania.** Mania integrates multiple pharmacogenomic data sources and represents each drug with a low-dimensional vector representation. Mania can be used in a variety of tasks, including constructing integrated drug-drug similarity network, predicting MoAs and targets for drugs, identifying drug communities that share common MoAs/targets, and target based drug repurposing.

2.1. *Construction of heterogeneous drug-drug similarity networks*

In this work, we integrate three different types of drug-related data, including drug structure, sensitivity, and perturbation data. Each type of data is presented in heterogeneous formats (e.g., sequences that represent the chemical structures of drugs, and matrices that represent

gene expression in response to different concentrations of a drug), and characterizes the properties of drugs from various aspects. To extract the information of drug-drug relationships encoded in the heterogeneous data, we construct a similarity network for each type of the data sources.

Drug perturbation. We obtained the drug perturbation data from the L1000 dataset[9] from the Integrated Network-Based Cellular Signatures (LINCS) Program (`http://www.lincsproject.org/`). The L1000 dataset produced over one million gene expression profiles of 1,000 landmark genes in response to the treatment of 20,413 unique compounds across many cancer cell lines, subject to various perturbation conditions. We used the PhamacoGx package[13] to download the transcriptional profiles, and compute a "signature" for each drug that quantifies the effect of drug concentration on the gene expression with a linear regression model. To characterize the relationships between a pair of drugs based on whether they perturb the same set of genes with similar patterns, we compute the pairwise drug similarity using the Pearson correlation between their drug perturbation signatures.

Drug structure. We collected the canonical SMILES strings for the small molecules in the L1000 dataset from the PubChem database.[14] We used the RDKit[15] library to parse the SMILES strings, generate fingerprints, and compute structure similarity. We generated the Morgan fingerprint[16] (also known as circular fingerprints) with radius 2 for each drug, which takes into account both the atomic properties and the neighborhood information of each atom. Given the fingerprints of a pair of drugs, the structure similarity between them was then calculated using the Dice coefficient,[17] which is a real value in $[0, 1]$ that measures the extent to which pairs of drugs share similar structure features.

Drug sensitivity. We used the drug sensitivity data released in a recent work,[18] which is also available at the Cancer Cancer Therapeutics Response Portal (`http://www.broadinstitute.org/ctrp/`). The dataset contains sensitivity patterns for 481 compounds (including FDA-approved drugs and clinical candidates) spanning 842 different human cancer cell lines encompassing 25 lineages. We extracted the area under curve (AUC) of the concentration-response curve as the metric of sensitivity, which measures the cellular response to individual compound. To quantify the relationships between a pair of drugs based on whether they cause similar responses to same cancer cell lines, we calculate the pairwise drug similarity using the Pearson correlation between their drug sensitivity profiles.

Drug MoAs and targets. The recently released Drug Repurposing Hub[19] is a repository that contains a drug screening collection of 4,707 compounds with extensively curated annotations (e.g., mechanism of action, target, SMILES string, drug indication, and disease area) for each drug. Drug Mechanism of actions and targets were exported from the repository and used as the ground truth in the experiments of this work.

Intersection of drugs in multi-omic data. Overlapping the drugs in the drug perturbation data (20,413 drugs) and sensitivity data (481 drugs), we obtained a common set of 277 drugs that are shared in the two types of data. The drug structures of these 277 drugs were collected from the PubChem database. Each of the 277 drugs was then searched in the Drug Repurposing Hub, and 170 of them were found to have annotated MoA and target information. The MoAs and targets of these drugs were extracted for evaluation in our experiments.

2.2. *Integration of multi-omics data*

Multi-omics data provide drug-related information from diverse data sources and integration methods can shed light on the properties of less-characterized drugs. Here, we used our recently developed network integration algorithm Mashup[20,21] to integrate three types of multi-omics data, including drug structure, drug sensitivity and drug perturbation profiles. Mashup has been demonstrated to achieve significantly improved prediction for protein function prediction, gene ontology reconstruction, genetic interaction prediction, and drug-target interaction prediction.[20–23] It takes one or more networks as input, performs random walk with restart (RWR)[24] and extracts topological information from the diffusion distributions using informative but low-dimensional vector representations of drugs.

Formally, let $\mathbf{A}$ denote the weighted adjacency matrix of a certain type of similarity network of n drugs (for example, let $A_{i,j}$ be the chemical structure similarity between drugs i and j). The transition matrix of the RWR can then be calculated as $\mathbf{B}_{i,j} = \mathbf{A}_{i,j} / \sum_{j'} \mathbf{A}_{i,j'}$. Let $\mathbf{s}_i^t$ be an n-dimensional distribution vector in which each element stores the probability of a node being visited from node i after t iterations of the random walk, the RWR process is then defines as $\mathbf{s}_i^{t+1} = (1 - p_r)\mathbf{s}_i^t\mathbf{B} + p_r\mathbf{e}_i$, where $\mathbf{e}_i$ stands for an n-dimensional vector with $\mathbf{e}_i(i) = 1$ and $\mathbf{e}_i(j) = 0$, $\forall j \neq i$, and p_r is the restart probability controlling the relative influence between local and global topological information in the diffusion process. At this fixed point of the RWR, we can obtain the "diffusion state" $\mathbf{s}_i^\infty$ for drug i (i.e., $\mathbf{s}_i = \mathbf{s}_i^\infty$), in which the jth element $\mathbf{s}_{ij}$ of the diffusion state stores the probability of RWR starting node i and ending up at node j in equilibrium.

The diffusion states resulting from the aforementioned RWR process may not be entirely accurate, partially due to the low-quality and high-dimensionality of biological data. One of the strengths of Mashup is that it teases functionally relevant topological patterns apart from noise in the diffusion states and jointly integrates heterogeneous information from L similarity networks by learning low-dimensional vector representations of drugs. With the goal of denoise and dimensionality reduction, Mashup approximates each the diffusion state $\mathbf{s}_i^{(l)}$ of drug i in network l with a multinomial logistic model parameterized by low-dimensional feature vectors:

$$\hat{\mathbf{s}}_{ij}^{(l)} = \frac{\exp\left(\mathbf{x}_j^T \mathbf{w}_i^{(l)}\right)}{\sum_{j'} \exp\left(\mathbf{x}_{j'}^T \mathbf{w}_i^{(l)}\right)}, \tag{1}$$

where $\forall i$, $\mathbf{w}_i^{(l)}, \mathbf{x}_i \in \mathbb{R}^d$ for $d \ll n$. For drug i, We refer to $\mathbf{w}_i^{(l)}$ as the *context feature* which is network-specific for network l, and $\mathbf{x}_i$ as the *node feature* which is shared globally across all networks. Finally, Mashup uses the Kullback-Leibler (KL) divergence to guide the learning of the two low-dimensional vectors,

$$\min_{\mathbf{w},\mathbf{x}} C(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} D_{KL}\left(\mathbf{s}_i^{(l)} \parallel \hat{\mathbf{s}}_i^{(l)}\right). \tag{2}$$

The vectors $\{\mathbf{x}_i\}$ are subsequently used as the low-dimensional vector representations of drugs. If two drugs have similar vector representations, it generally implies that they have similar positions with respect to other drugs in the network, and thus probably share similar functions.

2.3. *Prediction of MoAs and drug targets*

To predict the MoAs and drug targets, Mania first identifies similar drugs based on the low-dimensional vector representations of drugs. In the experiments throughout this work, we used the cosine distance between the feature vectors as the distance metric for a pair of drugs i and j, following the previous work:[21]

$$D_{cos}(i,j) = 1 - \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}, \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of drugs i and j, respectively.

After computing the distances, Mania is able to predict the MoAs and targets for drugs that are less well-characterized using a k-nearest neighbor approach, i.e., predicting the MoAs and targets for a drug by transferring the knowledge of its k most similar drugs based on the distances computed above. Specifically, Mania calculates the affinity score of drug i and MoA (or target) j as a weighted majority voting by the k most similar drugs of drug i:

$$\mathbf{s}_{i,j} = \sum_{d \in \mathcal{N}_i} \cos(\mathbf{x}_i, \mathbf{x}_j) \mathbb{I}[d \in M_j], \tag{4}$$

where $\mathcal{N}_i$ is the set of the k most similar drugs of drug i, $\mathbb{I}[\cdot]$ is the indicator function, and M_j is the set of drugs that are annotated with MoA j in the training data. We set $k = 10$ in our experiments.

3. Results

We evaluate the ability of our Mania framework on uncovering the drug-drug relationships and predicting drug MoAs and drug targets by integrating multi-omics data. The integrated drug-drug similarity network given by Mania achieved an AUPRC score of 0.892, which is a substantial improvement over DNF[11] (0.838), the state-of-the-art integration method for drug taxonomy. The low-dimensional vector representations of drugs learned by Mania can also be used as plug-in features for off-the-shelf machine learning algorithms. We show that by using its learned feature vectors as the input of a k-nearest neighbor (kNN) algorithm, Mania successfully recovering around 75% true MoAs associated with drugs when evaluated with five-fold cross-validation on the list of its top 10 predictions, which is remarkably 20% higher than the DNF method. The details of our experiments are described below.

3.1. *Mania improves the quantification of drug-drug similarity*

Accurate quantification of drug similarity can help elucidate the drug-drug relationships and predict new targets for existing drugs. To assess the ability of Mania on quantifying the drug similarity, we calculated the cosine similarity among the low-dimensional vector representations between pairs of drugs. The cosine similarity matrix was compared to a binary drug similarity matrix, where an entry was set to 1 if the pair of drugs shares at least one MoA, and 0 otherwise. We filtered the MoAs to retain only those MoAs that are associated with at least two drugs before computing the binary similarity matrix. We evaluate the performance by computing the area under the receiver operating characteristic curve (AUROC) and the

area under the precision-recall curve (AUPRC). Note the whole process is unsupervised and no MoA information was available to Mania when learning the low-dimensional vector representations and computing the similarity metric. We implemented Mania based on Mashup[20] (`http://mashup.csail.mit.edu/`). We set the dimensionality of the low-dimensional vector representations of drugs as $d = 10$. The restart probability p_r of RWR was set to 0.8. We observed stable performances a wide range of values of d and p_r in our experiments.

We first compared the integrated similarity network by Mania's integration of multi-omics data with three similarity networks that were computed based on individual omics data, including the drug structure, drug sensitivity and drug perturbation (Fig. 2). We noticed that although the individual similarity networks of drug structure and drug sensitivity achieved roughly the same AUROC score (around 0.80), there was a 20% gap between their performances and that of the similarity network computed based on drug perturbation, which means there was a 20% fraction of drug-drug relationships (i.e., drug pairs that share same MoAs) that cannot be accurately predicted by drug perturbation data only. The similar effects were also observed for the AUPRC scores, where there were noticeable gaps between the individual networks of drug structure, drug sensitivity, and drug perturbation. These findings suggested that each omics data are not redundant. Instead, these multi-omics data are complementary and an integration of them would improve the quantification of drug-drug similarities. Even if sensitivity data and structure data have roughly the same AUROC score, the similar drug pairs identified by these two data sources were inherently different, thus motivating us to further integrate them. The performance of the integrated similarity network by Mania confirmed this hypothesis, where the AUROC score was substantially improved to 0.892, an 11% improvement over the best individual similarity network, and the AUPRC score was also significantly improved to 0.423, 43% higher than the best individual similarity network. We also observed similar results for the evaluation on the target data, in which the binary drug similarity matrix was computed based on whether two drugs share at least one common target. Notably, the performance of perturbation-based network was the worst among all three similarity networks, possibly due to the noisy and batch effect in large-scale perturbation experiments.

Furthermore, we compared Mania with DNF,[11] a state-of-the-art integration method for drug taxonomy. DNF was built upon the similarity fusion network (SNF) method, which takes individual networks as input, iteratively updates every network by message passing until convergence to a single network. Unlike our method that outputs low-dimensional vector representations that can be used to compute any kind of similarity of distance metric, the DNF method directly outputs the converged single network as a similarity network. We found that although both DNF and Mania improved the performance over individual networks, the performances of Mania were substantially higher than that of DNF when evaluated on the binary similar matrices based on both MoA and drug target data (one-sided Wilcoxon rank-sum test $P < 0.001$). For example, on the MoA data, Mania achieved a 25% improvement on AUPRC over DNF (AUPRC of 0.423 and 0.339 for Mania and DNF, respectively) and a 5% improvement on AUROC over DNF (AUPRC of 0.892 and 0.849 for Mania and DNF, respectively). Further comparisons suggested that Mania also outperformed a recently proposed matrix factorization-based integration framework, Collective-Matrix Factorization (CMF).[25]

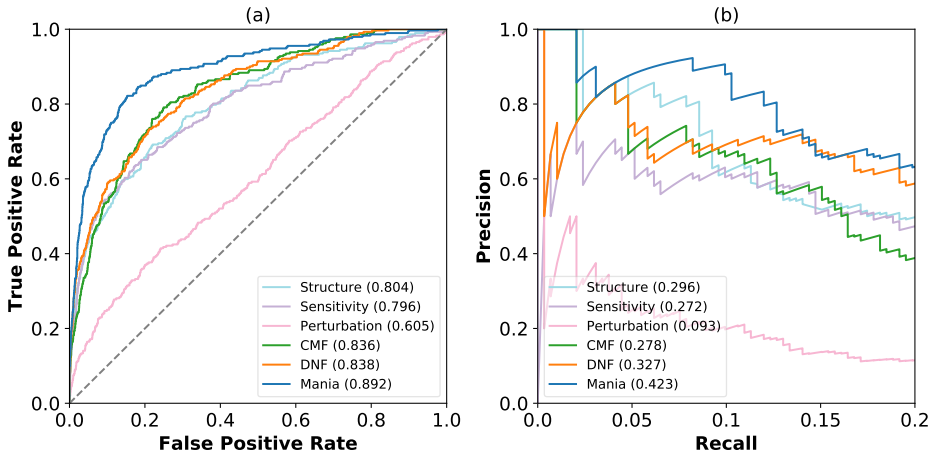


Fig. 2. **Accurate quantification of drug-drug similarity.** The similarity matrices based on individual data sources and the integration of DNF and Mania were compared to the binary similarity matrix derived from the ground truth data of drug MoA associations. Performance was evaluated using AUROC score (a) and AUPRC score (b).

These results suggested that Mania is capable of integrating drug properties from multi-omics data to provide a more comprehensive drug-drug similarity measure, which is helpful for elucidating the drug-drug relationships and potentially useful in MoA and target prediction of drugs, which we will demonstrate in the next section.

3.2. *Mania achieves accurate prediction of drug MoAs and targets*

In reality, one can collect and leverage the existing MoA (or target) information of well-studied drugs to infer the MoAs (or targets) of less well-characterized drugs. Therefore, it is also important to assess the ability of Mania on predicting drug MoAs and targets in a supervised way. As Mania outputs low-dimensional vector representations of drugs, we can use these vectors as plug-in features of drugs in off-the-shelf machine learning algorithms. In particular, we cast the prediction of drug MoAs and targets as a multi-label classification task and applied a k-nearest neighbor (kNN) method with the vector representations of drugs as input features. We assessed the performance of prediction using a five-fold cross-validation. For each test drug, we ran a weighted majority voting among its $k = 10$ most similar drugs based on the cosine distances imposed over the vector representations of drugs. Following the previous work,[26] we used the "recall@top-R" as the evaluation metric, which is defined as the fraction of true associated MoAs (or targets) that were retrieved in the list of top-R predictions for a drug. The motivation of using this metric was that a method that can recover the true MoAs (or targets) in the top-R predictions with high probability is desirable and useful in applications such as drug repurposing.

We compared the performance of Mania, CMF and DNF evaluated by recall@top-R for $R = 1, 5, 10$ on the MoA and target data (Fig. 3). We observed that Mania correctly recovered more MoAs and targets across all values of R. We would like to highlight the noticeable improvements of our method on smaller values of R (e.g., R =1 or 5), as evaluation under smaller

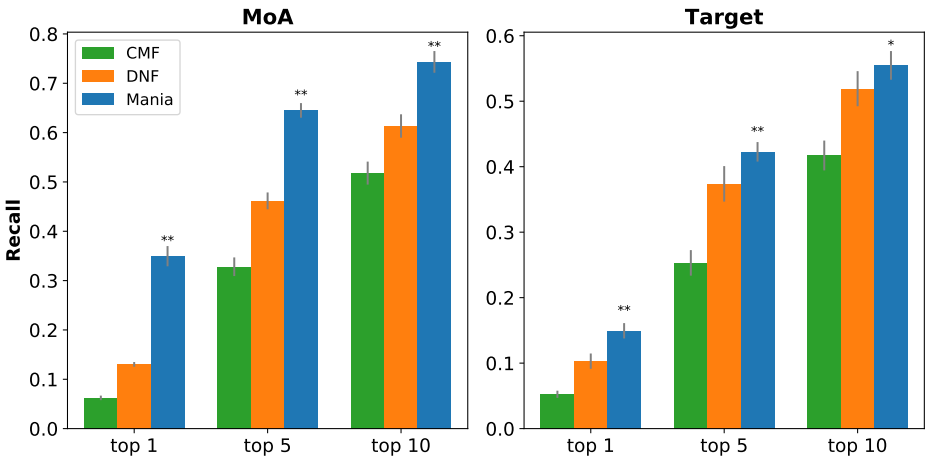


Fig. 3. **Comparison on the performance of drug MoAs and targets prediction.** We performed ten trials of five-fold cross-validation to compare the drug MoA/target prediction performance of Mania, CMF and DNF. The recall@top-R was used as the evaluation metric. *: $P < 0.01$ and **: $P < 0.001$, one-sided Wilcoxon rank-sum test.

values of R is a more challenging metric and reflects the precision of the very top predictions of a method. For instance, we observed Mania achieved a 0.350 recall@top-1 score, much higher than that of DNF (0.129) and CMF (0.062). We also found that the improvements of Mania over the other two methods were statistically significant (one-sided Wilcoxon rank-sum test $P < 0.01$). The superior performance of Mania demonstrates its potential on transferring the information of existing drugs to infer new MoAs or targets of drugs that are not well-studied, thus serving as a practical tool for drug repurposing.

3.3. *Identification of functionally-enriched drug communities*

The low-dimensional vector representations learned by Mania are able to capture the context-specific information and vectors of drugs that are functionally correlated will be co-localized in the low-dimensional space. To assess the pharmacological relevance among drugs exhibited by the vector representations of drugs learned by Mania, we applied an affinity propagation clustering algorithm on the low-dimensional vector representations of the common set of 277 drugs, with cosine distance as the distance metric. The affinity propagation algorithm requires an input "preference" parameter determines the likelihood of a particular drug to become an "exemplar" of a community (cluster). We set this preference parameter to be the 90% percentile of all pairwise similarities to encourage a relatively large number of communities to be produced, thus enabling each community to have more distinguishable pharmacological features. Note that the clustering was solely based on the low-dimensional vector representations that were learned by integrating multi-omics data, and no information of drug MoAs or targets was used to guide the clustering process.

We obtained 29 drug communities, with one drug selected as the representative (exemplar) drug in each community (Fig. 4). The size of the communities varies from 2 to 13, with a median size of 6 drugs. We observed that based on the low-dimensional vector representations

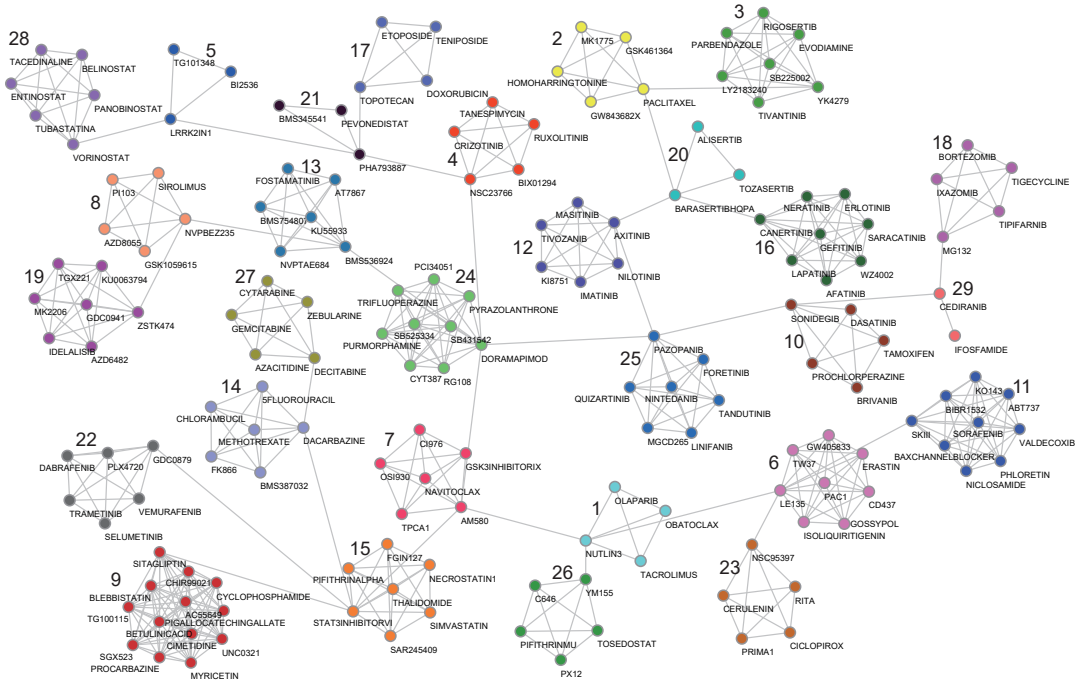


Fig. 4. **Network visualization of functionally-enriched drug communities identified by Mania.** Mania identified 29 drug communities by applying an affinity propagation clustering algorithm. One drug of each community was selected as the exemplar drug for that community. Inter-community edges were represented by exemplar-exemplar edges, which were obtained by building a minimal spanning tree among exemplar drugs.

of drugs, Mania produced various drug communities in which drugs with the same or similar functions were clustered together. For instance, Mania correctly identified the inhibitors for several targets, including the inhibitors of BRD4 (Cluster 5), PI3K/mTOR (Cluster 8), IGF-1R (Cluster 13), EGFR/ERBB (Cluster 16), TOP2 (Cluster 17), PSMB1 (Cluster 18), BRAD/MEK (Cluster 22), TGFBR1 (Cluster 24), and HDAC (Cluster 28). Among these, TOP2 (Topoisomerase II) has held great interest of researchers because of the discovery of active anti-cancer drugs that target TOP2,[27] and Mania identified four drugs (Teniposide, Etoposide, Topotecan, and Doxorubicin) targeting TOP2 (Cluster 17), which includes Etoposide and Doxorubicin, two clinically active agents.

We further conducted a Fisher's exact test between all the drugs grouped in a community and all drugs associated with a specific MoA or target, to assess whether the specific MoA or target is enriched in the community. We found that out of the 29 drug communities, 16 communities were significantly enriched ($P < 0.05$) for a direct target and 15 were significantly enriched for one MoA. For example, Cluster 28 were statistically enriched ($P < 10^{-5}$) for targets in the HDAC family and contained six known inhibitors for these targets, including Vorinostat and Belinostat, two FDA approved drugs. Another example is Cluster 16, where the 8 drugs were statistically enriched ($P < 10^{-7}$) for the ERBB and EGFR targets.

Taken together, the above results on clustering demonstrated the ability of Mania on illustrating the functional drug-drug relationships among drugs, by partitioning the drugs into

disjoint function-related communities based on the low-dimensional vector representations.

3.4. *Predictions of drugs for significantly mutated genes*

We then proceeded with explosive analysis to test the ability of Mania for drug repurposing. To this end, we obtained a list of 224 significantly mutated cancer genes across 21 tumor types in a recent analysis of The Cancer Genome Atlas.[28] Mania first predicted targets for each drug through a weighted majority voting process, and the top 10 scored targets for each drug were recorded. Each significantly mutated gene was then searched against the list of top 10 predicted targets for each drug, and we found that 20 genes had been predicted by Mania to have new corresponding drugs that were not included in the Drug Repurposing Hub. Among these, EGFR, a significantly mutated gene in lung adenocarcinoma, was predicted by Mania as the top 1 gene that can be targeted by the drug Saracatinib but has not been predicted by DNF in its top 10 list. Our prediction of the use of Saracatinib to treat lung cancer through reducing the activation of EGFR is also supported by studies in the literature,[29] where Saracatinib was found to be able to efficiently reduce the activation of EGFR. Interestingly, Saracatinib was clustered into Cluster 16, which contains several well-known EGFR inhibitors, including two launched drugs, Afatinib and Erlotinib. This demonstrated the ability of Mania on transferring the knowledge of well-studied drugs to other similar but less characterized ones and providing additional potential insights into drug repurposing.

4. Discussion

We have presented Mania, a method for integrating heterogeneous pharmacogenomic data, which can be used to predict drug MoAs and targets, as well as to study the drug-drug relationships. Mania integrates multiple data sources and learns low-dimensional vector representations of drugs, which encode the structural and functional information for drugs. We have demonstrated Mania accurately quantifies the drug-drug similarity and substantially improves the performance of MoA/target prediction.Furthermore, Mania identifies functionally-enriched drug communities and new drugs that potentially target cancer mutated genes.

In the future, we plan to pursue further improvements of Mania. First, besides the Pearson correlation as the similarity measure for perturbation and sensitivity data, we plan to explore other approaches that can better capture the drug-drug relationships. In addition, we will test our method on non-redundant data (redundancy may arise, for example, when structures of some drugs were derived from others) and analyze the prediction ability of each data source. Furthermore, we plan to integrate more types of pharmacogenomic data (e.g., Cancer Cell Line Encyclopedia[30]) to provide a more complete view of the relationships among drugs.

Acknowledgements: This work was supported in part by the NSF CAREER Award, the Sloan Research Fellowship, and the PhRMA Foundation Award in Informatics.

References

1. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *Journal of computational chemistry* **30**, 2785 (2009).
2. K. Bleakley and Y. Yamanishi, *Bioinformatics* **25**, 2397 (2009).

3. M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science* **321**, 263 (2008).
4. J. Li, X. Zhu and J. Y. Chen, *PLoS computational biology* **5**, p. e1000450 (2009).
5. F. Yang, J. Xu and J. Zeng, Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data, in *Pacific Symposium on Biocomputing*, 2014.
6. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics* **24**, i232 (2008).
7. F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi *et al.*, *Proceedings of the National Academy of Sciences* **107**, 14621 (2010).
8. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *science* **313**, 1929 (2006).
9. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu *et al.*, *bioRxiv* , p. 136168 (2017).
10. B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould *et al.*, *Cancer discovery* **5**, 1210 (2015).
11. N. El-Hachem, D. M. Gendoo, L. S. Ghoraie, Z. Safikhani, P. Smirnov, C. Chung, K. Deng, A. Fang, E. Birkwood, C. Ho, R. Isserlin, G. D. Bader, A. Goldenberg and B. Haibe-Kains, *Cancer Research* **77**, 3057 (2017).
12. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nature methods* **11**, 333 (2014).
13. P. Smirnov, Z. Safikhani, N. El-Hachem, D. Wang, A. She, C. Olsen, M. Freeman, H. Selby, D. M. Gendoo, P. Grossmann *et al.*, *Bioinformatics* **32**, 1244 (2015).
14. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, *Nucleic acids research* **44**, D1202 (2015).
15. Rdkit: Open-source cheminformatics `http://www.rdkit.org`.
16. D. Rogers and M. Hahn, *Journal of chemical information and modeling* **50**, 742 (2010).
17. L. R. Dice, *Ecology* **26**, 297 (1945).
18. M. G. Rees, B. Seashore-Ludlow, J. H. Cheah, D. J. Adams, E. V. Price, S. Gill, S. Javaid, M. E. Coletti, V. L. Jones, N. E. Bodycombe *et al.*, *Nature chemical biology* **12**, p. 109 (2016).
19. S. M. Corsello, J. A. Bittker, Z. Liu, J. Gould, P. McCarren, J. E. Hirschman, S. E. Johnston, A. Vrcic, B. Wong, M. Khan *et al.*, *Nature Medicine* **23**, 405 (2017).
20. H. Cho, B. Berger and J. Peng, *Cell systems* **3**, 540 (2016).
21. H. Cho, B. Berger and J. Peng, Diffusion component analysis: Unraveling functional topology in biological networks. *RECOMB* 2015.
22. S. Wang, H. Cho, C. Zhai, B. Berger and J. Peng, *Bioinformatics* **31**, i357 (2015).
23. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen and J. Zeng, *Nature Communications* **8** (2017).
24. H. Tong, C. Faloutsos and J.-y. Pan, *ICDM* , 613 (2006).
25. M. Žitnik and B. Zupan, *IEEE TPAMI* **37**, 41 (2015).
26. U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon and E. M. Marcotte, *PloS one* **8**, p. e58977 (2013).
27. J. L. Nitiss, *Nature reviews. Cancer* **9**, p. 338 (2009).
28. M. S. Lawrence, P. Stojanov, C. H. Mermel, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander and G. Getz, *Nature* **505**, p. 495 (2014).
29. L. Formisano, V. D'Amato, A. Servetto, S. Brillante, L. Raimondo, C. Di Mauro, R. Marciano, R. C. Orsini, S. Cosconati, A. Randazzo *et al.*, *Oncotarget* **6**, p. 26090 (2015).
30. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin *et al.*, *Nature* **483**, 603 (2012).

Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome

Emily K. Mallory[†,1], Ambika Acharya[†,2], Stefano E. Rensi[3], Peter J. Turnbaugh[4], Roselie A. Bright[5], and Russ B. Altman[5]

[1]*Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305, USA*

[2]*Computer Science Department, Stanford University, Stanford, CA 94305, USA*

[3]*Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*

[4]*Department of Microbiology & Immunology, University of California, San Francisco, CA 94143, USA*

[5]*Office of Health Informatics, Office of the Chief Scientist, Office of the Commissioner, Food and Drug Administration (FDA), Silver Spring, MD 20993, USA*

[6]*Departments of Bioengineering, Genetics, Medicine, and Biomedical Data Science, Stanford University, Stanford, CA 94305, USA*

Email: rbaltman@stanford.edu

Bacteria in the human gut have the ability to activate, inactivate, and reactivate drugs with both intended and unintended effects. For example, the drug digoxin is reduced to the inactive metabolite dihydrodigoxin by the gut Actinobacterium *E. lenta*, and patients colonized with high levels of drug metabolizing strains may have limited response to the drug. Understanding the complete space of drugs that are metabolized by the human gut microbiome is critical for predicting bacteria-drug relationships and their effects on individual patient response. Discovery and validation of drug metabolism via bacterial enzymes has yielded >50 drugs after nearly a century of experimental research. However, there are limited computational tools for screening drugs for potential metabolism by the gut microbiome. We developed a pipeline for comparing and characterizing chemical transformations using continuous vector representations of molecular structure learned using unsupervised representation learning. We applied this pipeline to chemical reaction data from MetaCyc to characterize the utility of vector representations for chemical reaction transformations. After clustering molecular and reaction vectors, we performed enrichment analyses and queries to characterize the space. We detected enriched enzyme names, Gene Ontology terms, and Enzyme Consortium (EC) classes within reaction clusters. In addition, we queried reactions against drug-metabolite transformations known to be metabolized by the human gut microbiome. The top results for these known drug transformations contained similar substructure modifications to the original drug pair. This work enables high throughput screening of drugs and their resulting metabolites against chemical reactions common to gut bacteria.

Keywords: Chemoinformatics; Matched molecular pair; Vector embedding; Drug metabolism; Microbiome.

1. Introduction

The trillions of microorganisms that colonize the human gastrointestinal tract (the gut microbiome) encode a diverse array of enzymes that catalyze the biotransformation of therapeutics drugs prior to or after absorption. The downstream microbial metabolites can have clinically relevant changes to their pharmacological properties, including the activation of prodrugs, drug

inactivation, and the reactivation of drugs subsequent to host metabolism.[1] The cardiac drug digoxin is a textbook example, wherein gut bacterial drug inactivation prior to drug absorption can reduce the bioavailability and thus efficacy of this essential medication. Digoxin is used to treat cardiac arrhythmia and heart failure and has a narrow therapeutic index. Although the bacterial metabolism of digoxin by the gut Actinobacterium *Eggerthella lenta* was originally described in 1983,[2] the enzymes responsible remained unknown for 30 years. Our prior work identified a 2-gene operon, referred to as the cardiac glycoside reductase (*cgr*) operon, unique to a digoxin metabolizing strain of *E. lenta*.[3; 4] Similar studies have implicated the gut microbiome in the metabolism of >50 distinct drugs, spanning multiple diseases,[1; 5] but no systematic experimental or computational analyses have been performed on the full set of FDA-approved compounds. Thus, the full scope of drugs that are metabolized or transformed by the human gut microbiome is currently unknown, representing a major gap in the scientific literature with immediate clinical implications.

The major bottleneck to a comprehensive view of gut microbial drug metabolism is the challenge of developing high-throughput analytical approaches to quantifying the parent compounds and all its possible metabolites. Typically, this is done by incubating cultured gut bacteria with a given drug and analyzing cell-free supernatants by mass spectrometry, a chemical-level technique used to detect quantities of molecules in a given substance.[6] Mass spectrometry interrogates the gut microbiome and its effects on forming metabolites in plasma, feces and urine.[7] While experimental techniques provide evidence of drug metabolism, they become time intensive and challenging when applied to large quantities of drugs. Therefore, there is a need for *in silico* methods that do not rely solely on experimental techniques. Quantitative structure-activity relationship (QSAR) modeling includes a set of computational techniques that are used for predicting the bioactivities of drugs by extrapolating from data observed for similar structures.[8] However, traditional QSAR methods have limited ability to address biotransformations, because they focus on individual molecules, while chemical transformations are defined over pairs of molecules.[9] Thus, there remains a need for efficient and effective *in silico* approaches for characterizing the properties of molecular transformations to enhance our understanding of drug metabolism in the human gut.

Matched molecular pair analysis (MMPA) is a specialized branch of QSAR modeling predicated on the concept of matched molecular pairs (MMPs) – two chemical structures that differ by a small, well-defined transformation.[10] For example, substrate-product pairs arising from hydroxylation by CYP3A4 are matched molecular pairs. A number of approaches to MMPA have been developed.[11] Fragment indexing based methods[12] are the most popular because they are efficient, but limited by exact matches. Such methods may fail to identify near-MMPs transformations relevant to an analysis, such as multiple site substitutions or transformations that do not occur at non-ring single bond sites.[13] Furthermore, they consider transformations independent of the surrounding molecular context.[14] We have reported an approach to address these limitations using kernel PCA embedded vector representations of molecules and principals of compositional semantics from computational linguistics.[15] While computational methods exist to compare, classify, and search enzymatic reactions[16-21], they frequently rely on direct comparison of molecular fingerprints as well as specific bond or atom changes within the

molecule. Our approach allows for the representation of chemical transformations as algebraic expressions of chemical structure vectors. *We hypothesize that molecules in chemical reactions can form analogous pairs with molecules in other reactions and be used to identify similar classes of reactions.* Furthermore, we can use such methods to identify chemical reactions with high similarity to drug-metabolite pairs. These methods could give us the tools to build a system that leverages the structural properties of chemical reactions and their enzymes as a proxy for drug metabolism.

While experimental methods for linking the human gut microbiome to drug metabolism are accelerating, there are still no high-throughput screening tools that could be broadly applied to all current drugs. Our work provides an important step towards this grand challenge by combining chemical reaction data with the concept of vector embeddings for molecules. We demonstrate the feasibility of detecting potential drug metabolism via bacteria in the human gut.

2. Methods

We introduce a pipeline for constructing a vector space for chemical reactions. This pipeline includes data processing, vector space construction and characterization, and chemical reaction and drug querying.

2.1. *Data sources and processing*

The chemical space and reaction set contained compounds and reactions from the MetaCyc metabolic pathway database.[22] We used the primary metabolic pathways provided by MetaCyc to generate a reaction list, react_list, that contained reaction name, direction, primary substrate compound, primary product, and Simplified Molecular Input Line Entry Specification codes (SMILES)[23] for each reaction. The unfiltered react_list contained 10,180 reactions, of which 8,981 were bacterial and 670 were *E. coli* specific. In addition, we constructed a list of 23 drug-metabolite pairs with identifiable structures from a curated list of known drugs modified by gut bacteria from Spanogiannopoulos et al.[1] These transformations were also added to react_list. Additionally, we removed reactions with high molecular weight compounds (>700) and those where the primary compounds are common types from a curated list, including “proton”, “coenzyme-A”, “water”, "NADP", "NADPH", etc. in order to include only relevant small molecules in the transformations. The final react_list contained 5,241 reactions, including 23 drug reactions, 5,116 bacterial reactions and 394 *E. coli*-specific reactions. To create the vector space we used all compounds from our dataset, not just those found in react_list. This compound set contained 11,893 unique compounds, a set we define as compound_dataset with size num_compounds.

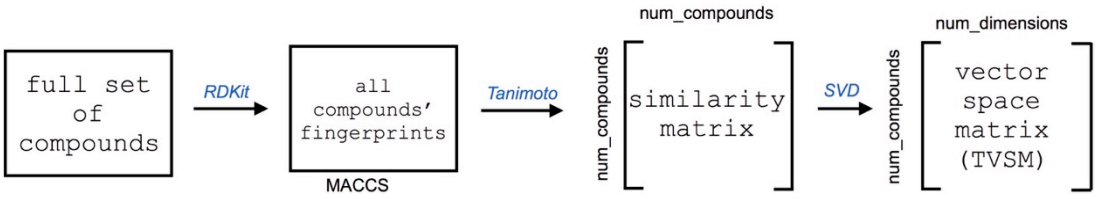


Fig. 1. Pipeline for constructing a vector space for a set of compounds. Starting with the full set of compounds, we generated fingerprints of all compounds, computed their pairwise Tanimoto similairities, and finally transformed the similarity matrix to the vector space matrix (TVSM) using kernel PCA. Matrix dimensions are included for both the similarity matrix and the vector space matrix (TVSM).

2.2. *Constructing molecular vector space*

The pipeline for constructing a molecular vector space is depicted in Figure 1 and previously described.[15] In summary, the pipeline takes SMILES as input, generates molecular fingerprints, and embeds the molecular fingerprints using kernel principal component analysis (KPCA).[24] From all compounds in reactions in compound_dataset, we used the corresponding SMILES string to generate chemical fingerprints. Specifically, we stored each compound using MACCS keys to encode molecular structure in a condensed bit vector.[25] To construct each vector, we used RDKit, an open source cheminformatics software for Python.[26] We used Tanimoto similarity (aka Jaccard index)[27] as the kernel function for kernel PCA using the molecular fingerprints. Therefore the resulting vector space matrix, Transformed Vector Space Matrix (TVSM), is of dimension (num_compounds, num_dimensions). We stored the mappings of row numbers in TVSM to compound names in a separate data structure. Next, we generated a scree plot to determine which components of the decomposition account for the majority of the variance in data (Figure S1). Since the scree plot plateaued at d = 8, we used that as a cutoff for the number of dimensions for each compound in TVSM. Using this cutoff, TVSM's final dimensions were 11,893 by 8.

2.3. *Characterizing vector spaces*

Next, we evaluated the effectiveness of the TVSM.

2.3.1. *Molecule-level Analysis*

To characterize types of chemical compound information stored in the TVSM, we clustered the vectors representing compounds using KMeans and performed an enrichment analysis to detect clusters of given chemical types. We computed the gap statistic[28], using a Python implementation[29], in order to find appropriate values of k at both the molecular and reaction levels.

To visualize the space, we used t-Distributed Stochastic Neighbor Embedding (t-SNE), a method which uses probability distributions to transform high dimensional data into 2 or 3-dimensions.[30]

We performed a hypergeometric enrichment analysis with a Bonferroni correction to determine enriched molecule types for each cluster. We used Chemical Entities of Biological

Interest (ChEBI)[31] ontology, which contains hierarchies for a large portion of the compounds found in our dataset. To give each molecule a ChEBI label, we observed that all molecules have the same top-level ChEBI term, either 72695 (for organic molecule) or 50860 (organic molecular entity). We then take the following three ChEBI terms downstream in the tree and create a tuple out of them. If there are multiple paths, we include all of these as descriptor types. An example tuple is depicted in Figure S2. For each compound in compound_dataset, we generated its ChEBI tuple and ran enrichment analysis on a clustering of the data.

2.3.2. *Reaction-Level Analysis*

To detect types of chemical reactions encoded in the vector space, we applied KMeans clustering to reaction vectors constructed using MetaCyc reactions. We constructed a vector for each reaction by subtracting vector A from vector B from TVSM for all reactions A → B in react_list. We applied the same KMeans methodology and series of experiments from the molecule clustering to these difference vectors. Additionally to evaluate the effectiveness of reaction clusters created using KMeans, we performed an enrichment analysis to characterize clusters by enzymes that catalyze the reactions. For this task, we wished to glean what reaction types were characteristic of each cluster using enzymes as a proxy for the reaction type. We performed these analyses using data from MetaCyc: both unigram and bigram enzyme names (see Suppl), Enzyme Consortium (EC) class numbers, and Gene Ontology (GO) codes for Molecular Function.[32]

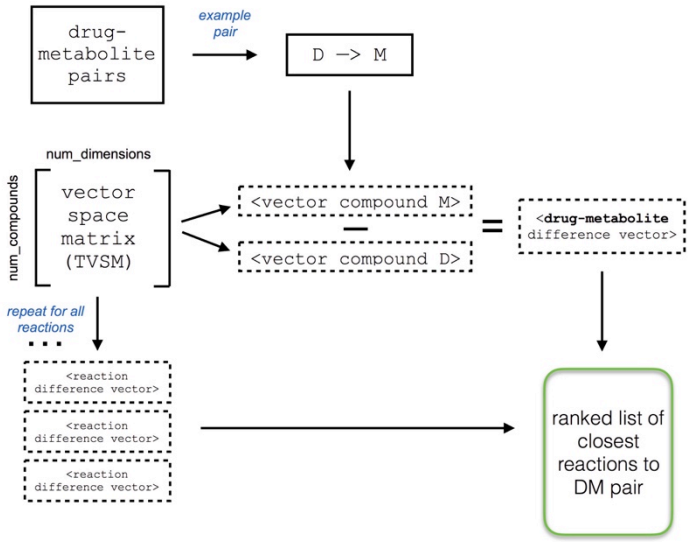


Fig. 2. Pipeline for querying reactions for drug-metabolite pairs. For each drug-metabolite pair, we subtract the drug vector from the metabolite vector to construct a difference vector. We repeat this process for all reactions in the dataset to create a ranked list of reactions most similar to the original drug-metabolite query.

2.4. *Querying drug-metabolite pairs against reaction vectors*

To find the most similar chemical reactions in the TVSM, we queried reactions against drug-metabolite pairs. The query pipeline is depicted in Figure 2. To detect the most similar reactions to

the query, we selected the k most similar difference vectors, using both Euclidean and cosine distance metrics. For each drug-metabolite pair, we constructed a difference vector by subtracting the drug vector from the metabolite vector. We next computed the similarity between the drug transformation vector and each reaction difference vector in our dataset. This resulted in a ranked list of all reactions for each drug-metabolite pair based on similarity between the drug and reaction difference vectors.

3. Results

3.1. *Molecule-level analysis*

KMeans clustering of all compounds using the TVSM resulted in clusters of similar compounds. Using the gap statistic, the optimal number of clusters was k=40 (from range k=1–50).

To visualize the high dimensional space of TVSM, we used t-SNE to visualize both 2D and 3D representations of the data. Molecules in this space, particularly at the 2D level, are clustered close to others in the cluster (Figure 3A). This suggests that the points in the clusters formed from this method have small intra-cluster distances, which is confirmed when adding another dimension (Figure 3B).

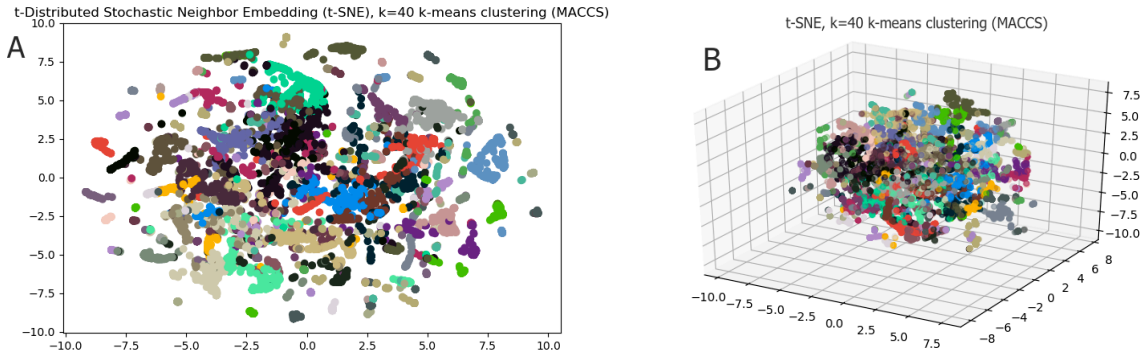


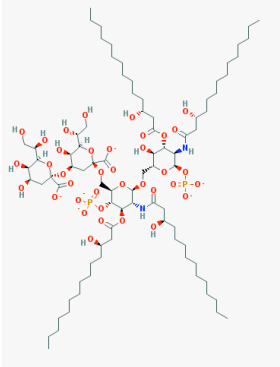
Fig. 3. t-SNE plots in 2D(A), 3D(B) of molecules in TVSM, clustered using KMeans with k=40. Colors identify clusters and are the same across Figures 3A and 3B.

After performing an enrichment test using ChEBI tuples, we found that many of the clusters contained similar ChEBI terms. For example, one cluster was predominantly made up of ring structures, another of lipids, and others, such as cluster 25 depicted in Figure 4, captured many different types. For cluster 25 specifically, the molecules in the cluster were a combination of the enriched terms. In addition, clusters contained high intra-cluster molecular similarity (mean pairwise Tanimoto similarity of molecular MACCS keys in Table S9).

3.2. *Reaction-level analysis*

To select k for KMeans clustering on the reaction vectors, we computed k = 32 using the gap statistic. To visualize the high dimensional space of reaction vectors, we used t-SNE to visualize a 2D representation of the data. After clustering the reaction vectors with k = 32, we discovered that the reaction vectors were not evenly distributed between different clusters, but instead one cluster

contained 38% of the data (labeled cluster 12 during clustering). During visualization using t-SNE (Figure 5A), this cluster spanned the entire two-dimensional space and did not contain signal for specific reactions. To detect further clusters within cluster 12, we performed k-means clustering (computed k = 45) on reactions occurring within this cluster. Using t-SNE for 2-D visualization in Figure 5B, points within individual clusters are closer to each other than those in other clusters. This is in direct contrast to the t-SNE plot for the full data in Figure 5A, where the points in cluster 12 spanned the entire space.



ChEBI descriptor	p-value
Carbonyl compound	1.19E-08
Organic aromatic compound	7.04E-09
Organophosphate oxoanion	4.61E-24
Sphingolipid	2.90E-36
Glycolipid	7.66E-29

Fig. 4. Results from enrichment analysis on cluster 25 from KMeans clustering with k=40 on TVSM. We also show an example structure (alpha-Kdo-(2->4)-alpha-Kdo-(2->6)-lipid IVA) from this cluster. Full results can be found in Supplementary Table S1.

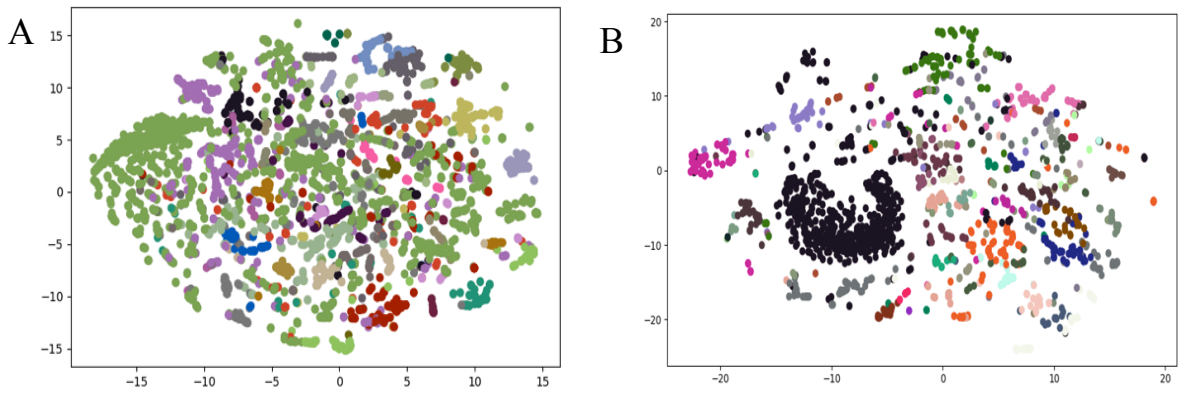


Fig. 5. t-SNE plots in 2D for all reaction vectors (A) and reactions in cluster 12 (B), clustered using KMeans. Colors identify clusters.

Reaction cluster enrichment results for GO Molecular Function terms and EC numbers from three of the clusters are found in Table 1. The enriched GO codes matched the functionality of the enzymes and were corroborated by the enriched EC numbers and expanded on the unigram, bigram, and GO enrichment results (Tables S2-S6). For example, cluster 20 mapped to GO:0047893, which is consistent with EC number 2.4.1 hexosyltransferase and cluster 23 mapped to GO:0008934 inositol monophosphate 1-phosphatase activity, which is consistent with EC number 3.1.3 (phosphoric monoester hydrolases). In addition, cluster 3 is enriched with GO term "3-beta-hydroxy- delta5-steroid dehydrogenase activity", consistent with enrichment with EC

class 1.1.1.- (oxidoreductases, acting on the CH-OH group of donors, with NAD(+) or NADP(+) as acceptor).

Table 1. Results from enrichment analysis on select clusters from KMeans clustering (k = 32). We report one enriched group for each category in three clusters. Full results can be found in Supplementary Tables S2-S5.

Cluster	EC number (p-value)	GO-code (p-value)
3	1.1.1 Oxidoreductases, acting on the CH-OH group of donors, with NAD(+) or NADP(+) as acceptor. (3.12e-36)	0003854 3-beta-hydroxy- delta5-steroid dehydrogenase activity (5.09e-10)
20	2.4.1 hexosyltransferase (2.28e-84)	0047893 flavonol 3-O-glucosyltransferase activity (1.22e-07)
23	3.1.3 Phosphoric monoester hydrolases (2.11e-50)	0008934 inositol monophosphate-1- phosphatase activity (8.16e-08)

For cluster 3 from Table 1, we show sample reactions in Figure 6. The oxidation of an OH group is found in A, B, and D in Figure 6. While the cluster is enriched for dehydrogenase reactions, the cluster is not composed solely of those reactions. In particular, reaction C in Figure 6 is methylation. Despite the inclusion of additional types of reactions in individual clusters, the clusters contained signal for specific types of reactions compared to the overall set of reactions in the dataset.

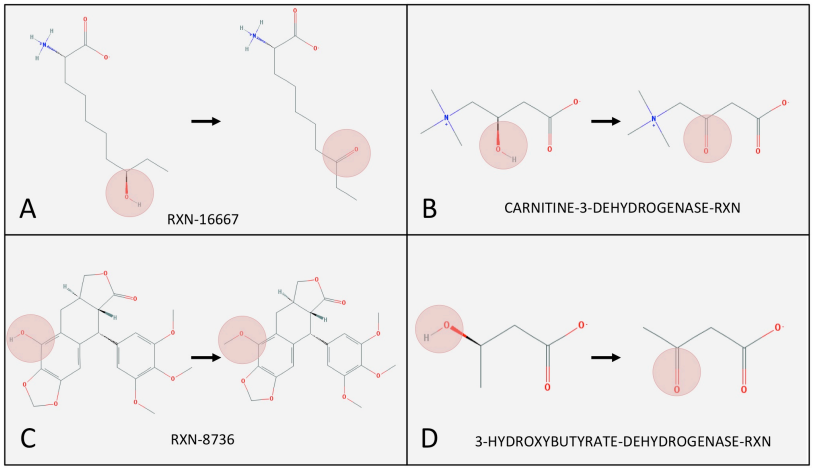


Fig. 6. Sample reactions from cluster #3 from KMeans clustering on reaction vectors. Reactions in this cluster were predominantly characterized by oxidation. The transformations in the sample reactions are highlighted. Reaction identifiers from MetaCyc are included for each reaction.

3.3. *Querying reaction vectors against drug-metabolite pairs*

For each drug metabolite pair, we ranked all reactions by similarity of their reaction vector to the drug-metabolite vector to find the top 10 closest reactions. Here, we show examples of the drugs Digoxin and Levodopa in Figures 7 and 8, respectively. Full results can be found in the Supplementary Tables S7 and S8. As we are particularly interested in bacterial reactions, we mapped each reaction to any bacterial pathway or more specifically *E. coli* as a representative gut bacterial species. While all top 10 reactions for all 23 drugs existed in bacterial pathways, several

top hits were present in *E. coli* pathways. For example, the transformation of sorivudine to E-5-(2-bromovinyl)uracil was similar to the transformation of beta-nicotinate D-ribonucleotide to nicotinate adenine dinucleotide. Similarly, the second closest reaction vector to the transformation of zonisamide to 2-sulfamoylacetylphenol was present in *E. coli* metabolic pathways.

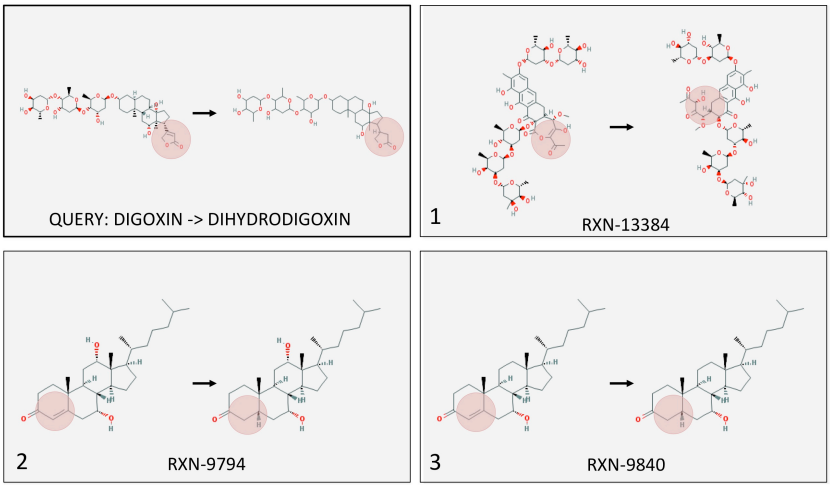


Fig. 7. The three closest reactions to drug-metabolite pair digoxin-dihydrodigoxin. The retrieved reactions are categorized by the hydrogenation of a double bond in a ring.

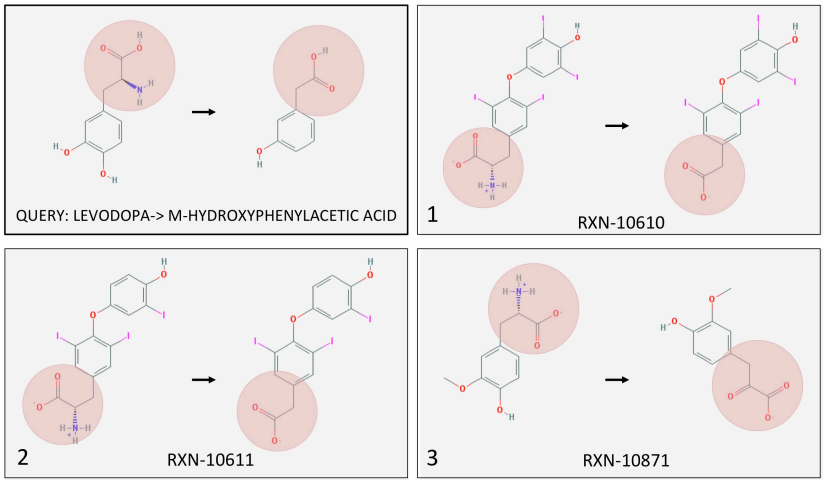


Fig. 8. The three closest reactions to drug-metabolite pair levodopa-m-hydroxyphenylacetic acid. The retrieved reactions are characterized by deamination to form a ketone and then decarboxylation of the ketone, occurring near a ring structure.

Figure 7 depicts the three closest reactions to digoxin and its inactive metabolite dihydrodigoxin. While our current experimental data suggests that the *cgr* operon of *E. lenta* is capable of reducing the double bond in the α,β-unsaturated lactone of digoxin, the top reaction represents a more complex ring opening reaction. The second and third ranked reactions were more in line with our expectations, depicting reduction of a double bond in a ring. For levodopa (another drug metabolized by gut bacteria), we see the deamination to form a ketone and then decarboxylation of the ketone, occurring near a ring structure (Figure 8). The second ranked reaction was nearly identical to levodopa, and the first and third reactions were highly similar to

each other. Because the reaction set contained all reactions regardless of similarity within the set, it is unsurprising that similar reactions rank highly to the same query drug-metabolite pair. In both figures it is apparent that while the top reactions have different overall structures, their local transformations are similar.

4. Discussion

Understanding the space of reactions that occur in the gut microbiome is a critical step towards predicting the intricacies of drug metabolism in the human gut. Bacteria can metabolize drugs via many different enzymes, in particular those catalyzing reduction and oxidation reactions.[1; 33; 34] Knowledge of types of reaction transformations occurring in bacteria as well as specific bacterial enzymes are necessary for predicting potential drug metabolism. Not only do we know of relatively few cases of drug metabolism, research is ongoing for detecting bacterial enzymes relevant to the gut microbiome.[35] In this work, we described a pipeline for constructing chemical embeddings for chemical reactions. In addition, we characterized the resulting reaction vectors using enzymes from MetaCyc. While all 23 drugs in our system had transformation vectors close to bacterial reaction vectors, four drugs had at least one reaction from *E. coli* in the three closest reaction vectors. While MetaCyc does not contain the complete set of enzymatic reactions occurring in the gut microbiome, similar reactions may provide hypotheses for drug metabolism and thus can be used for high throughput computational screening and hypothesis generation for drug metabolism in the human gut microbiome. In addition, known enzymes or transformations found in specific gut microbial species[36] can be used to screen for similar drug transformations in the vector space. In this way, one can computationally generate hypotheses for drug transformations that may occur via the gut microbiome.

We were able to transform molecules into a computational vector space, characterize and then fine-tune the space to best reflect properties at both the molecular and reaction level. Furthermore, we showed preliminary drug-metabolite queries inside a vector landscape. By detecting similar reaction and drug-metabolite vectors, we showed a first step toward modeling drug metabolism by gut microbes using the vector space. We found evidence of successful reaction vector clustering, as shown by trends of clusters enriched with enzymes with similar functions (Table 1, Supplementary Tables S2-S5). For example, cluster 3 is enriched for a specific type of oxidoreductases and cluster 20 is enriched for glucosyltransferases. Because enzymes can catalyze multiple types of reactions, we performed enrichment analyses using GO terms and EC classes for reactions. The enriched EC classes were consistent with the GO terms. Therefore, despite having one cluster that accounts for close to 40% of the data and some clusters sharing EC class, the smaller clusters have significant enrichment indicating that these methods can be used to differentiate reaction transformations.

Vector addition and subtraction in the vector space can describe transformation properties of drug metabolism. For example, the centroid for a highly enriched reductase cluster could be classified as a ‘reductase vector’. Using such enzyme vectors, we can add compound vectors to find compounds that may undergo the transformation. Additionally, drug metabolism does not occur in a single step, but occurs over a sequence of transformations in order for the drug to

become active in the body and eventually be eliminated. Through the use of transformation vectors with additional drug metabolites and similar compounds, we can use this technique to detect the transformation pathway from one compound to another, based on enzyme vectors. This automatic construction of drug-related pathways would aid current manual curation efforts for pathway construction at drug databases like PharmGKB[37] and provide an initial automatically constructed pathway for other users that do not require a high quality curated pathway for their work.

Characterizing reaction vectors was a more challenging task compared to the molecule vectors because the reaction vectors reflect the transformation between the two molecules. Observing the silhouette plots for clustering done with the best k for both TVSM and reaction vectors, we noticed that the former is significantly better distributed, with clusters around the same size. The reaction vector silhouette plot had one large cluster (cluster 12) that dominated the clustering and captured many different types of reactions. The reaction vector clusters we found within cluster 12 are closer to each other than the original reaction clusters.

In addition to the challenge of classifying reaction vectors effectively, limited data provided another obstacle for clustering. Although the MetaCyc database contains a large curated set of metabolic pathways, it is limited in examples especially critical to the understanding the metabolism of drugs in the human gut. Since this approach is completely data-driven, limited data in the types of transformations necessary for this type of metabolism hinders the model's ability to learn meaningful representations of molecules and their reactions. Thus, when querying drugs in the space, the resulting reactions may not be the most useful in terms of correlating with drug-metabolite interactions. One solution is to only use bacterial reactions in the drug queries; however, this approach is limited by the data available in MetaCyc. To add additional bacterial reactions to the database, one solution is to incorporate bacterial reactions described in the literature.

While this computational approach is more efficient and time-effective, supplementing the methods outlined here with experimental features would bolster the model. Even though we have shown that structure is a large component in making these predictions, incorporating empirical data would give us even more information to build on. Lastly, we queried from a very small subset of drugs, and for this proof-of-concept to be implementable for predictions, we must add in a larger set of drug-metabolite pairs. This remains challenging because most of the public information about drug metabolites is in text, image, or PDF format.

5. Conclusion

We developed a pipeline for computing similarities between chemical reactions and drug-metabolite transformations catalyzed by bacterial enzymes in the human gut microbiome. We show meaningful clusters for molecules and reactions in the transformed vector space based on chemical similarity, and how this data can be used to understand drug metabolism. Further development of these analytical pipelines and inclusion of larger chemical and reaction datasets pertaining specifically to the microbiome will enable high throughput screening of drugs and their resulting metabolites against chemical reactions common to gut bacteria.

6. Acknowledgments

The authors acknowledge Dr. Michael Fischbach for discussions regarding drug metabolism via the microbiome, and Dr. Larry Callahan, Dr. Frank Switzer, and Ms. Elaine Johanson for insights regarding chemistry and FDA work. This publication was made possible by grant U01FD004979 from the FDA, which supports the UCSF-Stanford Center of Excellence in Regulatory Science. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or FDA. EKM is supported by NIH NRSA F31 LM012354. SER is supported by NIH GM102365 and NIH GM61374. PJT is supported by NIH R01HL122593 and the Searle Scholars Program.

References

1. P. Spanogiannopoulos, *et al.*, *Nature reviews. Microbiology*. **14**, 273-287 (2016).
2. J. R. Saha, *et al.*, *Science*. **220**, 325-327 (1983).
3. H. J. Haiser, *et al.*, *Gut microbes*. **5**, 233-238 (2014).
4. H. J. Haiser, *et al.*, *Science*. **341**, 295-298 (2013).
5. N. Koppel, V. Maini Rekdal and E. P. Balskus, *Science*. **356**, (2017).
6. D. E. Lefebvre, *et al.*, *Nanotoxicology*. **9**, 523-542 (2015).
7. B. D. Wallace and M. R. Redinbo, *Current opinion in chemical biology*. **17**, 379-384 (2013).
8. A. Cherkasov, *et al.*, *Journal of medicinal chemistry*. **57**, 4977-5010 (2014).
9. R. P. Sheridan, P. Hunt and J. C. Culberson, *J Chem Inf Model*. **46**, 180-192 (2006).
10. A. G. Dossetter, E. J. Griffen and A. G. Leach, *Drug Discov Today*. **18**, 724-731 (2013).
11. C. Tyrchan and E. Evertsson, *Computational and structural biotechnology journal*. **15**, 86-90 (2017).
12. J. Hussain and C. Rea, *J Chem Inf Model*. **50**, 339-348 (2010).
13. E. Griffen, *et al.*, *Journal of medicinal chemistry*. **54**, 7739-7750 (2011).
14. G. Papadatos, *et al.*, *J Chem Inf Model*. **50**, 1872-1886 (2010).
15. S. Rensi and R. B. Altman, *Computational and structural biotechnology journal*. **15**, 320-327 (2017).
16. H. Kraut, *et al.*, *J Chem Inf Model*. **53**, 2884-2895 (2013).
17. Q. N. Hu, *et al.*, *PloS one*. **7**, e52901 (2012).
18. N. Schneider, *et al.*, *J Chem Inf Model*. **55**, 39-53 (2015).
19. S. A. Rahman, *et al.*, *Nature methods*. **11**, 171-174 (2014).
20. Q. N. Hu, *et al.*, *Bioinformatics*. **27**, 2465-2467 (2011).
21. V. Giri, *et al.*, *Bioinformatics*. **31**, 3712-3714 (2015).
22. R. Caspi, *et al.*, *Nucleic acids research*. **44**, D471-480 (2016).
23. E. Anderson, G. D. Veith and D. Weininger, *Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021*. (1987).
24. S. Bernhard, *et al.*, *Neural Computation*. **10**, 1299-1319 (1998).
25. J. L. Durant, *et al.*, *J Chem Inf Comput Sci*. **42**, 1273-1280 (2002).
26. G. Landrum (2016), http://www.rdkit.org.
27. D. Bajusz, A. Rácz and K. Héberger, *Journal of Cheminformatics*. **7**, 20 (2015).
28. R. Tibshirani, G. Walther and T. Hastie, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. **63**, 411-423 (2001).
29. K. Sang (2016), https://github.com/minddrummer/gap.
30. F. Pedregosa, *et al.*, *Journal of Machine Learning Research*. **12**, 2825-2830 (2011).
31. J. Hastings, *et al.*, *Nucleic acids research*. **41**, D456-463 (2013).
32. Gene Ontology Consortium, *Nucleic acids research*. **43**, D1049--D1056 (2015).
33. T. Sousa, *et al.*, *International journal of pharmaceutics*. **363**, 1-25 (2008).
34. R. Saad, M. R. Rizkallah and R. K. Aziz, *Gut pathogens*. **4**, 16 (2012).
35. B. J. Levin, *et al.*, *Science*. **355**, (2017).
36. A. Heinken, *et al.*, *Gut microbes*. **4**, 28-40 (2013).
37. M. Whirl-Carrillo, *et al.*, *Clinical pharmacology and therapeutics*. **92**, 414-417 (2012).

Loss-of-function of neuroplasticity-related genes confers risk for human neurodevelopmental disorders

Milo R. Smith[1-6], Benjamin S. Glicksberg[1,3,4,6]*, Li Li[3,6], Rong Chen[3], Hirofumi Morishita[1,2,4,5], Joel T. Dudley[3,6]*

Department of Neuroscience[1], Departments of Psychiatry and Opthamology[2], Department of Genetics and Genomic Sciences[3], Friedman Brain Institute[4], Mindich Child Health and Development Institute[5], Institute for Next Generation Healthcare[6]
Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.
New York City, NY 10029, USA
** Authors contributed equally*
Corresponding authors: joel.dudley@mssm.edu, hirofumi.morishita@mssm.edu

High and increasing prevalence of neurodevelopmental disorders place enormous personal and economic burdens on society. Given the growing realization that the roots of neurodevelopmental disorders often lie in early childhood, there is an urgent need to identify childhood risk factors. Neurodevelopment is marked by periods of heightened experience-dependent neuroplasticity wherein neural circuitry is optimized by the environment. If these critical periods are disrupted, development of normal brain function can be permanently altered, leading to neurodevelopmental disorders. Here, we aim to systematically identify human variants in neuroplasticity-related genes that confer risk for neurodevelopmental disorders. Historically, this knowledge has been limited by a lack of techniques to identify genes related to neurodevelopmental plasticity in a high-throughput manner and a lack of methods to systematically identify mutations in these genes that confer risk for neurodevelopmental disorders. Using an integrative genomics approach, we determined loss-of-function (LOF) variants in putative plasticity genes, identified from transcriptional profiles of brain from mice with elevated plasticity, that were associated with neurodevelopmental disorders. From five shared differentially expressed genes found in two mouse models of juvenile-like elevated plasticity (juvenile wild-type or adult *Lynx1*$^{-/-}$ relative to adult wild-type) that were also genotyped in the Mount Sinai Bio*Me* Biobank we identified multiple associations between LOF genes and increased risk for neurodevelopmental disorders across 10,510 patients linked to the Mount Sinai Electronic Medical Records (EMR), including epilepsy and schizophrenia. This work demonstrates a novel approach to identify neurodevelopmental risk genes and points toward a promising avenue to discover new drug targets to address the unmet therapeutic needs of neurodevelopmental disease.

Keywords: integrative genomics, neuroplasticity, neurodevelopment, risk genes, inflammation, drug targets

1. Introduction

Neurodevelopmental disorders place enormous personal and economic burdens on society [1,2]. In addition to environmental factors, genetic factors are known to be important predictors of neurodevelopmental outcomes, and the perinatal period comprises critical windows of disease susceptibility when mutations may express their deleterious effects on neurodevelopment. Particularly important windows of susceptibility are childhood critical periods that allow brain circuits to be refined by sensory and social experiences to establish normal perception and cognition [3–6].

Disruption of these critical periods can alter the developmental trajectory and confer risk for neurodevelopmental disorders [7,8]. Previous studies have found that genetic disruptions in neurodevelopmental disorder-related genes (MeCP2, Ube3a, Fmr1) led to disruptions in critical period plasticity [9–11]. The finding that alteration in neurodevelopmental genes disrupts developmental plasticity and leads to neurodevelopmental phenotypes calls for a comprehensive search for neurodevelopmental risk genes associated with plasticity.

To systematically identify plasticity gene variants, we used one of the best-studied models of childhood plasticity, namely critical period plasticity of visual cortex [12]. In response to deprivation of light to a single eye for a few days, neural responses in the cortex diminish correspondingly. Underlying this plasticity is major circuit remodeling [8] and only occurs naturally in youth - in adults there is minimal plasticity. Moreover, perturbations during the critical period are permanent - if the eye is occluded throughout juvenile life, there will be a persistent cortex-dependent reduction in visual acuity - a condition called amblyopia. Here, we use the mouse model of critical period visual plasticity [13] as our starting point to systematically identify genes related to neuroplasticity. This model has emerged as an indispensable model system to dissect the molecular mechanisms underlying functional cortical plasticity and whose transcriptional representation is functionally predictive [14]. Importantly, to control for age, elevated plasticity can be recapitulated in adult mice by genetically manipulating genes important for critical period plasticity. Here, we took a strategy of generating two transcriptional plasticity signatures from the visual cortex, one from juvenile and the other from adult *Lynx1*$^{-/-}$ mice, the latter upon release of the Lynx1 cholinergic plasticity brake exhibit juvenile-like plasticity [15]. These signatures represent the plasticity-permissive transcriptional landscape of visual cortex and the genes shared between these signatures are high confidence plasticity-related genes referred to here as "putative plasticity genes." Identifying putative plasticity genes in a data-driven, genome-wide manner sets the stage for high-throughput detection of potential novel risk variants associated to neurodevelopmental disease.

Past work to identify neurodevelopmental risk variants has traditionally focused on genome wide association studies (GWAS), family-based, or hereditary (e.g. twin, adoption) studies. While all are successfully used to identify risk variants, these approaches are inherently disease-centric rather than function-centric. By limiting discovery to a specific disease (e.g. microcephaly), discovery of cross-disease functional factors are missed. Our approach begins with functional plasticity-related genes and identifies any associated disease or phenotypic risk genes. This allows for greater biological insight downstream while increasing sensitivity by shrinking the search space to identify real associations between neurodevelopmental genes and disease. Moreover, by deriving putative plasticity genes using a genome-wide transcriptional approach across multiple models of elevated plasticity and coupling it with an integrative genomics methodology to identify risk genes across many diseases, we propose a highly systematic approach to identifying neurodevelopment risk genes, which does not depend on prior knowledge of either the specific functional role of the plasticity genes nor specific diseases.

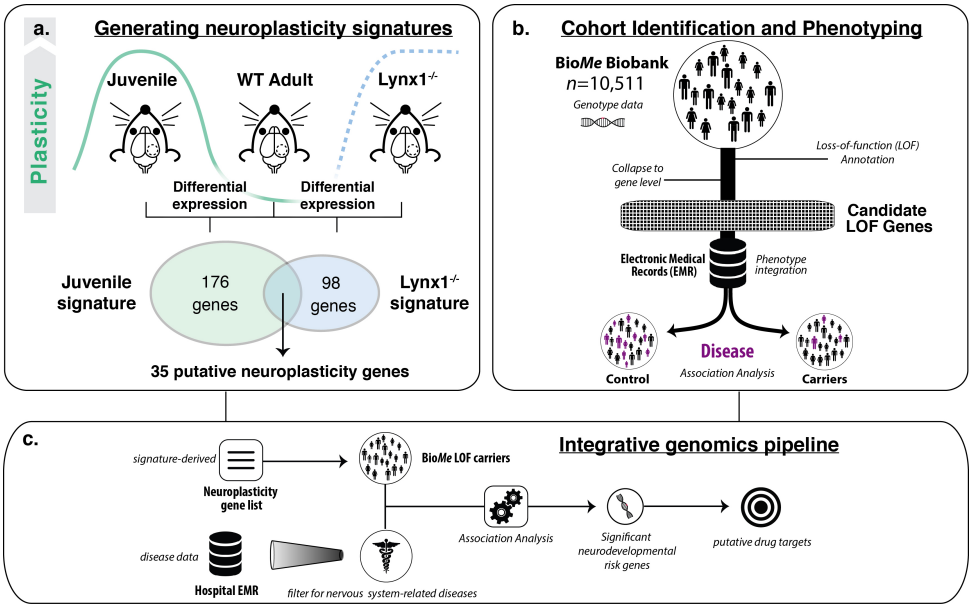


Figure 1. An integrative genomics approach to validate a role for putative neuroplasticity genes in human neurodevelopmental disorder. (**a**) We generated transcriptional neuroplasticity signatures from two mouse models of elevated, neurodevelopmental plasticity (juvenile and Lynx1-/-) to identify 35 shared putative plasticity genes used for downstream analysis. (**b**) We derived 2117 putative loss-of-function (LOF) variants in 1665 genes in a population of 10,510 patients from the Mount Sinai BioMe BioBank coupled with disease diagnosis data from EMR. (**c**) We applied an integrative genomics pipeline to identify associations between LOF of genotyped putative plasticity genes and nervous system diseases by logistic regression controlling demographic covariates to provide human-level evidence for multiple neurodevelopmental risk genes.

In order to assess the relationship between putative plasticity genes and neurodevelopment outcomes, we utilized a biobank of individuals with genetic data and longitudinal phenotype information from a large hospital system. To identify associations of large effect within this human dataset, we focused on the impact of loss-of-function (LOF) mutations on nervous system disease susceptibility. The study of LOF mutations in the human genome has played an important role in understanding etiologies of human disease, as these natural human knockouts shed light on gene function in the context of disease [16,17]. In a landmark study, MacArthur et al. identified LOF variants within protein coding genes using whole genome data from the 1000 Genomes Project [18,19]. They estimated that the typical human genome contains around 100 LOF variants and identified rare LOF variants that likely confer risk for disease. This work has been extended to elucidate the role and function of genes through LOF mutations in a variety of diseases: *ABCA1* with pancreatic β-cell dysfunction in Type 2 Diabetes [20]; *SETD5* with intellectual disability [21]; *APOC3* with reduced risk of both ischemic vascular disease and coronary disease [22,23]; *SLC30A8* with protection from Type 2 Diabetes [24, p. 30], among others. These findings have direct applications for identifying molecular targets to guide and accelerate drug discovery [25]. Using this strategy, Graham et al. found that antisense oligonucleotides targeting *ANGPTL3* transcripts reduced levels of atherogenic lipoproteins in humans [26]. In the current study, human findings may reveal novel drug targets relevant to neuroplasticity and neurodevelopment. Genes identified may be appropriate to directly target with small molecules. In addition, molecular editing of these targets in mouse could reveal novel molecular machinery important for disease phenotypes seen in human and lead to novel rescue therapeutics. In fact, Diamantopoulou *et al.* used

such a strategy to identify a LOF mutation in *Mirta22* that rescues schizophrenia-related phenotypes in a mouse model of 22q11.2 deletion [27].

Here, applying an integrative genomics approach we identified 35 putative plasticity genes across two mouse models, including those important in inflammatory processes. After identifying putative plasticity genes, we systematically identified LOF variants within these genes across the Charles Bronfman Institute of Personalized Medicine Mount Sinai Bio*Me* BioBank and linked Electronic Medical Record (EMR) cohort of 10,510 patients. We then assessed associations between putative plasticity genes and various neurodevelopment-related diseases using logistic regression that controls for demographic covariates (see **Figure 1** for the research design). This approach revealed potential risk variants in multiple putative plasticity genes for neurodevelopmental disorders, including epilepsy and schizophrenia. These findings provide human evidence for a role of plasticity-related genes in neurodevelopment and establish a novel approach to identify human neurodevelopment risk variants. Using model-derived putative plasticity genes as seeds to identify neurodevelopmental risk genes in human immediately sets the stage for pre-clinical studies to determine the mechanisms by which these novel risk genes disrupt neurodevelopment, and provides novel targets for therapeutic discovery.

2. Methods

All data processing and statistical analyses were conducted in R v 3.2.2 and Python v 2.7.10.

2.1 Neuroplasticity signatures

To identify putative neuroplasticity genes, we compared primary visual cortex transcriptomes of juvenile wild-type mice or adult *Lynx1*$^{-/-}$ compared to adult wild-type (n = 3 all groups). We used Limma [28] to quantile normalized raw microrray probe-level data and RankProd [29] to compute rank-based differential expression of mouse genes, which we mapped to orthologous human genes using the Mouse Genome Informatics homology reference to yield 176 and 98 gene signatures (juvenile wild-type and adult *Lynx1*$^{-/-}$ respectively), 35 of which were shared (Fisher Exact Test: OR=37.1, 95% CI = 23.8–58.0, $p < 2.2 \times 10^{-16}$, replication of comparison found in [14]) **(Figure 1a).** Both juvenile and *Lynx1*$^{-/-}$ mice have elevated experience-dependent plasticity, whereas adult wild-type mice have reduced plasticity. Transcriptional data was derived from publicly available data (GSE89757 [14]). We used the well-established gene set enrichment approach from Enrichr [30] to determine known Gene Ontology Biological Processes relevant to the 35 putative plasticity genes (using a FDR < 0.05) and further assessed relevance of individual genes that mapped to genotyped variants using a literature-based approach.

2.2 Hospital and biobank cohort

The Mount Sinai Hospital, located in Upper Manhattan, NY, has EMR that are de-identified and stored within the Mount Sinai Data Warehouse. These records contain clinical (e.g. disease diagnoses) and demographic data for over four million patients as of February 2015. The Charles

Bronfman Institute of Personalized Medicine Bio*Me* biobank (http://icahn.mssm.edu/research/ipm) within the Icahn School of Medicine at Mount Sinai has collected genetic data for over 30,000 patients with linked EMR as of 2016. For the current analysis, we utilized a subset of Bio*Me*, consisting of over 11,000 individuals that were genotyped using the Illumina Human Omni Express Exome Bead-8 BeadChip v1.1 array. This cohort consists of 61.2% females and 38.8% males and the self-reported racial breakdown is as follows: 46.3% Hispanic/Latino, 33.6% African American, 18.6% Caucasian, and 1.5% Other (merged from several smaller racial group categories). To account for relatedness within this cohort, we used PLINK v1.9 [31] to identify pairs of directly related individuals (PI-HAT scores > 0.25). From these pairs, we randomly selected one from each to exclude (n=612), resulting in 10,510 individuals used for the analyses.

2.3 Variant annotation

We adapted the protocol used by Glicksberg et al. to annotate genotyped variants as LOF [32]. Briefly, we ran 906,917 genotyped variants through three different public annotation tools, namely Variant Annotation Tool (VAT) [33], ANNOVAR (v. 2015Apr14) [34], and SnpEff (v. 3.6) [35]. Following established procedures [19,36], we restricted output from these annotators to "High" effect and relevant types: stop gain, frameshift/indel, and splice site. We performed further quality control by excluding variants that were in the final exon of the transcript and those with >2% alternate allele frequency. To enhance confidence of these annotations, we only included variants that passed these criteria in at least two out of three of the annotators for at least one overlapping transcript. Following these steps, we derived 2,117 putative LOF variants in 1,665 genes. For the purposes of this study, we collapsed variants to the gene level. When intersecting with the 35 neurodevelopmental genes of interest, there were five (*IL33, INMT, MAP9, LCN2, LRG1*) overlapping with at least one LOF variant (**Table 1**) used for subsequent analyses.

Table 1. Loss-of-function variants mapped to five of 35 putative neuroplasticity genes

Gene	Chr	Position	RSID	Ref	Alt	VAT effect type	snpEff effect type	ANNOVAR effect type	BioMe aaf	ExAC aaf
IL33	9	6253575	rs145735086	G	T	.	STOP_GAINED	stopgain	0.000143	0.000025
INMT	7	30791800	rs190694809	C	T	stop_gained	STOP_GAINED	stopgain	0.000190	0.000033
MAP9	4	156294610	rs149881598	T	G	splice_site	SPLICE_SITE_ACCEPTOR	.	0.000476	0.000190
LCN2	9	130913998	rs139329518	T	C	splice_site	SPLICE_SITE_DONOR	.	0.001332	0.000289
LRG1	19	4540004	rs116733978	G	A	stop_gained	.	stopgain	0.000999	0.000189

2.4 Neurodevelopmental disease phenotyping

Disease diagnoses are encoded in the Mount Sinai Hospital de-identified EMR as International Classification of Diseases (ICD)-9 codes. In order to increase power for our analyses, we mapped these codes to the Clinical Classification Software (CCS; https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp) for ICD-9-CM single level categories. In total, there are 283 single level categories. As our focus is neurodevelopment and the nervous system, we restricted this list to 38 disease categories where the nervous system is considered the primary affected organ, specifically: "Meningitis (except that caused by tuberculosis or sexually transmitted disease)",

"Inflammation; infection of eye (except that caused by tuberculosis or sexually transmitted disease)", "Other CNS infection and poliomyelitis", "Otitis media and related conditions", "Cancer of brain and nervous system", "Delirium, dementia, and amnestic and other cognitive disorders", "Alcohol-related disorders", "Substance-related disorders", "Schizophrenia and other psychotic disorders", "Mood disorders", "Anxiety disorders", "Personality disorders", "Screening and history of mental health and substance abuse codes", "Developmental disorders", "Adjustment disorders", "Attention-deficit, conduct, and disruptive behavior disorders", "Impulse control disorders, NEC", "Other nervous system disorders", "Other hereditary and degenerative nervous system conditions", "Parkinson`s disease", "Headache; including migraine", "Multiple sclerosis", "Paralysis", "Epilepsy; convulsions", "Acute cerebrovascular disease", "Coma; stupor; and brain damage", "Spinal cord injury", "Other eye disorders", "Retinal detachments; defects; vascular occlusion; and retinopathy", "Glaucoma", "Cataract", "Blindness and vision defects", "Other ear and sense organ disorders", "Conditions associated with dizziness or vertigo", "Transient cerebral ischemia", "Nervous system congenital anomalies", "Poisoning by psychotropic agents", "Suicide and intentional self-inflicted injury".

2.5 LOF gene and disease association analysis

With the genotype and disease data processed, we assessed associations between LOF in these five putative plasticity genes and the 38 nervous system-related disease categories of interest (**Figure 1b**). Specifically, we performed a logistic regression for all gene-disease combinations for which there were at least three carriers of the gene afflicted with the disease. We also controlled for demography in the form of age, self-reported sex, and genetic ancestry using Principal Component Analysis (PCA) in the form of the first five Principal Components, which constituted the majority of variance explained (Eq. 1). The use of PCA on genetic data for determining and controlling for genetic ancestry in association studies is well established [37]. We focused on the significance of the gene term and magnitude and direction of the associated β_l value, which represents effect size after controlling for other covariates (positive values indicate increased risk and vice versa).

$$P(\text{disease} \mid \beta_0 + \beta_l \cdot \text{gene} + \beta_g \cdot \text{sex} + \beta_a \cdot \text{age} + \beta_{pc1} \cdot \text{PC1} \ldots + \ldots \beta_{pc5} \cdot \text{PC5}) \quad (1)$$

where disease is a binary Yes/No outcome, gene is binary Yes/No indicating presence of LoF mutation, age is a continuous constant per year, sex is binary piecewise Female/Male, and PC# is continuous.

3. Results

3.1 Identifying putative neuroplasticity genes

To identify putative plasticity genes, we generated transcriptional signatures of plasticity by comparing primary visual cortex transcriptomes of juvenile wild-type or adult *Lynx1*$^{-/-}$ compared to adult wild-type mice yielding 176 and 98 differentially expressed genes (**Figure 1a**). *Lynx1*$^{-/-}$ mice have elevated, juvenile-like plasticity [15] and were used to control for non-plasticity aspects of the

juvenile signature. We defined putative plasticity genes as the 35 shared between the two signatures (Fisher Exact Test: OR=37.1, 95% CI = 23.8–58.0, $p < 2.2 \times 10^{-16}$, this statistic reproduced as in [14]). Interestingly, using gene set enrichment we found that these genes are predominantly enriched for immune processes, including gene sets related to neutrophil degranulation, defense response to fungus, immune cell chemotaxis, apoptotic pathways, and cytokine production (**Table 2**).

Table 2. Enrichment of biological pathways across 35 putative neuroplasticity genes reveals inflammatory pathways

Term	P-value	Z-score	Comb. Score	Genes
neutrophil degranulation	7.8E-04	-4.99	22.6	LRG1;LCN2;PDAP1;S100A9;S100A8
defense response to fungus	1.0E-04	-3.45	20.3	S100A9;S100A8
antimicrobial humoral response	6.7E-05	-2.68	16.1	LCN2;S100A9;S100A8
negative regulation of transcription from RNA polymerase II promoter	5.9E-03	-5.03	15.6	MTF2;ARID5B;TBL1X;CPEB3
activation of cysteine-type endopeptidase activity involved in apoptotic process	1.3E-04	-2.46	14.5	ACER2;S100A9;S100A8
cytokine production	4.9E-05	-2.01	12.0	S100A9;S100A8
regulation of cytoskeleton organization	3.9E-04	-2.34	11.5	S100A9;S100A8
leukocyte migration involved in inflammatory response	3.5E-05	-1.87	11.2	S100A9;S100A8
positive regulation of intrinsic apoptotic signaling pathway	6.3E-04	-2.34	10.8	S100A9;S100A8

3.2 *LOF variants in putative plasticity genes confer risk for neurodevelopmental and nervous system-related disorders*

Applying an integrative genomics approach (**Figure 1c**), we determined that five of 35 putative plasticity genes (*IL33, INMT, MAP9, LCN2, LRG1*) contained a LOF variant (**Table 1**) that had been genotyped in the Bio*Me* biobank and were included in subsequent analyses. Using a disease carrier minimum frequency of three, we were able to perform 27 association tests for three genes (*MAP9, LCN2, LRG1*) across 15 nervous system-related diseases. We found that two genes, *LRG1* and *LCN2,* conferred risk for five nervous system diseases (**Table 3**). Strikingly, two of these diseases, schizophrenia and epilepsy, have putative etiologies based in perinatal and childhood neurodevelopment (*LRG1* - schizophrenia: $\beta = 1.27$, $p = 0.04$; *LCN2* - epilepsy: $\beta = 1.22$, $p = 0.03$). Additionally, we identified a trending association between *MAP9* and blindness and vision defects *(*$\beta = 1.15$, $p = 0.08$).

Table 3. Putative neuroplasticity genes confer risk for neurodevelopmental and brain-related diseases

Disease	Gene	Disease Carriers	Carriers	Disease	P-value	Beta
Schizophrenia and other psychotic disorders	LRG1	3	21	407	0.042	1.27
Epilepsy; convulsions	LCN2	4	28	455	0.025	1.22
Conditions associated with dizziness or vertigo	LCN2	10	28	1646	0.013	1.01
Blindness and vision defects	LRG1	8	21	1745	0.042	0.93
Anxiety disorders	LCN2	8	28	1672	0.049	0.83

4. Discussion

We demonstrate an innovative use of genes relevant to neuroplasticity to identify potential human neurodevelopmental risk genes. By applying an integrative genomics approach, we identified *LRG1* and *LCN2* as putative plasticity genes associated with the neurodevelopmental diseases epilepsy and schizophrenia in a human population. These genes are correlated to experience-dependent neural plasticity across two mouse models, suggesting that LOF in human may confer risk for neurodevelopmental disorders by disrupting plasticity. Moreover, these genes are regulated by inflammation via lipopolysaccharide, which also disrupts experience-dependent plasticity in juvenile mouse [14], suggesting LOF in these genes may disrupt a component of neural-immune interaction to confer risk for human neurodevelopment. Consistent with that perspective, schizophrenia and epilepsy have numerous aberrations in immune function [38,39] and neural plasticity [40–42] and this work suggests that a nexus of these aberrations may be juvenile experience-dependent plasticity, which is increasingly postulated as an important locus of neurodevelopmental risk [7,8].

LRG1 has been previously identified as dysregulated in the choroid plexus of individual's with schizophrenia [43] and marks early granulocyte maturation [44], consistent with gene set enrichments indicating the 35 putative plasticity genes are enriched for granulocyte function (see Table 2). In contrast, antipsychotics appear to induce an immature granulocytic phenotype [45]. This has generally been considered a side-effect (and is separate from potentially fatal agranulocytosis, as induced by clozapine [46]), but neutrophils in drug-free individuals with schizophrenia generate elevated reactive oxygen species (ROS) [47,48] and ROS levels can be normalized by antipsychotics [49,50]. It should be noted, however, that one study found antipsychotics did not decrease ROS in patients with schizophrenia [51]. In animal models and postmortem brains of individuals with schizophrenia, there is evidence of elevated oxidative stress associated with the parvalbumin interneuron cell type [52]. Moreover, genetically reducing the antioxidant glutathione specifically in parvalbumin cells (which elevates ROS) leads to dysregulated critical period plasticity [53]. Therefore, we speculate that neutrophils may be a source of oxidative stress (i.e. ROS) in schizophrenia and that the suppressive effective of antipsychotics on neutrophil function may in fact be a therapeutic phenomenon. Together this suggests neutrophils and *LRG1* as previously unrecognized components of schizophrenia pathophysiology and as putative therapeutic targets that should be explored further.

LCN2 is an important cell-autonomous marker of astrocyte activation - a phenotype that shifts astrocytes away from their resting-state role in maintaining neural circuit homeostasis to an active watchfulness against cellular damage and other forms of danger. In epilepsy, abberations in astrocytic regulation of neurotransmitters (i.e. glutamate and GABA) and ions (i.e. K^+) likely contribute to excitotoxicity and reduced threshold for induction of seizure [54]. Therefore, we hypothesize that LOF mutations in *LCN2* could cause astrocytes to exit their normal resting-state wherein they homeostatically support neural equilibrium, leading to chronic neurotransmitter and ionic dysregulations. Moreover, Lcn2 is an exogenous activator of microglia [55] and microglia are critically important to juvenile experience-dependent plasticity *per se* [56]. Together, this suggests mutations in *LCN2* may confer risk for epilepsy via dysregulation of multiple glial types to produce a multi-faceted disruption across neurodevelopment and suggests glia may be a promising therapeutic target at the intersection of inflammation and plasticity in epilepsy.

Given that neurodevelopmental disease is highly polygenic, it may be unsurprising that in addition to epilepsy, *Lcn2* is dysregulated in the *Disc1*-L100P mouse model of schizophrenia [57]. There is a growing but unclear role of both microglia and astrocytes in schizophrenia [58]. Functionally, activated microglia go on to secrete soluble inflammatory cytokines C1q, Tnf, and Il1-α to activate astrocytes, which then secrete Lcn2 and an unknown toxic substance that inhibits synaptic efficacy and kills neurons [59]. We speculate that a microglia-astrocyte-neural circuit may be involved in plasticity aberrations in schizophrenia and future work should explore this possibility. Consistent with this hypothesis, *Gfap* expression is elevated in the *Disc1*-L100P model, indicating a reactive astrocyte phenotype [57]. Moreover, sodium valproate normalized *Gfap* and *Lcn2* levels, as well as functional correlates of schizophrenia, indicating *Lcn2* may be a novel drug target or biomarker of successful treatment in schizophrenia. More generally, astrocytes and microglia may be an inflammatory hub in epilepsy and schizophrenia that could be targeted for therapeutic intervention.

We provide here a highly systematic and high-throughput integrative genomics approach to identify neurodevelopmental risk genes. This approach is complementary to existing approaches including GWAS, family-based, and hereditary (e.g. twin, adoption) studies. Those approaches have been extremely useful for identifying risk variants in a disease-focused manner; our integrative genomics approach extends on these by liberating from disease-centric constraints to orient the analysis on a function-based approach to identify relevant risk genes across multiple diseases. Implementing this approach here, we find that two genes implicated in neural plasticity, *LRG1* and *LCN2*, are associated with the neurodevelopmental diseases epilepsy and schizophrenia and may play a pathophysiological role at the nexus of immune-brain function. As such, we believe these genes may be biomarkers for such neurodevelopmental-related diseases and candidates for drug targets. On the other hand, there are a few caveats and limitations to our integrative genomics approach. We used two models of plasticity (juvenile and *Lynx1*$^{-/-}$), but transcriptional changes in other models could further contribute to the identification of neurodevelopmental risk genes in human. Limiting to the models used here could exclude genes relevant to neuroplasticity (i.e. false negatives). Additionally, though we used a strict FDR threshold to identify putative plasticity genes, the possibility of including genes that are not directly relevant to plasticity (i.e. false positives) is possible given the variable nature of gene expression profiling. In addition, the specific molecular function of these genes in plasticity is not yet established, making interpretation of their role in developmental neuroplasticity *per se* more challenging. Moreover since these genes were identified using differential expression analysis, making interpretations of LOF in a given gene challenging and robust experimental work should follow. Given the input set of plasticity genes used in this study, we were limited by the number of genes that were genotyped and the number of LOF variants in these genes. Separately, there are known issues surrounding the accuracy of defining a disease by ICD codes. While robust, multimodal electronic phenotyping algorithms exist for many diseases (e.g. PheKB; https://phekb.org/), we utilized ICD-based definitions for diseases (via CCS) because there are not many algorithms that exist for our disease domain of interest (nervous-system and neurodevelopment). Finally, this study used only a single cohort (Mount Sinai Bio*Me*) and given the relatively low sample sizes for the diseases for which we identified LOF variants in putative plasticity genes (see Table 3) we considered a nominal *p* value threshold of 0.05 as appropriate for discovery. Follow up studies in larger, independent cohorts using a multiple test correction approach, as well as

functional experiments to elucidate the specific neurobiological relevance, is critical to validate these findings. We further discuss potential approaches to address these issues in follow-up studies within the next section.

5. Conclusions and Future Directions

This study provides significant impact to the field by identifying unrecognized neurodevelopment risk genes for schizophrenia and epilepsy through a novel systematic approach leveraging Mount Sinai's BioMe BioBank and linked Mount Sinai Hospital's Electronic Medical Record (EMR) data. This integrative genomics approach facilitates high-throughput identification of LOF risk variants that may have a deleterious impact on neurodevelopment and the findings set the groundwork for functional studies to determine the mechanisms by which these novel risk genes disrupt neurodevelopment and to investigate their utility for therapeutic discovery. Using putative plasticity genes as the seed genes to identify neurodevelopmental risk genes immediately sets the stage to rigorously test the hypothesis that these genes play a role in childhood neurodevelopmental. Using the ocular dominance animal model of developmental neuroplasticity [13] from which the plasticity genes were derived allows investigators to rapidly return to the mouse to test the effect of gene perturbation in neurodevelopment and neuroplasticity.

There are several future directions we will pursue to extend and further assess the implications of our findings. The relatively low sample sizes of nervous-system related diseases in our cohort (for example, see Table 3) coupled with the rare nature of these LOF mutations, limits power to detect associations. As such, we plan to perform a cross-validation experiment using genotype and clinical data for the 500,000 individuals in the UKBioBank (http://www.ukbiobank.ac.uk/). Additionally, in the hopes of exploring associations for the entire original set of 35 putative neuroplasticity genes, we will leverage the UK10K (http://www.uk10k.org) whole exome sequencing data to identify putative LOF variants for these genes within the neurodevelopmental cohort (N=3,000). In addition, we will increase our collection of genes related to neuroplasticity using other models, such as calorie restriction-induced plasticity [60], exercise-induced plasticity [61], drug-induced plasticity [62], as well as other plasticity-enhancing gene perturbation models, depending on available transcriptional data. Relatedly, we aim to extend these analyses to confirmed plasticity genes whose molecular mechanisms in plasticity are well-established, to yield a hypothesis-driven iteration of our approach. While it is important to increase the number of starting plasticity genes and use larger quantities of human data, it would be additionally valuable to reassess the associations made here using PheKB algorithms for nervous-system related diseases to address the limitations of ICD-code based phenotyping. Finally, we expect this integrative genomics approach will be generalizable to identify risk genes and facilitate focused biological inquiry in other disease contexts to enable drug target and biomarker identification.

6. Acknowledgments

We would like to sincerely thank the Mount Sinai Data Warehouse for facilitating data accessibility and the Mount Sinai Scientific Computing team for infrastructural support. We would

also like to thank the Charles Bronfman Institute of Personalized Medicine for collecting and housing the biobank genetic data as well as Eli Stahl and Douglas Ruderfer for performing quality control on the genotyping array.

This work was funded by a Traineeship, National Institute of Child Health and Human Development - Interdisciplinary Training in Systems and Developmental Biology and Birth Defects Grant T32H-D0-75735 (to M.R.S.); the Mindich Child Health and Development Institute Pilot Fund (to J.T.D. and H.M.); the Knights Templar Eye Foundation (to H.M.); the March of Dimes (to H.M.); the Whitehall Foundation (to H.M.); the Harris Center for Precision Wellness (to J.T.D); and National Institutes of Health Grants P30-ES-023515 (to J.T.D. and H.M.), R01-DK-098242 (to J.T.D.), and R01-EY- 024918, R01-EY-026053, and R21 MH106919 (to H.M.).

References

1. C. A. Boyle, S. Boulet, L. A. Schieve, R. A. Cohen, et al., *Pediatrics*, peds.2010-2989 (2011).
2. P. Grandjean and P. J. Landrigan, *Lancet Neurol.* **13**, 330–338 (2014).
3. J. S. Johnson and E. L. Newport, *Cognit. Psychol.* **21**, 60–99 (1989).
4. T. P. Nikolopoulos, G. M. O'Donoghue, and S. Archbold, *The Laryngoscope* **109**, 595–599 (1999).
5. T. L. Lewis and D. Maurer, *Dev. Psychobiol.* **46**, 163–183 (2005).
6. C. A. Nelson, C. H. Zeanah, N. A. Fox, P. J. Marshall, et al., *Science* **318**, 1937–1940 (2007).
7. J. J. LeBlanc and M. Fagiolini, *Neural Plast.* **2011** (2011).
8. A. E. Takesian and T. K. Hensch, in *Prog. Brain Res.* **207**, M. N. and T. M. V. V. Michael M. Merzenich, Ed. (Elsevier, 2013).
9. D. Tropea, E. Giacometti, N. R. Wilson, C. Beard, et al., *Proc. Natl. Acad. Sci.* **106**, 2029–2034 (2009).
10. K. Yashiro, T. T. Riday, K. H. Condon, A. C. Roberts, et al., *Nat. Neurosci.* **12**, 777–783 (2009).
11. E. G. Harlow, S. M. Till, T. A. Russell, L. S. Wijetunge, et al., *Neuron* **65**, 385–398 (2010).
12. T. N. Wiesel and D. H. Hubel, *J. Neurophysiol.* **26**, 1003–1017 (1963).
13. J. A. Gordon and M. P. Stryker, *J. Neurosci.* **16**, 3274–3286 (1996).
14. M. R. Smith, P. Burman, M. Sadahiro, B. A. Kidd, et al., *eNeuro* **3**, ENEURO.0240-16.2016 (2016).
15. H. Morishita, J. M. Miwa, N. Heintz, and T. K. Hensch, *Science* **330**, 1238–1240 (2010).
16. F. S. Alkuraya, *Genome Med.* **7** (2015).
17. F. E. Dewey, M. F. Murray, J. D. Overton, L. Habegger, et al., *Science* **354** (2016).
18. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, et al., *Nature* **467**, 1061–1073 (2010).
19. D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, et al., *Science* **335**, 823–828 (2012).
20. M. Vergeer, L. R. Brunham, J. Koetsveld, J. K. Kruit, et al., *Diabetes Care* **33**, 869–874 (2010).
21. D. Grozeva, K. Carss, O. Spasic-Boskovic, M. J. Parker, et al., *Am. J. Hum. Genet.* **94**, 618–624 (2014).
22. A. B. Jørgensen, R. Frikke-Schmidt, B. G. Nordestgaard, and A. Tybjærg-Hansen, *N. Engl. J. Med.* **371**, 32–41 (2014).
23. *N. Engl. J. Med.* **371**, 22–31 (2014).
24. J. Flannick, G. Thorleifsson, N. L. Beer, S. B. R. Jacobs, et al., *Nat. Genet.* **46**, 357–363 (2014).
25. N. O. Stitziel and S. Kathiresan, *Trends Cardiovasc. Med.* **27**, 352–359 (2017).
26. M. J. Graham, R. G. Lee, T. A. Brandt, L.-J. Tai, et al., *N. Engl. J. Med.* **377**, 222–232 (2017).
27. A. Diamantopoulou, Z. Sun, J. Mukai, B. Xu, et al., *Proc. Natl. Acad. Sci.* **114**, E6127–E6136 (2017).
28. G. K. Smyth, in *Bioinforma. Comput. Biol. Solut. Using R Bioconductor* (Springer, New York, 2005).
29. F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, et al., *Bioinformatics* **22**, 2825–2827 (2006).

30. E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, et al., *BMC Bioinformatics* **14**, 128 (2013).
31. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, et al., *GigaScience* **4** (2015).
32. B. S. Glicksberg, L. Amadori, N. K. Akers, K. Sukhavasi, et al., *Rev.*
33. G. T. Wang, B. Peng, and S. M. Leal, *Am. J. Hum. Genet.* **94**, 770–783 (2014).
34. K. Wang, M. Li, and H. Hakonarson, *Nucleic Acids Res.* **38**, e164 (2010).
35. P. Cingolani, A. Platts, L. L. Wang, M. Coon, et al., *Fly (Austin)* **6**, 80–92 (2012).
36. A. H. Li, A. C. Morrison, C. Kovar, L. A. Cupples, et al., *Nat. Genet.* **47**, 640–642 (2015).
37. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, et al., *Nat. Genet.* **38**, 904–909 (2006).
38. J. C. Leza, B. García-Bueno, M. Bioque, C. Arango, et al., *Neurosci. Biobehav. Rev.* **55**, 612–626 (2015).
39. A. Vezzani, J. French, T. Bartfai, and T. Z. Baram, *Nat. Rev. Neurol.* **7**, 31–40 (2011).
40. D. Ben-Shachar and D. Laifenfeld, *Int. Rev. Neurobiol.* **59**, 273–296 (2004).
41. H. E. Scharfman, *The Neuroscientist* **8**, 154–173 (2002).
42. J. W. Swann and J. J. Hablitz, *Ment. Retard. Dev. Disabil. Res. Rev.* **6**, 258–267 (2000).
43. S. Kim, Y. Hwang, D. Lee, and M. J. Webster, *Transl. Psychiatry* **6**, e964 (2016).
44. L. C. O'Donnell, L. J. Druhan, and B. R. Avalos, *J. Leukoc. Biol.* **72**, 478–485 (2002).
45. J. M. Delieu, M. Badawoud, M. A. Williams, R. W. Horobin, et al., *J. Psychopharmacol. (Oxf.)* **15**, 191–194 (2001).
46. J. M. J. Alvir, J. A. Lieberman, A. Z. Safferman, J. L. Schwimmer, et al., *N. Engl. J. Med.* **329**, 162–167 (1993).
47. P. Sirota, R. Gavrieli, and B. Wolach, *Psychiatry Res.* **121**, 123–132 (2003).
48. Y. Melamed, P. Sirota, D. R. Dicker, and P. Fishman, *Psychiatry Res.* **77**, 29–34 (1998).
49. F. Péters, T. Franck, M. Pequito, G. De La REBIÈRE, et al., *J. Vet. Pharmacol. Ther.* **32**, 541–547 (2009).
50. F. Vargas, V. Chávez, and K. Pérez, *Rev. Colomb. Cienc. Quím. - Farm.* **38**, 5–18 (2009).
51. M. Cosentino, A. Fietta, E. Caldiroli, F. Marino, et al., *Prog. Neuropsychopharmacol. Biol. Psychiatry* **20**, 1117–1129 (1996).
52. J.-H. Cabungcal, P. Steullet, H. Morishita, R. Kraftsik, et al., *Proc. Natl. Acad. Sci.* **110**, 9130–9135 (2013).
53. H. Morishita, J.-H. Cabungcal, Y. Chen, K. Q. Do, et al., *Biol. Psychiatry* **78**, 396–402 (2015).
54. N. C. de Lanerolle, T.-S. Lee, and D. D. Spencer, *Neurotherapeutics* **7**, 424–438 (2010).
55. E. Jang, S. Lee, J.-H. Kim, J.-H. Kim, et al., *FASEB J.* **27**, 1176–1190 (2013).
56. G. O. Sipe, undefined R. L. Lowery, M.-È. Tremblay, E. A. Kelly, et al., *Nat. Commun.* **7**, 10905 (2016).
57. T. V. Lipina, F. N. Haque, A. McGirr, P. C. Boutros, et al., *PLOS ONE* **7**, e51562 (2012).
58. H.-G. Bernstein, J. Steiner, and B. Bogerts, *Expert Rev. Neurother.* **9**, 1059–1071 (2009).
59. S. A. Liddelow, K. A. Guttenplan, L. E. Clarke, F. C. Bennett, et al., *Nature* **541**, 481–487 (2017).
60. M. Spolidoro, L. Baroncelli, E. Putignano, J. F. Maya-Vetencourt, et al., *Nat. Commun.* **2**, 320 (2011).
61. E. Kalogeraki, F. Greifzu, F. Haack, and S. Löwel, *J. Neurosci.* **34**, 15476–15481 (2014).
62. D. Silingardi, M. Scali, G. Belluomini, and T. Pizzorusso, *Eur. J. Neurosci.* **31**, 2185–2192 (2010).

Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders

Gregory P. Way

Genomics and Computational Biology Graduate Program,
University of Pennsylvania,
Philadelphia, PA 19104, USA
E-mail: gregway@mail.med.upenn.edu

Casey S. Greene*

Department of Systems Pharmacology and Translational Therapeutics
University of Pennsylvania,
Philadelphia, PA 19104, USA
E-mail: csgreene@mail.med.upenn.edu

The Cancer Genome Atlas (TCGA) has profiled over 10,000 tumors across 33 different cancer-types for many genomic features, including gene expression levels. Gene expression measurements capture substantial information about the state of each tumor. Certain classes of deep neural network models are capable of learning a meaningful latent space. Such a latent space could be used to explore and generate hypothetical gene expression profiles under various types of molecular and genetic perturbation. For example, one might wish to use such a model to predict a tumor's response to specific therapies or to characterize complex gene expression activations existing in differential proportions in different tumors. Variational autoencoders (VAEs) are a deep neural network approach capable of generating meaningful latent spaces for image and text data. In this work, we sought to determine the extent to which a VAE can be trained to model cancer gene expression, and whether or not such a VAE would capture biologically-relevant features. In the following report, we introduce a VAE trained on TCGA pan-cancer RNA-seq data, identify specific patterns in the VAE encoded features, and discuss potential merits of the approach. We name our method "Tybalt" after an instigative, cat-like character who sets a cascading chain of events in motion in Shakespeare's "*Romeo and Juliet*". From a systems biology perspective, Tybalt could one day aid in cancer stratification or predict specific activated expression patterns that would result from genetic changes or treatment effects.

Keywords: Deep Learning; Gene Expression; Variational Autoencoder, The Cancer Genome Atlas

1. Introduction

Deep learning has improved the state of the art in many domains, including image, speech, and text processing, but it has yet to make significant enough strides in biomedicine for it to be considered transformative.[1] Nevertheless, several studies have revealed promising results. For instance, Esteva *et al.* used convolutional neural networks (CNNs) to diagnose melanoma from skin images and Zhou and Troyanskaya trained deep models to predict the impact of non-

*To whom correspondence should be addressed.

coding variants.[2,3] However, several domain specific limitations remain. In contrast to image or text data, validating and visualizing learning in biological datasets is particularly challenging. There is also a lack of ground truth labels in biomedical domains, which often limits the efficacy of supervised models. New unsupervised deep learning approaches such as generative adversarial nets (GANs) and variational autoencoders (VAEs) harness the modeling power of deep learning without the need for accurate labels.[4–6] Unlike traditional CNNs, which model data by minimizing inaccurate class predictions, autoencoder models, including VAEs, learn through data reconstruction. Reconstructing gene expression input data using autoencoder frameworks has been previously shown to reveal novel biological patterns.[7–9]

VAEs and GANs are generative models, which means they learn to approximate a data generating distribution. Through approximation and compression, the models have been shown to capture an underlying data manifold — a constrained, lower dimensional space where data is distributed — and disentangle sources of variation from different classes of data.[10,11] For instance, a recent group trained adversarial autoencoders on chemical compound structures and their growth inhibiting effects in cancer cell lines to learn manifold spaces of effective small molecule drugs.[12,13] Additionally, Rampasek *et al.* trained a VAE to learn a gene expression manifold of reactions of cancer cell lines to drug treatment perturbation.[14] The theoretical basis for modeling cancer using lower dimensional manifolds is established, as it has been previously hypothesized that cancer exists in "basins of attraction" defined by specific pathway aberrations that drive cells toward cancer states.[15] These states could be revealed by data driven manifold learning approaches.

The Cancer Genome Atlas (TCGA) has captured several genomic measurements for over 10,000 different tumors across 33 cancer-types.[16] TCGA has released this data publicly, enabling many secondary analyses, including the training of deep models that predict survival.[17] One data type amenable to modeling manifold spaces is RNA-seq gene expression because it can be used as a proxy to describe tumor states and the downstream consequences of specific molecular aberration. Biology is complex, consisting of multiple nonlinear and often redundant connections among genes, and when a specific pathway aberration occurs, the downstream response to the perturbation is captured in the transcriptome. In the following report, we extend the autoencoder framework by training and evaluating a VAE on TCGA RNA-seq data. We aim to demonstrate the validity and specific latent space benefits of a VAE trained on gene expression data. We do not aim to comprehensively profile all learned pan-cancer VAE features nor survey clinical implications. We also do not compare our approach to alternate dimensionality reduction algorithms, but instead present our model as an additional tool in the toolkit for extracting knowledge from gene expression. We shall name this model "Tybalt".

2. Methods

2.1. *Model Summary*

VAEs are data driven, unsupervised models that can learn meaningful latent spaces in many contexts. In this work, we aim to build a VAE that compresses gene expression features and reveals a biologically relevant latent space. The VAE is based on an autoencoding framework, which can discover nonlinear explanatory features through data compression and nonlinear

activation functions. A traditional autoencoder consists of an encoding phase and a decoding phase where input data is projected into lower dimensions and then reconstructed.[18] An autoencoder is deterministic, and is trained by minimizing reconstruction error. In contrast, VAEs are stochastic and learn the *distribution* of explanatory features over samples. VAEs achieve these properties by learning two distinct latent representations: a mean and standard deviation vector encoding. The model adds a Kullback-Leibler (KL) divergence term to the reconstruction loss, which also regularizes weights through constraining the latent vectors to match a Gaussian distribution. In a VAE, these two representations are learned concurrently through the use of a reparameterization trick that permits a back propagated gradient.[4] Importantly, new data can be projected onto an existing VAE feature space enabling new data to be assessed.

2.2. *Model Implementation*

VAEs have been shown to generate "blurry" data compared with other generative models, including GANs, but VAEs are also generally more stable to train.[19] We trained our VAE model, Tybalt, with the following architecture: 5,000 input genes encoded to 100 features and reconstructed back to the original 5,000 (Figure 1A). The 5,000 input genes were selected based on highest variability by median absolute deviation (MAD) in the TCGA pan-cancer dataset.

We initially trained Tybalt without batch normalization,[20] but observed that when we included batch normalization in the encoding step, we trained faster and with heterogeneous feature activation. Batch normalization in machine learning is distinct from normalizing gene expression batches together in data processing. In machine learning, batch normalization adds additional feature regularization by scaling activations to zero mean and unit variance, which has been observed to speed up training and reduce batch to batch variability thus increasing generalizability. We trained Tybalt with an Adam optimizer,[21] included rectified linear units[22] and batch normalization in the encoding stage, and sigmoid activation in the decoding stage. We built Tybalt in Keras (version 2.0.6)[23] with a TensorFlow backend (version 1.0.1).[24] For more specific VAE illustrations and walkthroughs refer to an extended tutorial[25] and these intuitive blog posts.[26,27]

2.3. *Parameter Selection*

We performed a parameter sweep over batch size (50, 100, 128, 200), epochs (10, 25, 50, 100), learning rates (0.005, 0.001, 0.0015, 0.002, 0.0025) and warmups (κ) (0.01, 0.05, 0.1, and 1). κ controls how much the KL divergence loss contributes to learning, which effectively transitions a deterministic autoencoder to a VAE.[28,29] For instance, a $\kappa = 0.1$ would add 0.1 to a weight on the KL loss after each epoch. After 10 epochs, the KL loss will have equal weight as the reconstruction loss. We did not observe κ to influence model training (Figure 1B), so we kept $\kappa = 1$ for downstream analyses. We evaluated train and test set loss at each epoch. The test set was a random 10% partition of the full data. In general, training was relatively stable for many parameter combinations, but was consistently worse for larger batches, particularly with low learning rates. Ultimately, the best parameter combination based on validation loss was

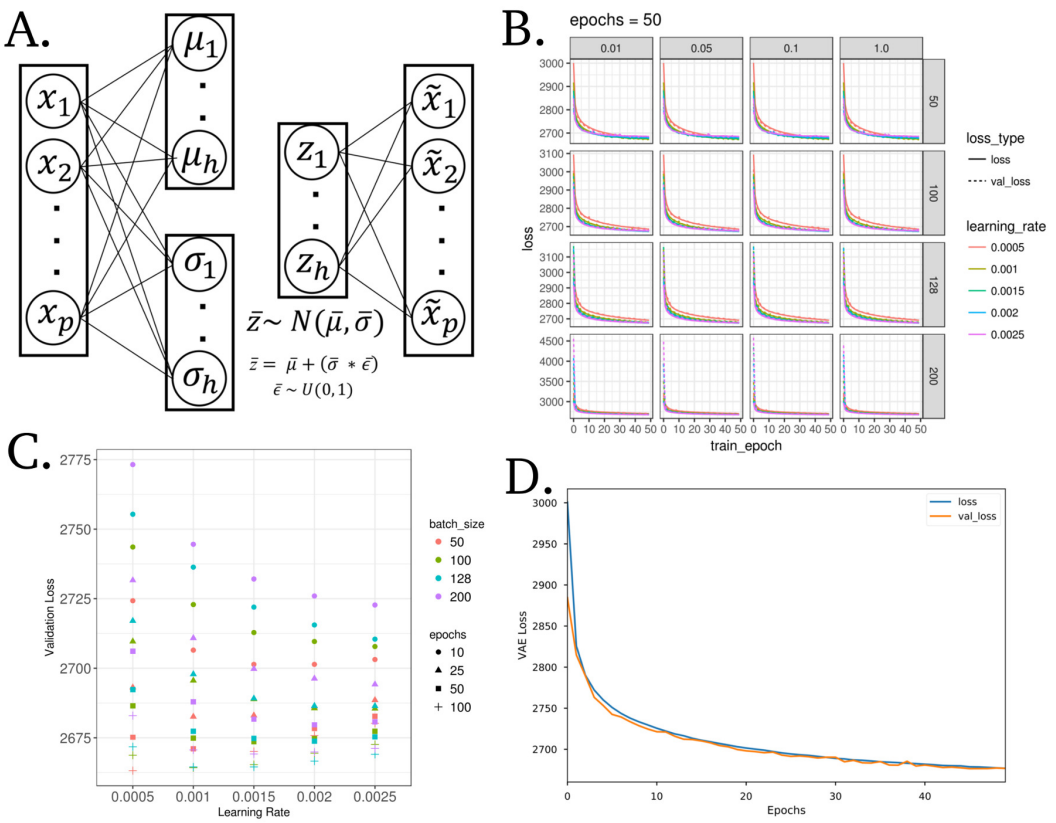


Fig. 1. *A variational autoencoder (VAE) applied to model gene expression data.* **(A)** Model wire diagram of Tybalt encoding a gene expression vector ($p = 5{,}000$) into mean (μ) and standard deviation (σ) vectors ($h = 100$). A reparameterization trick[4,5] enables learning z, which is then reconstructed back to input ($\tilde{x}$). **(B)** Training and validation VAE loss across training epochs (full pass through all training data). Shown across vertical and horizontal facets are values of κ and batch size, respectively. **(C)** Final validation loss for all parameters with $\kappa = 1$. **(D)** VAE loss for training and testing sets through optimized model training.

batch size 50, learning rate 0.0005, and 100 epochs (Figure 1C). Because training stabilized after about 50 epochs, we terminated training early. Training and testing loss across all 50 epochs is shown in Figure 1D. We performed the parameter sweep on a cluster of 8 NVIDIA GeForce GTX 1080 Ti GPUs on the PMACS cluster at The University of Pennsylvania.

2.4. *Input Data*

The input data consisted of level 3 TCGA RNA-seq gene expression data for 9,732 tumors and 727 tumor adjacent normal samples (10,459 total samples) measured by the 5,000 most variably expressed genes. The full dataset together is referred to as the pan-cancer data. The level 3 RNA-seq data consists of a preprocessed and batch-corrected gene abundance by sample matrix measured by log2(FPKM + 1) transformed RSEM values. The most variably expressed genes were defined by median absolute deviation (MAD). In total, there were 33 different cancer-types (including glioblastoma, ovarian, breast, lung, bladder cancer, etc.) profiled, each with varying number of tumors. We accessed RNA-seq data from the UCSC Xena data browser

on March 8th, 2016 and archived the data in Zenodo.[30] To facilitate training, we min-maxed scaled RNA-seq data to the range of $0-1$. We used corresponding clinical data accessed from the Snaptron web server.[31]

2.5. *Interpretation of Gene Weights*

Much like the weights of a deterministic autoencoder, Tybalt's decoder weights captured the contribution of specific genes to each learned feature.[7,8,32] For most features, the distribution of gene weights was similar: Many genes had weights near zero and few genes had high weights at each tail. In order to characterize patterns explained by selected encoded features of interest, we performed overrepresentation pathway analyses (ORA) separately for both positive and negative high weight genes; defined by greater than 2.5 standard deviations above or below the mean, respectively. We used WebGestalt,[33] with a background of the 5,000 assayed genes, to perform the analysis over gene ontology (GO) biological process terms.[34] P values are presented after an Benjamini-Hochberg FDR adjustment.

2.6. *The Latent Space of Ovarian Cancer Subtypes*

Image processing studies have shown the remarkable ability of generative models to mathematically manipulate learned latent dimensions.[35,36] For example, subtracting the image latent representation of a neutral man from a smiling man and adding it to a neutral woman, resulted in a vector associated with a smiling woman. We were interested in the extent to which Tybalt learned a manifold representation that could be manipulated mathematically to identify state transitions across high grade serous ovarian cancer (HGSC) subtypes. The TCGA naming convention of these subtypes is mesenchymal, proliferative, immunoreactive, and differentiated.[37]

To characterize the largest differences between the mesenchymal/immunoreactive and proliferative/differentiated HGSC subtypes, we performed a series of mean HGSC subtype vector subtractions in Tybalt latent space:

$$\bar{\theta}_k = \frac{\sum_{i=1}^{n} z_{i,1}(i_k = k)}{n_k}, \ldots, \frac{\sum_{i=1}^{n} z_{i,100}(i_k = k)}{n_k} \tag{1}$$

$$\bar{\theta}_{\text{immunoreactive}} - \bar{\theta}_{\text{mesenchymal}} = \bar{\theta}_{\text{immuno-mes}} \tag{2}$$

$$\bar{\theta}_{\text{differentiated}} - \bar{\theta}_{\text{proliferative}} = \bar{\theta}_{\text{diff-prolif}} \tag{3}$$

Where $(i_k = k)$ is an indicator function if sample i has membership with subtype k and z is the encoded layer. We used tumor subtype assignments provided for TCGA samples in Verhaak *et al.* 2013.[38] If Tybalt learned a biological manifold, this subtraction would result in the identification of biologically relevant features stratifying tumors of specific subtypes with a continuum of expression states.

2.7. *Enabling Exploration through Visualization*

We provide a Shiny app to interactively visualize activation patterns of encoded Tybalt features with covariate information at `https://gregway.shinyapps.io/pancan_plotter/`.

3. Results

Tybalt compressed tumors into a lower dimensional space, acting as a nonlinear dimensionality reduction algorithm. Tybalt learned which genes contributed to each feature, potentially capturing aberrant pathway activation and treatment vulnerabilities. Tybalt was unsupervised; therefore, it could learn both known and unknown biological patterns. In order to determine if the features captured biological signals, we characterized both sample- and gene-specific activation patterns.

3.1. *Tumors were encoded in a lower dimensional space*

The tumors were encoded from original gene expression vectors of 5,000 MAD genes into a lower dimensional vector of length 100. To determine if the sample encodings faithfully recapitulated large, tissue specific signals in the data, we visualized sample-specific Tybalt encoded features (z vector for each sample) by t-distributed stochastic neighbor embedding (t-SNE).[39] We observed similar patterns for Tybalt encodings (Figure 2A) as compared to $0-1$ normalized RNA-seq data (Figure 2B). Tybalt geometrically preserved well known relationships, including similarities between glioblastoma (GBM) and low grade glioma (LGG). Importantly, the recapitulation of tissue-specific signal was captured by non-redundant, highly heterogeneous features (Figure 2C). Based on the hierarchical clustering dendrogram, the features appeared to be capturing distinct signals. For instance, tumor versus normal and patient sex are large signals present in cancer gene expression, but they were distributed uniformly in the clustering solution indicating non-redundant feature activations.

3.2. *Features represent biological signal*

Our goal was to train and evaluate Tybalt on its ability to learn biological signals in the data and not to perform a comprehensive survey of learned features. Therefore, we investigated whether or not Tybalt could distinguish patient sex and patterns of metastatic activation. We determined that the model extracted patient sex robustly (Figure 3A). Feature encoding 82 nearly perfectly separated samples by sex. Furthermore, we identified a set of nodes that together identified skin cutaneous melanoma (SKCM) tumors of both primary and metastatic origin (Figure 3B).

The weights used to decode the hidden layer (z vector) back into a high-fidelity reconstruction of the input can capture important and consistent biological patterns embedded in the gene expression data.[7,8,32] For instance, there were only 17 genes needed to identify patient sex (Figure 3C). These genes were mostly located on sex chromosomes. The two positive weight genes were X inactivation genes *XIST* and *TSIX*, while the negative weight genes were mostly Y chromosome genes such as *EIF1AY*, *UTY*, and *KDM5D*. This result served as a positive control that the unsupervised model was able to construct a feature that described a clearly biological source of variance in the data.

There were several genes contributing to the two encoded features that separated the SKCM tumors (Figure 3D). Several genes existed in the high weight tails of each distribution for feature encodings 53 and 66. We performed an ORA on the high weight genes. In general, several pathways were identified as overrepresented in the set as compared to random. The

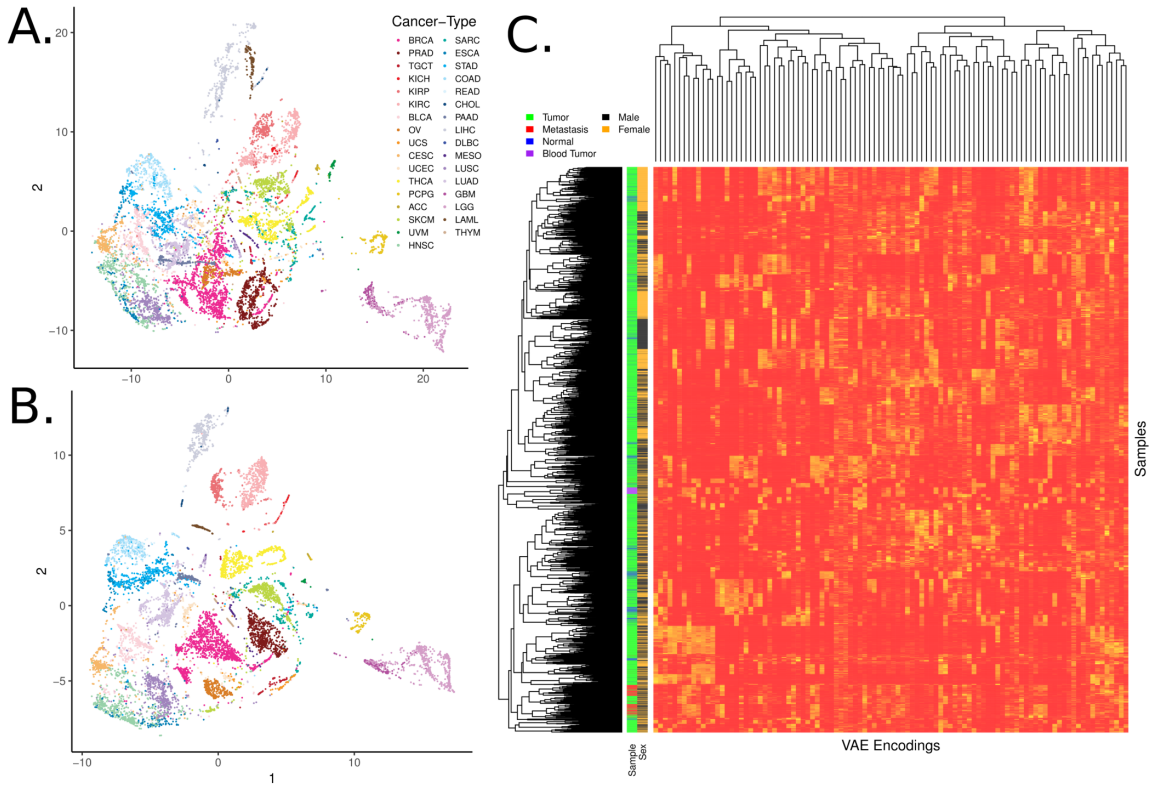


Fig. 2. *Samples encoded by a variational autoencoder retain biological signals.* **(A)** t-distributed stochastic neighbor embedding (t-SNE) of TCGA pan-cancer tumors with Tybalt encoded features. **(B)** t-SNE of 0-1 normalized gene expression features. Tybalt retains similar signals as compared to uncompressed gene expression data. **(C)** Full Tybalt encoding features by TCGA pan-cancer sample heatmap. Given on the y axis are the patients sex and type of sample.

samples had intermediate to high levels of feature encoding 53, which did not correspond to any known GO term, potentially indicating an unknown but important biological process. The samples also had intermediate to high levels of encoding 66 which implicated GO terms related to cholesterol, ethanol, and lipid metabolism including regulation of intestinal cholesterol absorption ($adj.\ p = 3.0e^{-2}$), ethanol oxidation ($adj.\ p = 4.0e^{-02}$), and lipid catabolic process ($adj.\ p = 4.0e^{-02}$). SKCM samples had consistently high activation of both encoded features, which separated them from other tumors. Nevertheless, more research is required to determine how VAE features could be best interpreted in this context.

3.3. *Interpolating the lower dimensional manifold of HGSC subtypes*

We performed an experiment to test whether or not Tybalt learned manifold differences of distinct HGSC subtypes. Previously, several groups identified four HGSC subtypes using gene expression.[37,40,41] However, the four HGSC subtypes were not consistently defined across populations; the data suggested the presence of three subtypes or fewer.[42] The study observed that the immunoreactive/mesenchymal and differentiated/proliferative tumors consistently collapsed together when setting clustering algorithms to find 2 subtypes.[42] This observation

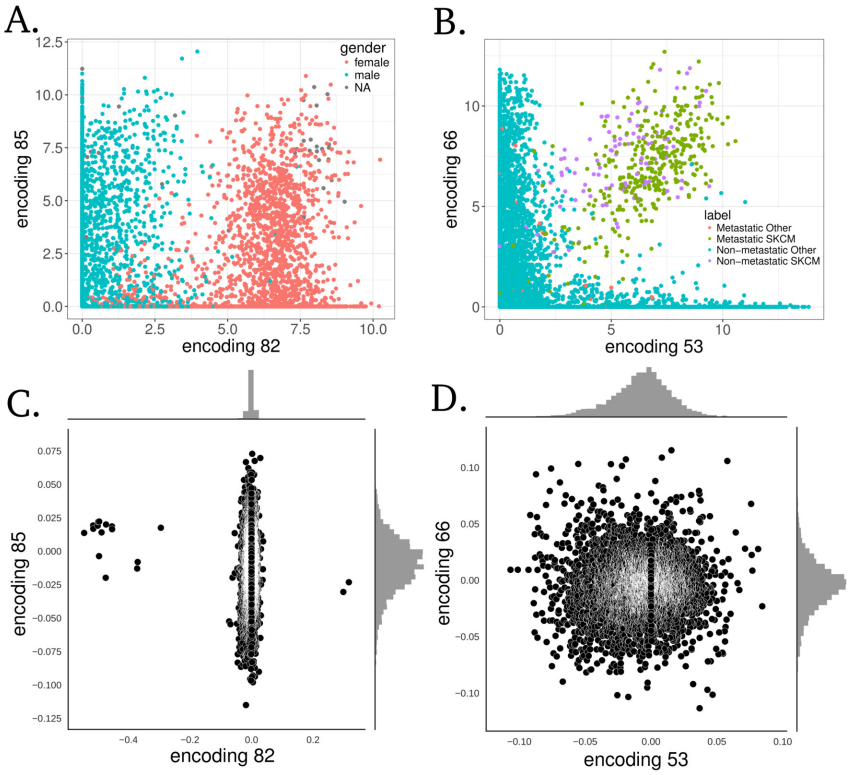


Fig. 3. *Specific examples of Tybalt features capturing biological signals.* **(A)** Encoding 82 stratified patient sex. **(B)** Together, encodings 53 and 66 separated melanoma tumors. Distributions of gene coefficients contributing to each plot above for **(C)** patient sex and **(D)** melanoma. The gene coefficients consist of the Tybalt learned weights for each feature encoding.

may suggest the presence of distinct gene expression programs existing on an activation spectrum driving differences in these subtypes. Therefore, we hypothesized that Tybalt would learn the manifold of gene expression spectra existing in differential proportions across these subtypes.

The largest feature encoding difference between the mean HGSC mesenchymal and the mean immunoreactive subtype ($\bar{\theta}_{\text{immuno-mes}}$) was encoding 87 (Figure 4A). Encoding 77 and encoding 56 (Figure 4B) also distinguished the mesenchymal and immunoreactive subtypes. The largest feature encoding differences between the mean proliferative and the mean differentiated subtype ($\bar{\theta}_{\text{diff-prolif}}$) were contributed by encoding 79 (Figure 4C) and encoding 38 (Figure 4D). Interestingly, encoding 38 had high mean activation in both the immunoreactive and differentiated subtypes.

The mesenchymal subtype had the highest encoding 87 activation. Encoding 87 was associated with the expression of genes involved in collagen and extracellular matrix processes (Table 1), which has been previously observed to be an important marker of the mesenchymal subtype.[37,40] Encoding 56 was associated with immune system responses (Table 1), and the immunoreactive subtype displayed the highest activation. Encoding 79 is mostly expressed in the proliferative subtype and has low activation in differentiated tumors. The high weight negative

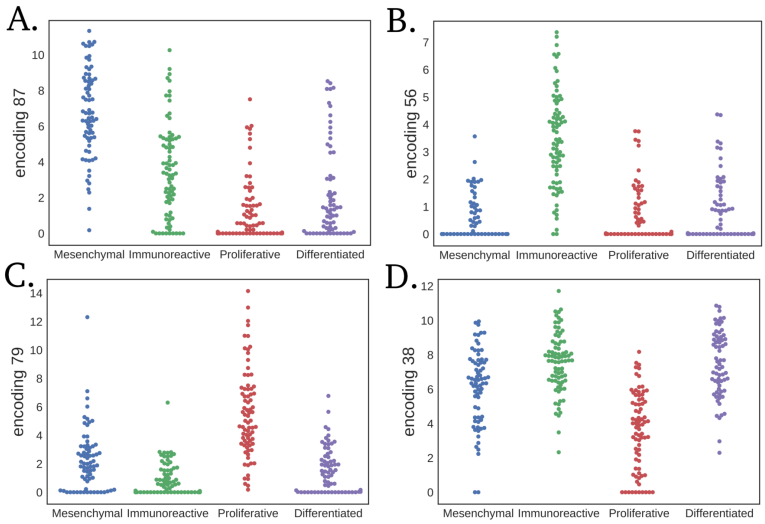


Fig. 4. *Largest mean differences in HGSC subtype vector subtraction for each subtype.* Subtracting the mesenchymal subtype by the immunoreactive results in distribution differences in **(A)** feature encoding 87 and **(B)** encoding 56. Subtracting the proliferative subtype by the differentiated subtype results in differences between **(C)** feature encoding 79 and **(D)** encoding 38.

genes of encoding 79 were associated with glucuronidation processes (Table 1). The negative genes of encoding 38, which also distinguished differentiated from proliferative tumors but in the opposite direction, were also associated with glucuronidation. Previously, glucuronidation processes were observed to be associated with response to chemotherapy and survival in colon cancer patients.[43,44] Our results indicate that differential activation of glucuronidation is a strong signal distinguishing HGSC subtypes. This observation may also help to explain increased survival in HGSC patients with differentiated tumors.[41] Lastly, encoding 77 also separated immunoreactive from mesenchymal tumors and did not display any significant terms, which may indicate novel biology explaining undiscovered subtype differences.

4. Conclusion

Tybalt is a promising model but still requires careful validation and more comprehensive evaluation. We observed that the encoded features recapitulated tissue specific patterns. We determined that the learned features were generally non-redundant and could disentangle large sources of variation in the data, including patient sex and SKCM. It is also likely that the features learn tissue specific patterns distinguishing other cancer-types (our shiny app enables full exploration of VAE features by cancer-type). While we identified specific features separating HGSC subtypes, there are likely several other features that describe other important biological differences across cancer-types including differentiation state and activation states of specific pathways. Interpretation of the decoding layer weights helped to identify the contribution of different genes and pathways promoting disparate biological patterns. However, interpretation by pathway analysis must be performed with caution as these analyses rely on incomplete pathway databases and may contain many false positive results.

Table 1. Summary of significantly overrepresented pathways separating HGSC subtypes

Encoding	Tail	Subtype Enrichment	Pathway	Adj. p value
87	+	Mesenchymal	Collagen Catabolic Process	$1.8e^{-09}$
87	+	Mesenchymal	Extracellular Matrix Organization	$4.2e^{-06}$
87	-	Immunoreactive	Urate Metabolic Process	$1.5e^{-02}$
56	+	Immunoreactive	Immune Response	$1.3e^{-12}$
56	+	Immunoreactive	Defense Response	$2.9e^{-12}$
56	+	Immunoreactive	Regulation of Immune System Process	$8.0e^{-07}$
56	-	Mesenchymal	*No significant pathways identified*	
79	+	Proliferative	Chemical Synaptic Transmission	$9.1e^{-03}$
79	-	Differentiated	Xenobiotic Glucuronidation	$2.1e^{-09}$
38	+	Differentiated	*No significant pathways identified*	
38	-	Proliferative	Xenobiotic Glucuronidation	$7.2e^{-06}$

VAEs provide similar benefits as autoencoders, but they also have the ability to learn a manifold with meaningful relationships between samples. This manifold could represent differing pathway activations, transitions between cancer states, or indicate particular tumors vulnerable to specific drugs. We performed initial testing to determine if we could traverse the underlying manifold by subtracting out cancer-type specific mean activations. While we identified several promising functional relationships existing in a spectrum of activation patterns, rigorous experimental testing would be required to draw strong conclusions about the biological implications. The specific subtype associations must be confirmed in independent datasets and the processes must be confirmed experimentally. It must also be assessed if Tybalt features learned from TCGA pan-cancer are generalizable to other, potentially more heterogeneous datasets. Further testing is required to confirm that Tybalt catalogued an interpretable manifold capable of interpolation between cancer states. In the future, we will develop higher capacity models and increased evaluation/interpretation efforts to catalog Tybalt encoded RNA-seq expression patterns present in specific cancer-types. This effort will lead to widespread stratification of expression patterns and enable accurate detection of samples who may benefit from specific targeted therapies.

5. Reproducibility

We provide all scripts to reproduce and to build upon this analysis under an open source license at `https://github.com/greenelab/tybalt`.[45]

Acknowledgments

This work was supported by NIH grant T32 HG000046 (GPW) and GBMF 4552 from the Gordon and Betty Moore Foundation (CSG). We would like to thank Brett K. Beaulieu-Jones for helpful discussions and Jaclyn N. Taroni and David Nicholson for code review. We would also like to thank four anonymous reviewers for their insightful comments. This is a preprint of

References

1. T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, A. Gitter and C. S. Greene, *bioRxiv* (May 2017).
2. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, *Nature* **542**, 115 (February 2017).
3. J. Zhou and O. G. Troyanskaya, *Nature Methods* **12**, 931 (October 2015).
4. D. P. Kingma and M. Welling, *arXiv:1312.6114 [cs, stat]* (December 2013).
5. D. J. Rezende, S. Mohamed and D. Wierstra, *arXiv:1401.4082 [cs, stat]* (January 2014).
6. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *arXiv:1406.2661 [cs, stat]* (June 2014).
7. J. Tan, J. H. Hammond, D. A. Hogan and C. S. Greene, *mSystems* **1**, e00025 (February 2016).
8. L. Chen, C. Cai, V. Chen and X. Lu, *BMC Bioinformatics* **17**, p. S9 (January 2016).
9. J. Tan, G. Doing, K. A. Lewis, C. E. Price, K. M. Chen, K. C. Cady, B. Perchuk, M. T. Laub, D. A. Hogan and C. S. Greene, *Cell Systems* **5**, 63 (July 2017).
10. I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed and A. Lerchner, *arXiv:1606.05579 [cs, q-bio, stat]* (June 2016).
11. E. Park, *http://www.cs.unc.edu/˜eunbyung/papers/manifold_variational.pdf* .
12. A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, A. Zhavoronkov, A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov and A. Zhavoronkov, *Oncotarget* **8**, 10883 (December 2016).
13. A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper and A. Zhavoronkov, *Molecular Pharmaceutics* (July 2017).
14. L. Rampasek, D. Hidru, P. Smirnov, B. Haibe-Kains and A. Goldenberg, *arXiv:1706.08203 [stat]* (June 2017).
15. S. Huang, I. Ernberg and S. Kauffman, *Seminars in cell & developmental biology* **20**, 869 (September 2009).
16. J. N. Weinstein, E. A. Collisson, G. B. Mills, K. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander and J. M. Stuart, *Nature genetics* **45**, 1113 (October 2013).
17. K. Chaudhary, O. B. Poirion, L. Lu and L. Garmire, *bioRxiv* , p. 114892 (March 2017).
18. P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (ACM, New York, NY, USA, 2008).
19. A. Lamb, V. Dumoulin and A. Courville, *arXiv:1602.03220 [cs, stat]* (February 2016), arXiv: 1602.03220.
20. S. Ioffe and C. Szegedy, *arXiv:1502.03167 [cs]* (February 2015).
21. D. P. Kingma and J. Ba, *arXiv:1412.6980 [cs]* (December 2014).
22. V. Nair and G. E. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10 (Omnipress, USA, 2010).
23. F. Chollet and others, *Keras* (GitHub, 2015).
24. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Joze-

fowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *arXiv:1603.04467 [cs]* (March 2016).
25. C. Doersch, *arXiv:1606.05908 [cs, stat]* (June 2016).
26. K. Franz, *Variational Autoencoders Explained*, 2016).
27. H. Saghir, *An intuitive understanding of variational autoencoders without any formula*, 2017).
28. T. Raiko, H. Valpola, M. Harva and J. Karhunen, *J. Mach. Learn. Res.* **8**, 155 (May 2007).
29. C. K. Snderby, T. Raiko, L. Maale, S. K. Snderby and O. Winther, *arXiv:1602.02282 [cs, stat]* (February 2016).
30. G. Way, *Data Used For Training Glioblastoma Nf1 Classifier* (Zenodo, June 2016).
31. C. Wilks, P. Gaddipati, A. Nellore and B. Langmead, *bioRxiv* , p. 097881 (January 2017).
32. J. Tan, M. Ung, C. Cheng and C. S. Greene, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* , 132 (2015).
33. J. Wang, S. Vasaikar, Z. Shi, M. Greer and B. Zhang, *Nucleic Acids Research* **45**, W130 (July 2017).
34. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nature Genetics* **25**, 25 (May 2000).
35. A. Dosovitskiy, J. T. Springenberg and T. Brox, Learning to generate chairs with convolutional neural networks (IEEE, June 2015).
36. A. Radford, L. Metz and S. Chintala, *arXiv:1511.06434 [cs]* (November 2015).
37. T. C. G. A. R. Network, *Nature* **474**, 609 (June 2011).
38. R. G. Verhaak, P. Tamayo, J.-Y. Yang, D. Hubbard, H. Zhang, C. J. Creighton, S. Fereday, M. Lawrence, S. L. Carter, C. H. Mermel, A. D. Kostic, D. Etemadmoghadam, G. Saksena, K. Cibulskis, S. Duraisamy, K. Levanon, C. Sougnez, A. Tsherniak, S. Gomez, R. Onofrio, S. Gabriel, L. Chin, N. Zhang, P. T. Spellman, Y. Zhang, R. Akbani, K. A. Hoadley, A. Kahn, M. Kbel, D. Huntsman, R. A. Soslow, A. Defazio, M. J. Birrer, J. W. Gray, J. N. Weinstein, D. D. Bowtell, R. Drapkin, J. P. Mesirov, G. Getz, D. A. Levine, M. Meyerson and The Cancer Genome Atlas Research Network, *Journal of Clinical Investigation* (December 2012).
39. L. v. d. Maaten and G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).
40. R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y.-E. Chiew, I. Haviv, Australian Ovarian Cancer Study Group, D. Gertig, A. DeFazio and D. D. L. Bowtell, *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **14**, 5198 (August 2008).
41. G. E. Konecny, C. Wang, H. Hamidi, B. Winterhoff, K. R. Kalli, J. Dering, C. Ginther, H.-W. Chen, S. Dowdy, W. Cliby, B. Gostout, K. C. Podratz, G. Keeney, H.-J. Wang, L. C. Hartmann, D. J. Slamon and E. L. Goode, *Journal of the National Cancer Institute* **106** (October 2014).
42. G. P. Way, J. Rudd, C. Wang, H. Hamidi, B. L. Fridley, G. E. Konecny, E. L. Goode, C. S. Greene and J. A. Doherty, *G3: Genes, Genomes, Genetics* , p. g3.116.033514 (January 2016).
43. J. Cummings, B. T. Ethell, L. Jardine, G. Boyd, J. S. Macpherson, B. Burchell, J. F. Smyth and D. I. Jodrell, *Cancer Research* **63**, 8443 (December 2003).
44. E. Cecchin, F. Innocenti, M. D'Andrea, G. Corona, E. De Mattia, P. Biason, A. Buonadonna and G. Toffoli, *Journal of Clinical Oncology* **27**, 2457 (May 2009).
45. G. Way and C. Greene, *greenelab/tybalt: Initial Development Release*, tech. rep., Zenodo (July 2017).

Diffusion mapping of drug targets on disease signaling network elements reveals drug combination strategies

Jielin Xu†, Kelly Regan-Fendt†, Siyuan Deng
Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, U.S.A.
Email: Jielin.Xu@osumc.edu, Kelly.Regan@osumc.edu, Siyuan.Deng@osumc.edu

William E. Carson III
Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, U.S.A.
Email: William.Carson@osumc.edu

Philip R.O. Payne
Institute for Informatics, Washington University in St. Louis, St. Louis, MO 63110, U.S.A.
Email: prpayne@wustl.edu

Fuhai Li
Department of Biomedical Informatics, Translational Data Analytics Institute, The Ohio State University, Columbus, OH 43210, U.S.A. Email: Fuhai.Li@osumc.edu

The emergence of drug resistance to traditional chemotherapy and newer targeted therapies in cancer patients is a major clinical challenge. Reactivation of the same or compensatory signaling pathways is a common class of drug resistance mechanisms. Employing drug combinations that inhibit multiple modules of reactivated signaling pathways is a promising strategy to overcome and prevent the onset of drug resistance. However, with thousands of available FDA-approved and investigational compounds, it is infeasible to experimentally screen millions of possible drug combinations with limited resources. Therefore, computational approaches are needed to constrain the search space and prioritize synergistic drug combinations for preclinical studies. In this study, we propose a novel approach for predicting drug combinations through investigating potential effects of drug targets on disease signaling network. We first construct a disease signaling network by integrating gene expression data with disease-associated driver genes. Individual drugs that can partially perturb the disease signaling network are then selected based on a drug-disease network "impact matrix", which is calculated using network diffusion distance from drug targets to signaling network elements. The selected drugs are subsequently clustered into communities (sub-groups), which are proposed to share similar mechanisms of action. Finally, drug combinations are ranked according to maximal impact on signaling sub-networks from distinct mechanism-based communities. Our method is advantageous compared to other approaches in that it does not require large amounts drug dose response data, drug-induced "omics" profiles or clinical efficacy data, which are not often readily available. We validate our approach using a BRAF-mutant melanoma signaling network and combinatorial in vitro drug screening data, and report drug combinations with diverse mechanisms of action and opportunities for drug repositioning.

Keywords: Drug repositioning; drug combination; signaling network; network diffusion

† Co-first authors.

This work is supported in part by the NLM Pre-doctoral training grant to KRF, and supported in part by startup funding from OSU Translational Data Analytics to FL.

1. Introduction

Despite the discovery of many disease-causing molecular aberrations, the vast majority are not successfully targeted by approved drugs. Furthermore, widespread drug resistance to targeted therapies is still a major challenge in cancer treatment [1]. Thus, the design of multi-target agents and rationale drug combinations seeks to address some of these issues and accomplish specific objectives: increased overall efficacy, improved initiation for first-line therapies, reduced drug resistance, reduced required doses and reduced drug toxicities. However, the high costs and low success rates of high-throughput drug screening are exponentially prohibitive to screen drug combinations across different cellular contexts and doses [2]. Therefore, computational methods have the potential to focus research efforts on optimal drug combination in the preclinical testing setting, and eventually to aid in clinical decision-making [3, 4].

Malignant melanoma represents an important use case for precision medicine research and systematizing the design of rationale combination therapies, including recent developments in targeted and immune-based therapies. Melanoma tumors are primarily driven by two oncogenes, *BRAF* (~50%) and *NRAS* (~25%), that converge on the MAPK signaling pathway to promote growth, survival and evade apoptosis. Currently approved targeted therapies for patients with *BRAF*-mutant melanoma include first-line treatment with BRAF inhibitors vemurafenib and dabrafenib, which have improved survival by 6-12 months [5]. However, most patients eventually become resistant to BRAF inhibitor therapies, and heterogenous resistance mechanisms have been observed [6]. For instance, while mutations in *BRAF* and *NRAS* genes are observed to be mutually exclusive across primary and metastatic tumors, acquired *NRAS* mutations have been described as a mechanism of BRAF-inhibitor resistance [7]. Reactivation of the MAPK signaling pathway can also occur via MEK over-activation, and the first combination of targeted therapy including BRAF and MEK inhibition was recently approved for patients with *BRAF*-mutant melanoma. While this combination extends patient survival an additional 5-10 months, additional non-MAPK pathway resistance mechanisms arise, and new, more durable drug combination regimens are needed [8].

In our previous work, we described a novel computational method, SynGeNet [9], which i) integrates transcriptomics and protein-protein interaction data into a comprehensive disease signaling network to map signal flow from an initial set of disease-related “root” genes; and ii) determines drug combinations that maximally reverse disease-associated gene expression signals and targets topologically important nodes in the overall network. We showed that SynGeNet outperformed two other transcriptomics-based drug combination methods in predicting drug combinations validated *in vitro*, and that it could recapitulate genotype-specific results across diverse melanoma cell lines. Importantly, we observed that the network-mining step, which utilized an average of centrality metrics for drug target pathways, was the most crucial aspect of our method in validating drug combination efficacy. Due to the importance of modeling drug-target network connections, we sought to evaluate additional methods that exploit drug-target and disease signaling network structure. Additionally, we sought to overcome the limitation of requiring drug-

induced gene expression profiles as part of the original SynGeNet method. Compared with other unsupervised approaches based on correlating gene expression profiles alone, network-based approaches can more explicitly indicate possible mechanism of action in terms of inhibited signaling targets, and consequently specify a measure for predicting efficacy.

In this study, we propose a novel approach to prioritize drug combinations that can potentially impact a *BRAF*-mutant melanoma signaling network that is constructed from the integration of gene expression and protein-protein interaction data. We employed a random walk with restart (RWR) model to traverse a *BRAF*-mutant melanoma disease signaling network to derive a drug-disease "impact matrix". We then selected drugs that can maximally perturb the disease signaling network for subsequent drug combination modeling. Additionally, we hypothesized that drugs with different mechanisms of action or targets on distinct network modules may have a higher potential for synergy according to the independent mechanism theory for drug combinations. Therefore, we divided prioritized drugs into communities (sub-groups) based on drug target similarity matrices and subsequently ranked drug combinations representing different drug communities. To our knowledge, this is the first study to apply this paradigm to evaluate drug combination hypotheses. Furthermore, we apply the RWR method to determine drug mechanisms by delineating shortest local paths within the network.

2. Methods

An overview figure of the RWR approach to predict and validate drug combinations is shown in **Supplemental Figure S1**. All supplemental materials can be found at the following URL: https://www.kaggle.com/osubmi/diffusion-mapping-for-drug-combinations. All code is made available upon request.

2.1. ***Melanoma disease signaling network construction.*** In our previous work, we defined a melanoma disease signaling network integrating gene expression data from a publicly available dataset of melanoma patient tumors harboring driver $BRAF^{V600E/K}$ mutations (GSE15605) with protein-protein interaction data from the BioGRID database [10]. Briefly, the network was constructed using the belief propagation approach to map signal flow from a set of frequently mutated "root" melanoma disease genes (n=30) from the DisGeNET database [11, 12]. We hypothesized that an estimated driver disease network could be constructed by minimizing the cost function that weighs the trade-off of including highly activated genes via gene expression fold-changes against including experimentally validated protein-protein interactions with decreasing confidence. This resulted in a disease signaling network of 131 genes to be integrated with drug target information. The mathematical function is reproduced here: Given the BioGRID background network, G = (V, E), where V and E represent the vertices and edges in the BioGRID network, the sub-network, G' = (V', E'), is constructed to minimize the objective function:

$$\min_{E' \subseteq E, V' \subseteq V} \sum_{e \in E'} c_e - \lambda \sum_{i \in V'} b_i \qquad (1)$$

where *Ce* (cost of edge) is set to 0.2, b_i is the patient tumor gene expression fold change representing "activated" state of signaling components (i.e. positive fold change), and λ (set as 0.02) regulates the size of the sub-network. Detailed information regarding disease signaling network construction and empirical rationale for parameter selection can be found in our previous work [13].

2.2. ***Drug target-disease signaling impact matrix construction using Random Walk with Restart Model (RWR)***. To estimate the potential impact of inhibiting a drug target on the disease signaling network, the random walk with restart model (RWR) was employed. RWR describes a stochastic process of network signaling flow as follows: at each iteration step, the network signal beginning at an individual gene travels randomly, with equal probability, to a neighboring gene or remains in its current location. In this application, the RWR model is initiated for a gene representing a known direct drug target. The updated location of the network signal can be viewed as a probability expectation, which is defined mathematically in Eq. (2).

$$\vec{r_i} = (1-c)W\vec{r_i} + c\vec{e_i} \tag{2}$$

Here, $\vec{r_i}$ is a probability vector with elements r_{ij} that denotes the probability of signal flow at gene *i* travels to gene *j*, and the sum of all r_{ij} with respect to j should equal to 1. In our method, we set r_{ij} as the impact of gene i to gene j. Here *c* denotes the restart probability, and W is the normalized adjacency matrix, which is constructed on edge connections (protein-protein interactions) with respect to the BioGRID network. $\vec{e_i}$ represents the starting vector, and has all elements equal to 0 except the *i*-th element, which is set equal to 1. By solving (2) iteratively, we can extract the target-disease impact matrix $M \in \mathbb{R}^{|V_D|\times|V_T|}$, with M_{ij} denoting the impact of *j-th* drug target on the *i-th* gene on the disease signaling network.

2.3. ***Single drug scoring model.*** The single drug scoring model is constructed in the following sequential steps. First, with target-disease impact matrix M_{ij} with $M \in \mathbb{R}^{|V_D|\times|V_T|}$, the drug-disease matrix is defined by summing all related target-disease impacts, i.e., we define drug-disease matrix $\widetilde{M} \in \mathbb{R}^{|V_D|\times|DRUG|}$, as:

$$\widetilde{M}_{ij} = \sum_{s\in\{V_T[DRUG_j]\}} M_{is} \tag{3}$$

In other words, Eq. (3) gives us the overall impact of a set of drugs across disease genes within the network. Second, considering the relative "influence" (i.e. number of connections) of disease genes with respect to the topological structure of disease network, we weight drug-disease impact by the degree of disease genes. Thus, the larger the disease gene degree, the higher magnitude the drug impact is amplified. Mathematically speaking, we define a weighted drug-disease impact matrix $\widehat{M} \in \mathbb{R}^{|V_D|\times|DRUG|}$ as follows:

$$\widehat{M}_{ij} = \widetilde{M}_{ij} \cdot \overrightarrow{DEG}[i] \tag{4}$$

with $\overrightarrow{DEG}[i] = deg\left(V_D^i, N_D\right)$, where $1 \le i \le |V_D|$. Finally, a score for an individual drug is defined as its average impact on all genes in disease network.

$$DS[i] = \frac{\sum_{s=1}^{|V_D|} \widehat{M}_{si}}{|V_D|} \quad (5)$$

with $\overrightarrow{DS} \in \mathbb{R}^{|DRUG| \times 1}$. The computed drug scores are ranked in decreasing order.

2.4. *Drug combination scoring model*. The first assumption embedded in the drug combination model is independent mechanism theory, which states that drugs with different mechanisms of action or targets on different disease signaling modules have a higher potential for synergy. Based on this assumption, we first divided drugs into communities based on target similarity, and then estimate the drug combination synergy based on their impact disease signaling elements. The second assumption for the drug combination score model is that isolated disease genes do not contribute to the disease signaling, and that a disease gene can only influence the disease network if it is connected to other nodes inside the disease network along "important" paths. Relative importance for network paths is described in subsection 2.4.2.

2.4.1. Drug community clustering. We clustered drugs that passed criteria described in the single drug score model into different functional groups via the affinity propagation (AP) clustering algorithm [14]. Drug-target interaction information was extracted from the DrugBank database, and the resulting Jaccard index coefficient was used as the basis for affinity propagation clustering. Drug-drug similarity was evaluated by Jaccard index:

$$S_{ij} = \frac{\left|V_T[SEL\_DRUG_i] \cap V_T\left[SEL\_DRUG_j\right]\right|}{\left|V_T[SEL\_DRUG_i] \cup V_T\left[SEL\_DRUG_j\right]\right|} \quad (6)$$

with S_{ij} denotes the similarity between selected drug i and selected drug j, and $S = \{S_{ij}\} \in \mathbb{R}^{|SEL\_DRUG| \times |SEL\_DRUG|}$.

***2.4.2. Drug combination prediction*.** We select drug pairs from different communities as candidate combinations. We then determine the disease genes that are highly impacted by different drug combinations. The disease genes are scored by truncating impact above a certain threshold (*T*) as described below:

$$T = alpha * \max_{i,j} M \quad (7)$$

with $1 \leq i \leq |V_D|$ and $1 \leq j \leq |V_T|$. An ***alpha*** value is determined as follows: We define a parameter *p* that represents the percentage of disease genes exclusively impacted by drugs involved in a given drug combination. The optimal alpha should provide the highest '*p*' as defined below. In other words, the optimal alpha allows both drugs within a combination to provide as much unique information as possible. In this way, we prioritize drug pairs that influence non-redundant disease sub-networks.

$$p = \frac{|D_1| + |D_2|}{|D_1| + |D_2| + |D_c|} \quad (8)$$

With D_1, D_2, D_c represent the number of exclusive disease genes impacted by drug 1, drug 2, and both drugs, respectively. Based on the optimal alpha selected, for a given drug combination $\left(DRUG_i, DRUG_j\right)$, we can extract highly impacted disease genes for both

drugs, i.e., IDG_i and IDG_j, with $IDG_i = \bigcup_{k=1}^{|IDG_i|} V_D^k$ and $IDG_j = \bigcup_{l=1}^{|IDG_j|} V_D^l$. The drug combination score $CS(DRUG_i, DRUG_j)$ is then defined as follows:

$$CS(DRUG_i, DRUG_j) = \begin{cases} 0, & |IDG_i| = 1, |IDG_j| = 1 \\ PS_{1D}(DRUG_i), & |IDG_i| > 1, |IDG_j| = 1 \\ PS_{1D}(DRUG_j), & |IDG_i| = 1, |IDG_j| > 1 \\ PS_{2D}(DRUG_i, DRUG_j), & |IDG_i| > 1, |IDG_j| > 1 \end{cases} \quad (9)$$

with $PS_{1D}(DRUG_k)$ and $PS_{2D}(DRUG_k, DRUG_l)$ defined below respectively. Here, $|IDG_k| = 1$ refers to $DRUG_k$ impacts only one isolated gene in the entire disease network, and therefore, it does not contribute to the drug combination score. Using the RWR model, we then evaluate $DRUG_k$ impact to all Dijkstra's shortest paths that connects two arbitrary disease genes V_D^m and V_D^n in set IDG_k for $DRUG_k$, and then rank them in decreasing order as I_{DRUG_k}, then we have:

$$PS_{1D}(DRUG_k) = \sum_{m=1}^{|PATH_{NT}|[k]} I_{DRUG_k}[m] \quad (10)$$

with

$$|PATH_{NT}|[k] = \begin{cases} |I_{DRUG_k}|, & |I_{DRUG_k}| < 3 \\ 3, & |I_{DRUG_k}| \geq 3 \end{cases} \quad (11)$$

Here, $|I_{DRUG_k}|$ denotes the number of all shortest paths that connects all non-isolated disease genes that are highly impacted by the kth drug, and $|PATH_{NT}|[k]$ denotes the truncated numbers of all shortest paths, which are used for drug combination score evaluation. For construction of $PS_{2D}(DRUG_k, DRUG_l)$, considering a certain path score threshold T_{PS} which is chosen as the magnitude separator, i.e., local path scores above T_{PS} stay within the highest magnitude, and local path scores below T_{PS} are with lower magnitudes, then we have:

$$PS_{2D}(DRUG_k, DRUG_l) = \begin{cases} \sum_{k,l} PS_{1D}(DRUG_m), |PATH_T|[k] + |PATH_T|[l] = 0 \\ PS_{1D}(DRUG_k), |PATH_T|[k] > 0, |PATH_T|[l] = 0 \\ PS_{1D}(DRUG_l), |PATH_T|[k] = 0, |PATH_T|[l] > 0 \\ \sum_{k,l} PS_{1D}(DRUG_m), |PATH_T|[k] * |PATH_T|[l] > 0 \end{cases} \quad (12)$$

and

$$|PATH_T|[k] = \begin{cases} |I_{DRUG_k} > T_{PS}|, & |I_{DRUG_k} > T_{PS}| < 3 \\ 3, & |I_{DRUG_k} > T_{PS}| \geq 3 \end{cases} \quad (13)$$

Here, $\left|I_{DRUG_k} > T_{PS}\right|$ denotes the number of all shortest paths above T_{PS} for the kth drug, and $|PATH_T|[k]$ denotes the truncated numbers of shortest paths with highest magnitude which are used for drug combination score evaluation.

3. Results

3.1. ***Evaluation of single drug predictions.*** To evaluate the proposed approach, we first obtained all the FDA approved drugs with target information available from the DrugBank [15], which resulted in 1,433 drugs selected for this study. Based on the single drug score model, we selected the top 20% ranked single drugs (287 drugs) predicted by our method targeting a *BRAF*-mutant melanoma disease signaling network constructed via integration of gene expression and protein-protein interaction data (described in **Methods**). We first sought to distinguish single drug predictions on a mechanistic basis by applying affinity

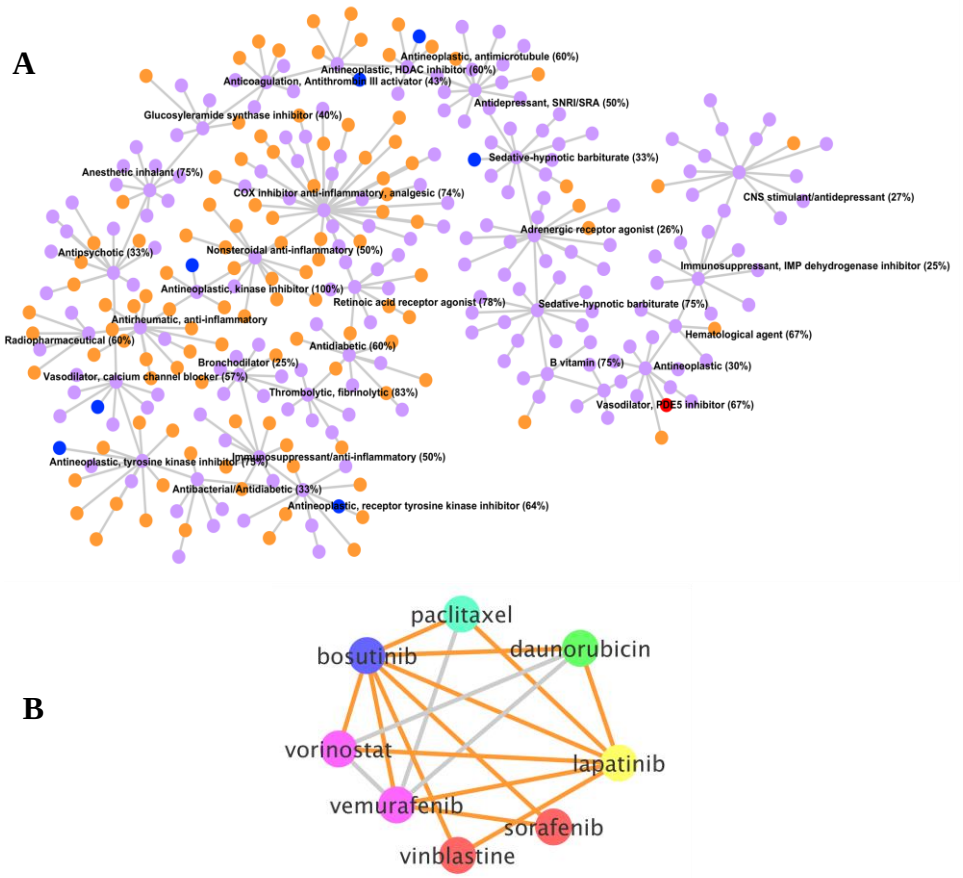


Figure 1A. Drug communities of prioritized single drugs (n = 287) organized by affinity propagation (AP) clustering. Blue nodes represent genes both validated in the screening study and with literature evidence. Orange nodes represent drugs with reported anti-melanoma effects in the literature. One red node represents a drug validated in the high-throughput combinatorial screening study without literature evidence. All other drugs are shown in purple. The dominant drug class per community is shown with the percentage of drugs belonging to a dominant drug function class. **Figure 1B.** Validated drugs combinations (n=17 pairs) from 8 drugs recovered in the top 30% of predictions by our method. Orange edges denote the 13 validated drug combinations prioritized by our method. Light grey edges denote validated combinations missed by our method. Drugs are color-coded according to drug community.

propagation clustering on the Jaccard Index of the drug-drug similarity matrix to define drug community clusters. This resulted in 30 non-overlapping drug communities, where each drug was assigned to a single community (**Figure 1A**). The average size of a drug community was 9.6 drugs, with an average of 54.8% drugs belonging to a predominant drug function class within a community (**Figure 1A**). While anti-neoplastics was the most frequent drug function category observed (6/30 communities), other common drug functions included anti-inflammatory, antidiabetic, vasodilation and anti-depressant drugs, indicating a high potential for drug repositioning. To evaluate these selected drugs, we first manually searched the literature for associations between these drugs and melanoma. We found literature associations for 112/287 drugs with melanoma having 10 or more citations in PubMed (**Supplemental Table S1**). Additionally, in order to evaluate our predictions in the *BRAF*-mutant melanoma context, we used the reported GI50 (50% growth inhibition) values from a previous drug screen testing 40 agents in combinations *BRAF*-mutant melanoma cell lines [16]. Drugs were first screened individually across melanoma cell lines (**validation dataset 1**). Of these 40 tested drugs, 13 overlapped with our initial set of FDA-approved therapies, and of these 13 drugs, we note that 8 were ranked in the top 20% of

3.2. ***Evaluation of drug combination predictions.*** We hypothesized that drugs paired from different drug communities would have non-redundant functions to inhibit the overall disease signaling network, and thus could represent efficacious drug combinations. From the aforementioned combinatorial drug screening study, a total of 650 drug combinations (with different doses) were considered validated by the previous criteria defined by the authors (**validation dataset 2**): both ≥ 50% growth inhibition in *BRAF*-mutant melanoma cell lines and ≥ 15% growth inhibition observed specifically in *BRAF*-mutant cell lines vs. other genetic backgrounds. Among the total 650 validated drug combinations, there were 28 drug combinations overlapping between the original subset of FDA-approved drugs. Considering the top 20% of predicted drugs, our method recovered 17 out of 28 validated drug combinations (**Figure 1B**). Interestingly and consistent with our hypothesis, 16 out of 17 of the validated drug combinations included both drugs from distinct communities in our clustering analysis. Only one validated drug combination contained two drugs from the same community and was discarded by our method for this reason. Furthermore, the drug combination score model predicted 13 out of 17 drug combinations that ranked in the top 30% of all qualified drug combinations (**orange** edges in **Figure 1B**), and the first 10 drug combinations ranked by GI50 score were all prioritized by our approach.

3.3. ***Drug combination mechanism discovery.*** We further sought to extend the RWR network diffusion model to test the independent drug mechanism hypothesis for drug combinations from a signaling pathway perspective by determining local signaling paths within the network. First, we aimed to define an optimum threshold that provides the most exclusive information between two drugs, and tested several alpha values that control the percentage of disease genes exclusively impacted by drugs paired in combination (alpha = 0.0001, 0.0003, 0.0005, 0.0007, 0.001) for all validated drug combinations as plotted in **Supplemental Figure S2**. According to our observations, we empirically selected 0.0005 as the optimal alpha.

To demonstrate the independent mechanism hypothesis in a clinically relevant context, we first considered the two FDA approved first-line targeted therapies for BRAF-mutant melanoma, including a BRAF and MEK inhibitor, respectively: vemurafenib + cobimetinib, and dabrafenib + trametinib. It is worth noting that all four single drugs are selected among the top 20% by our method. For the sake of simplicity, only the local path plot for vemurafenib + cobimetinib is presented (**Figure 2A**). As is shown in **Figure 2A,** the two shortest local paths for this drug combination originate from the *BRAF* and *MAP2K1* genes (targets of vemurafneib and cobimetinib, respectively), and they do not intersect, thus fulfilling the independent mechanism hypothesis. Among the 17 selected validated drug combinations (drug edges in Figure 1B) involving the 8 prioritized single drugs (drug nodes in Figure 1B), we note that 5 of these drug combinations (**Table 1**, right) shared common genes with impact scores within the same order of magnitude, thus permitting testing of the independent mechanism hypothesis via RWR of shortest paths.

Of these five combinations, the top four drug combinations all exhibited independent local paths derived from each drug. For the sake of simplicity, we only show the local path plot for the bosutinib + sorafenib combination in **Figure 2B** to compare to the clinically relevant example of vemurafenib + cobimetinib. As seen with vemurafenib + cobimetinib, the bosutinib + sorafenib combination also demonstrates local paths connecting to *BRAF* and *MAP2K1* (MEK1) genes through independent mechanisms. Another interesting observation for the bosutinib + sorafenib combination is that of the 20 total disease genes that are highly impacted by both drugs (union of shared and unique genes), only 6 genes (*BRAF*, *APP*, *FLT1*, *CDK2*, *SRC* and *MAP2K1*) are connected through local paths that reveal mechanisms by which both drugs impact the *BRAF*-mutant melanoma disease signaling network. Finally, we note that although both drugs share four highly impacted disease genes *BRAF*, *BSG*, *IPO13*, and *MC1R*, the only connected gene in a local path of these is *BRAF*, suggesting the major role for *BRAF* signaling in the network. Furthermore, we note that this local path for *BRAF* is connected to genes highly impacted by sorafenib, which unlike bosutinib, directly targets the BRAF protein.

Table 1. (Left): Highest truncated local path scores of prioritized single drugs (see drug nodes in Figure 2B) and BRAF/MEK inhibitors. We note that several drugs, e.g., daunorubicin and vemurafenib, have only two highly impacted disease genes based on our algorithm, and therefore returns only one shortest path. (Right): The 5 drug combinations (see drug edges in Figure 1B) assessed for independent mechanism hypothesis based on single drugs with local paths scores presented on the left.

bosutinib	**vorinostat**	**vinblastine**	**daunorubicin**	**paclitaxel**	**lapatinib**	**Drug A**	**Drug B**
0.5336	0.0038	0.0015	0.0032	0.0020	0.4006	bosutinib	sorafenib
0.4002	0.0031	0.0013		0.0020	0.2006	bosutinib	lapatinib
0.4000	0.0029	0.0012		0.0020	0.2005	bosutinib	vemurafenib
sorafenib	**vemurafenib**	**dabrafenib**	**trametinib**	**cobimetinib**		vemurafenib	lapatinib
0.3201	0.2010	0.2038	0.2676	0.2676		vemurafenib	sorafenib
0.3200		0.2013	0.2016				
0.3200		0.2003	0.0023				

3.4. ***Method comparison and robustness evaluation.*** First, we provide a comparison between our RWR model and our previously published SynGeNet model in **Table 2**. We observed that the RWR model recovered 76% of validated drug combinations in the top 30% of predictions, whereas SynGeNet recovered 69% of possible validated drug combinations at in the top 30% of predictions. Notably, RWR found 9 unique drug combinations not predicted by SynGeNet, while SynGeNet found 2 unique drug combinations not predicted by RWR. Additionally, to evaluate the robustness of single drug and drug combination prediction models, and consequently the confirm the validity of the highlighted signal flow path as presented in Figures 2A and 2B, we evaluated single and combination drug prediction results by re-wiring the underlying network randomly. The randomly re-wired networks were constructed by keeping the same nodes (genes) in the disease network and randomly assigning connections (edges), while maintaining the same number of connected genes in the original network. We generated 100 random disease networks, and evaluated single drug and drug combination predictions respectively (**Supplemental Figure S3 and Table 2**). We observed that the disease network implemented in our model consistently gave the best performance compared with the 100 randomly generated disease networks. For single drug predictions, our model could predict 8 out of 13 validated drugs, while the highest number of validated drugs predicted by a random network was 6 drugs. Furthermore, the highest number of validated drug combinations predicted by the random networks was 11, compared to the 17 out of 28 predictable, validated drug combinations.

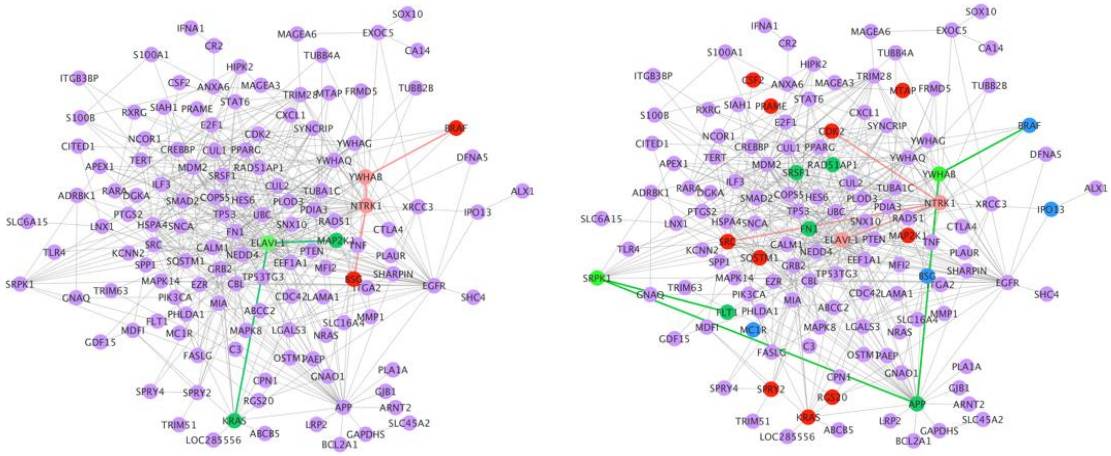


Figure 2A (Left). Local pathways highly impacted by vemurafenib (red) and cobimetinib (green). **Figure 2B (Right).** Local pathways highly impacted by bosutinib (red + blue) and sorafenib (green + blue). Dark red and green nodes represent target genes that are highly impacted exclusively by first and second drugs, respectively; light red and green nodes represent intermediate genes necessary for connections between target genes. Dark blue nodes represent common genes that are highly impacted by both drugs, and light blue nodes represent common intermediate genes necessary for connection between target genes. Purple nodes represent other network nodes not highly impacted by drugs.

Table 2. Comparison between the SynGeNet model and the RWR model proposed in this paper.

	SynGeNet Model	RWR Model
Genomic Data Input	GSE15605 BRAF melanoma gene expression	GSE15605 BRAF melanoma gene expression
Candidate Drug Input	633 FDA approved drugs with gene expression data and known targets	1,433 FDA approved drugs with known targets
Validation Set input	Melanoma cell line drug combination screening (Held *et al*, 2013)	Melanoma cell line drug combination screening (Held *et al*, 2013)
Single Drug Model	GSEA-based connectivity score for signaling network gene signature reversal	Utilize RWR model to rank single drugs according to their average impact to the entire disease network
Drug Combination Model	Drug targets that independently target highly central nodes using averaged topology score	Effective drug combinations are filtered out according to their signal flow path strength evaluation inside the entire disease network
Drug Combination Prediction Evaluation	Among top 30% drug combination predictions, 9 out 13 (69%) validated drug combinations	Among top 30% drug combination predictions, 13 out 17 (76%) validated drug combinations

4. Discussion

In summary, we present a novel approach to predict drug combinations using a RWR model to traverse a heterogeneous network integrating drug target and disease signaling elements. Drug combinations are prioritized that contain individual drugs that can partially impact the disease signaling network through distinct modules and represent different drug communities in order to fulfill the independent mechanism theory. The main advantage of this approach is that it permits a coherent framework to integrate different levels of drug, target and disease information and exploits the entire network structure, including multiple points of entry and pathways to traverse. Therefore, the issue of missing data can be resolved, where RWR can predict drug targets for drugs with no known drug-target interactions through intermediate paths. Additionally, our approach is advantageous over other methods, as it does not require large amounts of drug-induced genomics profiles or other sources of preclinical or clinical efficacy information, which are often lacking. Ultimately, we observed that the RWR model provides a precise delineation of the mechanism of action of drug combinations using propagated signal information as compared to other network-based approaches that may only highlight isolated, central genes. With increasing publicly available patient genomics datasets and ubiquitous drug-target databases, our approach can be applied to a variety of disease contexts and is highly scalable to complex drug-target-gene interaction networks.

There are several limitations of our method to be improved and other challenges for further investigation. First, the constructed signaling network is an undirected graph. The estimation of the signaling diffusion process of signaling on the network may not be entirely accurate. For instance, future studies should implement directed signaling networks (e.g., KEGG signaling network). Also, while affinity propagation is a well-established unsupervised clustering method that has been successfully applied to constructing communities based on drug-drug similarity metrics [17], other methods may be explored including Markov clustering and other energy-model layout algorithms [18]. Second, we can include more extensive drug datasets, as expanding known drug-target interactions will also likely increase the method performance. It will be important to further validate importance of disease-specific signaling networks using larger drug combination validation datasets. Future work could evaluate the effect of incorporating different types

of biological and drug information to construct similarity matrixes (e.g. sequence, chemical structure), and to evaluate the approach in a pan-cancer setting. Another interesting aspect to explore in future work would be the impact of drug community structure and the similarity of toxicity profiles to derive drug combination models that balance efficacy and toxicity simultaneously [19]. Finally, it will be important to confirm predicted mechanisms of drug synergy in prospective *in vitro* experiments (e.g., CRISPR gene editing) to assess the impact of local paths and gene sub-networks in the overall disease signaling network.

References

1. Holohan, C., et al., *Cancer drug resistance: an evolving paradigm.* Nat Rev Cancer, 2013. **13**(10): p. 714-26.
2. Bunnage, M.E., *Getting pharmaceutical R&D back on target.* Nat Chem Biol, 2011. **7**(6): p. 335-9.
3. Madani Tonekaboni, S.A., et al., *Predictive approaches for drug combination discovery in cancer.* Brief Bioinform, 2016.
4. Huang, L., et al., *DrugComboRanker: drug combination discovery based on target network analysis.* Bioinformatics, 2014. **30**(12): p. i228-36.
5. Chapman, P.B., et al., *Improved survival with vemurafenib in melanoma with BRAF V600E mutation.* N Engl J Med, 2011. **364**(26): p. 2507-16.
6. Rizos, H., et al., *BRAF inhibitor resistance mechanisms in metastatic melanoma: spectrum and clinical impact.* Clin Cancer Res, 2014. **20**(7): p. 1965-77.
7. Lo, R.S. and H. Shi, *Detecting mechanisms of acquired BRAF inhibitor resistance in melanoma.* Methods Mol Biol, 2014. **1102**: p. 163-74.
8. Long, G.V., et al., *Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma.* N Engl J Med, 2014. **371**(20): p. 1877-88.
9. Regan, K., P. Payne, and F. Li, *Integrative network and transcriptomics-based approach predicts genotype-specific drug combinations for melanoma.* 2017 Joint Summits on Translational Bioinformatics, San Francisco, March 27 ~ 30, 2017, 2017.
10. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
11. Pinero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database (Oxford), 2015. **2015**: p. bav028.
12. Bailly-Bechet, M., et al., *Finding undetected protein associations in cell signaling by belief propagation.* Proc Natl Acad Sci U S A, 2011. **108**(2): p. 882-7.
13. Regan, K., et al., *Drug Repurposing Hypothesis Generation Using the "RE:fine Drugs" System.* J Vis Exp, 2016(118).
14. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.
15. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
16. Held, M.A., et al., *Genotype-selective combination therapies for melanoma identified by high-throughput drug screening.* Cancer Discov, 2013. **3**(1): p. 52-67.
17. Sirci F., et al., *Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses*. NPJ Syst Biol Appl, 2017. **3**(23).
18. Udrescu L., et al., *Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing*. Sci Rep, 2016. **6**: p.32745.
19. Huang H., et al., *Systematic prediction of drug combinations based on clinical side-effects*. Sci Rep, 2014. **24**(4):p.7160.

Session Introduction: Challenges of Pattern Recognition in Biomedical Data

Shefali Setia Verma
Geisinger Health System
The Huck Institute of the Life Sciences, The Pennsylvania State University,
328 Innovation Blvd Ste 210
State College, PA 16803

Anurag Verma
Geisinger Health System
The Huck Institute of the Life Sciences, The Pennsylvania State University
328 Innovation Blvd Ste 210
State College, PA 16803

Anna Okula Basile
Department of Biochemistry and Molecular Biology, The Pennsylvania State University
328 Innovation Blvd Ste 210
State College, PA 16803

Marta-Byrska Bishop
Geisinger Health System
328 Innovation Blvd Ste 210
State College, PA 16803

Christian Darabos
Research Computing Services, Dartmouth College,
HB 6129
Hanover, NH 03755

The analysis of large biomedical data often presents with various challenges related to not just the size of the data, but also to data quality issues such as heterogeneity, multidimensionality, noisiness, and incompleteness of the data. The data-intensive nature of computational genomics problems in biomedical informatics warrants the development and use of massive computer infrastructure and advanced software tools and platforms, including but not limited to the use of cloud computing. Our session aims to address these challenges in handling big data for designing a study, performing analysis, and interpreting outcomes of these analyses. These challenges have been prevalent in many studies including those which focus on the identification of novel genetic variant-phenotype associations using data from sources like Electronic Health Records (EHRs) or multi-omic data. One of the biggest challenges to focus on is the imperfect nature of the biomedical data where a lot of noise and sparseness is observed. In our session, we will present research articles that can help in identifying innovative ways to recognize and overcome newly arising challenges associated with pattern recognition in biomedical data.

1. Introduction:

Machine learning methods are designed to identify regularities in datasets and then use the identified patterns in a subset of the data to make predictions for the rest of the data. Supervised and unsupervised machine learning methods for pattern recognition have been widely applied in many fields such as image and speech recognition, medical diagnosis, business analytics, finance, as well as in social media, movie recommendations (Netflix), retails, to name a few[2]. With technological advancements, biomedical data is increasing exponentially in size, and there is a high demand to apply these techniques to understand the etiologies of complex diseases[3]. To achieve this goal, it is important to address the challenges of big data analytics and develop optimized methods for pattern recognition that can handle complexities of biomedical data.

The biomedical field, in the current era of precision medicine, is recognized for the interest of researchers in elucidating the genetic architecture of human traits/diseases to improve clinical care. Some of the publicly available 'Big Data' datasets include but are not limited to The 1000 Genomes Project, The Cancer Genome Atlas (TCGA), UK Biobank, Encyclopedia of DNA Elements (ENCODE), Gene Expression Omnibus (GEO), the Library of Integrated Network-based Cellular Signatures (LINCS), the database of Genotypes and Phenotypes (dbGaP), and many other[4–7]. These resources consist of metadata from association analyses, variant information from commercial genotyping chips, whole exome and genome sequencing data, phenotype information, structural variation, gene expression, and among others. Challenges for identifying patterns arise in one data type and increases more in attempts to integrate multiple aforementioned data types/omics[8]. This expanding knowledge is both a blessing and a curse for identifying patterns. Traditional methods of analyzing biomedical data obtained from various high throughput sources are inadequate to handle the ever-increasing wealth of knowledge that is gathered about genotype and phenotype. In this session, we will address the challenges arising from attempts to integrate biomedical data from various sources (including, but not limited to, one or across more species, use of raw data, or summary level statistics) and identify patterns from these multi-omic datasets[9].

The data-intensive nature of computational genetics problem sets in the biomedical informatics field warrants the development and use of vast computer infrastructure and advanced software tools and platforms. Many existing technologies, e.g., Hadoop, Spark, MongoDB, Neo4j, make storage and analytics of large-scale datasets feasible[10]. Additionally, many such technologies are also available via various cloud-computing platforms such as Amazon Web Services (AWS), Google Cloud Platform, Cloudera, as well as vendors, such as DNAnexus, BaseSpace, SevenBridges, Cypher Genomics[11,12]. However, these options are often costly and out of reach for the majority of modest size research groups. While cloud computing aids in analytical performance by improving computing time and storage, there is considerable room for improvement in current software design in biomedical research for cloud-based big-data analysis.

The manuscripts in this session highlight the importance of network-based methods in identifying patterns and address the diverse range of challenges associated with machine

learning techniques. The applications of these methods are well demonstrated in EHR, next-generation sequencing data, as well as in simulated datasets as described below.

2. Session Contributions:

2.1 Network-based approaches

Network-based methods for pattern and data mining have gained popularity as efficient computational approaches[13,14]. For example, networks can be used for explaining associations among genetic variants and diseases where diseases and variants are represented as nodes, and associations are represented as edges. Applying various network analysis techniques has also helped in identifying hidden patterns in datasets which are otherwise not visible when results are evaluated in a tabular form[15,16]. The utility of networks is critical in integrating results from various association analyses as well as integrating multi-omic data sets in identifying combinatorial effects of variations on phenotypes. Along with representing associations, networks can also be used in identifying a non-constituent effect of different variables on a phenotype.

In the manuscript titled *"Functional network community detection can disaggregate and filter multiple underlying pathways in enrichment analysis"*, **Harrington** et al. address challenges with identifying pathways from differential expression analyses in non-network based methods. They demonstrate how applying a network based approach that combines community detection with functional networks can help in identifying true positive pathways. They applied the proposed method on simulated dataset and showed its utility on a biological dataset to discover pathways enriched across high grade serous ovarian cancer (HGSC).

Agarwal et al. address the challenge of dealing with imperfect and noisy molecular network data to uncover disease pathways and proteins in their manuscript titled *"Large-Scale Analysis of Disease Pathways in the Human Interactome"*. The authors conducted a comprehensive network analysis on publicly available data from human protein-protein interaction (PPI) network and DisGeNET database containing protein-disease associations. They observed that several proteins associated with a disease tend to fall in different pathways that are not necessarily well connected. This analysis could be useful in the future development of network-based methods to identify robust pathways.

Biomedical data is highly heterogeneous and incomplete, making extraction of meaningful biologically information a major challenge[17,18]. In their manuscript titled *"OWL-NETS: Transforming OWL Representations for Improved Network Inference"*, **Callahan** et al. propose a novel method for abstracting complex, heterogeneous biological knowledge into lossless network representations that facilitate network inference. The OWL-NETS method could help in enhancing network inference where multi-omic and complex biological information is utilized.

2.2 Machine learning approaches

Deep learning and machine learning techniques are an integral component of evaluating biomedical data, and their use has been increasing dramatically over the past decade[19,20]. Machine learning methods are used extensively for identification of correlations between variables, e.g. between different phenotypes, or between phenotypes and genotypes. Moreover, many methods have demonstrated their applications in other-omics datasets such as proteomics, transcriptomics, and metabolomics[21–23]. Methods such as unsupervised learning are independent of any set rules for identifying patterns to associate or correlate variables. These methods have also gained popularity in the field of biomedical data to improve prediction of health outcomes by mining biologically relevant data.

Recently, machine learning methods, both supervised and unsupervised, have been widely used in the field of biomedical informatics. Though the concept of machine learning is not new, researchers still struggle to identify the best method suited for identifying viable solutions to their problems. In their manuscript titled *"Data-driven Advice for Applying Machine Learning to Bioinformatics Problems",* **Olson** et al., present a wide range of comparisons among various machine learning methods, and show how effectively tuning the methods could enable identification of true positive results.

Machine learning methods are also extensively used in analyses of medical imaging data, such us in cancer radiomics, an emerging field focused on quantification of tumor phenotypes using various imaging features. In the manuscript titled *"Tree-based Methods for Characterizing Tumor Density Heterogeneity",* **Shoemaker** et al. propose a novel decision tree-based approach to quantify heterogeneous tumor characteristics from imagining data, using CT scans of solid adrenal lesions as an example.

In the manuscript titled "*Improving the Explainability of Random Forest Classifier – User Centered Approach"*, **Petkovic** et al. propose a novel approach to explain complex models generated from one of the most popular machine learning classifier methods, random forests. Through their method and its application, authors provide an effortless way to generate summary reports of data to enhance the interpretability of complex random forest classifiers.

2.2 Application of methods to identify patterns in EHR data

EHR data consists of a wealth of information about patients. These datasets are present in forms of patient records on disease diagnosis, lab tests which include blood tests as well as imaging data, demographic information, medication information, as well as physicians' clinical notes. Patients' EHRs can be linked with their genetic data in a form of biobanks (for example Geisinger's Mycode Community Health Initiative, Vanderbilt's BioVU, eMERGE Network, UK BioBank)[24–26], which provides a great opportunity for uncovering novel disease associations and, ultimately, improving health care.

EHR data are a great source of phenotype information. The criteria used for assigning a disease status (case and control status) to a patient sample vary greatly across different studies. Some studies use extensive manual curation and development of a phenotypic algorithm to assign patients disease status, whereas other studies use instances of disease diagnosis codes (ICD-9 codes) to assign case-control status to patients[27,28]. High-throughput techniques to generate phenotypes are necessary to bridge the gap between the two techniques described above. In the manuscript titled *"Automated Disease-Cohort Selection using Word Embeddings from Electronic Health Records",* **Glicksberg** et al. address this problem by evaluating the performance of automated feature learning method, word2vec, with the established research-based electronic phenotyping algorithms in extracting cohorts for five diseases.

In manuscript titled *"Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database"*, **Beaulieu-Jones** et al. apply deep learning techniques to map patient time series data to the preventive care that a patient receives in the EHR. The challenge of utilizing dense longitudinal information from EHR data is addressed in this manuscript. Machine learning methods and the comparisons of methods highlighted in this paper also provide useful insights towards deep learning techniques and their applications in pattern identification.

In the manuscript titled "*Causal Inference on Electronic Health Records to Assess Blood Pressure Treatment Targets: An Application of the Parametric g Formula",* **Johnson** et al. use EHRs to extract longitudinal blood pressure information from patients suffering from hypertension. They use this data to demonstrate the utility of an established causal inference technique, parametric g-formula, for the first time in EHR data in the context of cardiovascular preventative medicine.

2.3 Applications in transcriptome and next-generation sequencing data

With the availability of next-generation sequencing data (NGS), research focused on pattern recognition for identification of functional elements in the human genome has become widespread. Analyzing gene expression information is helpful in understanding the influence of such elements on a trait or disease.

Jeong et al. hypothesized that analyzing transposable elements (TE), which for a long time had been incorrectly labeled as junk DNA, could provide useful functional insights for biomedical data. In their manuscript, *"An Ultra-Fast and Scalable Quantification Pipeline for Transposable Elements from Next Generation Sequencing Data"*, the authors propose a pipeline to quantify TE in the genome from NGS data. This pipeline could be useful for the biomedical informatics community to discover hidden association among TE expression and diseases.

One of the major goals of association analysis is to identify the proportion of phenotypic variance explained by genetic variations. Transcriptome-wide association analysis (TWAS)

are becoming popular methods in explaining the proportion of phenotypic variance that cannot be explained by single nucleotide variations alone[9,29,30]. In the paper titled *"How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures?",* **Veturi** et al. use a simulation study to analyze two major approaches for conducting TWAS, TWAS-MP (multi-SNP prediction) and TWAS-SMR (summary-based Mendelian Randomization). The paper describes a comprehensive power analysis for detecting gene-trait analysis, which in the future could be expanded to other kinds of omics datasets.

3. References

1. Pattern Recognition and Machine Learning | Christopher Bishop | Springer.
2. Minelli, M., Chambers, M. & Dhiraj, A. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. (John Wiley & Sons, 2012).
3. Iddamalgoda, L. et al. Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: Challenges and Implications. Front. Genet. **7,** (2016).
4. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. **19,** A68–A77 (2015).
5. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature **526,** 68–74 (2015).
6. Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. **12,** e1001779 (2015).
7. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. **39,** 1181–1186 (2007).
8. Bourne, P. E. et al. The NIH Big Data to Knowledge (BD2K) initiative. J. Am. Med. Inform. Assoc. JAMIA **22,** 1114 (2015).
9. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. **18,** 117–127 (2017).
10. Hadoop & Spark | IBM Big Data & Analytics Hub. Available at: /technology/hadoop-and-spark. (Accessed: 20th September 2017)
11. Hashem, I. A. T. et al. The rise of "big data" on cloud computing: Review and open research issues. Inf. Syst. **47,** 98–115 (2015).
12. Big Data is a Big Deal for Biomedical Research. whitehouse.gov (2013). Available at: https://obamawhitehouse.archives.gov/blog/2013/04/23/big-data-big-deal-biomedical-research. (Accessed: 20th September 2017)
13. Silva, T. C. & Zhao, L. Network-based high level data classification. IEEE Trans. Neural Netw. Learn. Syst. **23,** 954–970 (2012).
14. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. Bioinforma. Oxf. Engl. **27,** 431–432 (2011).
15. Barabási, A.-L. Network Medicine — From Obesity to the "Diseasome". N. Engl. J. Med. **357,** 404–407 (2007).
16. Goh, K.-I. et al. The human disease network. Proc. Natl. Acad. Sci. **104,** 8685–8690 (2007).
17. Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P. & Burgun, A. Integrating Heterogeneous Biomedical Data for Cancer Research: the CARPEM infrastructure. Appl. Clin. Inform. **7,** 260–274 (2016).

18. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the Missing Link for Big Biomedical Data. JAMA **311,** 2479–2480 (2014).
19. Deng, L. & Yu, D. Deep Learning: Methods and Applications. Found. Trends® Signal Process. **7,** 197–387 (2014).
20. Special Issue: Deep Learning for Biomedical and Health Informatics. Available at: https://jbhi.embs.org/2016/12/30/special-issue-deep-learning-biomedical-health-informatics/. (Accessed: 20th September 2017)
21. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene-gene interactions: a review. Appl. Bioinformatics **5,** 77–88 (2006).
22. Zheng, T. et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int. J. Med. Inf. **97,** 120–127 (2017).
23. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. **16,** 85–97 (2015).
24. Gottesman, O. et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. Off. J. Am. Coll. Med. Genet. (2013). doi:10.1038/gim.2013.72
25. Carey, D. J. et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. (2016). doi:10.1038/gim.2015.187
26. Development of a large-scale de-identified DNA biobank to enable personalized medicine. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18500243. (Accessed: 20th September 2017)
27. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat. Rev. Genet. **17,** 129–145 (2016).
28. Kirby, J. C. et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J. Am. Med. Inform. Assoc. JAMIA **23,** 1046–1052 (2016).
29. Gusev, A. et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. Am. J. Hum. Genet. **95,** 535–552 (2014).
30. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. **45,** 580–585 (2013).

Large-scale analysis of disease pathways in the human interactome

Monica Agrawal[1,‡], Marinka Zitnik[1,‡], and Jure Leskovec[1,2]
[1]*Department of Computer Science, Stanford University, Stanford, CA, USA*
[2]*Chan Zuckerberg Biohub, San Francisco, CA, USA*
[‡]*Equal contribution; Email: {agrawalm, marinka, jure}@cs.stanford.edu*

Discovering disease pathways, which can be defined as sets of proteins associated with a given disease, is an important problem that has the potential to provide clinically actionable insights for disease diagnosis, prognosis, and treatment. Computational methods aid the discovery by relying on protein-protein interaction (PPI) networks. They start with a few known disease-associated proteins and aim to find the rest of the pathway by exploring the PPI network around the known disease proteins. However, the success of such methods has been limited, and failure cases have not been well understood. Here we study the PPI network structure of 519 disease pathways. We find that 90% of pathways do not correspond to single well-connected components in the PPI network. Instead, proteins associated with a single disease tend to form many separate connected components/regions in the network. We then evaluate state-of-the-art disease pathway discovery methods and show that their performance is especially poor on diseases with disconnected pathways. Thus, we conclude that network connectivity structure alone may not be sufficient for disease pathway discovery. However, we show that higher-order network structures, such as small subgraphs of the pathway, provide a promising direction for the development of new methods.

Keywords: disease pathways, disease protein discovery, protein-protein interaction networks

1. Introduction

Computational discovery of disease pathways aims to identify proteins and other molecules associated with the disease.[1–3] Discovered pathways are systems of interacting proteins and molecules that, when mutated or otherwise altered in the cell, manifest themselves as distinct disease phenotypes (Figure 1A).[4] Disease pathways have the power to illuminate molecular mechanisms but their discovery is a challenging computational task. It involves identifying all disease-associated proteins,[2,5,6] grouping the proteins into a pathway,[7–10] and analyzing how the pathway is connected to the disease at molecular and clinical levels.[11,12] Many of the main challenges facing the task arise from the interconnectivity of a pathway's constituent proteins.[2,13–15] This interconnectivity implies that the impact of altering one protein is not restricted only to the altered protein, but can spread along the links of the protein-protein interaction (PPI) network[14] and affect the activity of proteins in the vicinity.[4,15]

As understanding each disease protein in isolation cannot fully explain most human diseases, numerous computational methods were developed to predict which proteins are associated with a given disease, and to bring them together into pathways using the PPI network (Figure 1B).[2,5–10,16,17] These methods have accelerated the understanding of diseases, but have not yet fully succeeded in providing actionable knowledge about them.[1] For example, recent studies[5,7,18] found that only a relatively small fraction of disease-associated proteins physically interact with each other, suggesting that methods, which predict disease proteins by searching for dense clusters/communities of interacting proteins in the network, may be limited in discovering disease pathways. Analytic methods may thus be hindered by such issues, and unless specifically tuned, can lead to an expensive and time-consuming hunt for new disease proteins. Furthermore, although numerous methods exist, protein-protein interaction and connectivity patterns of disease-associated proteins remain largely unexplored.[4,12] Because of the

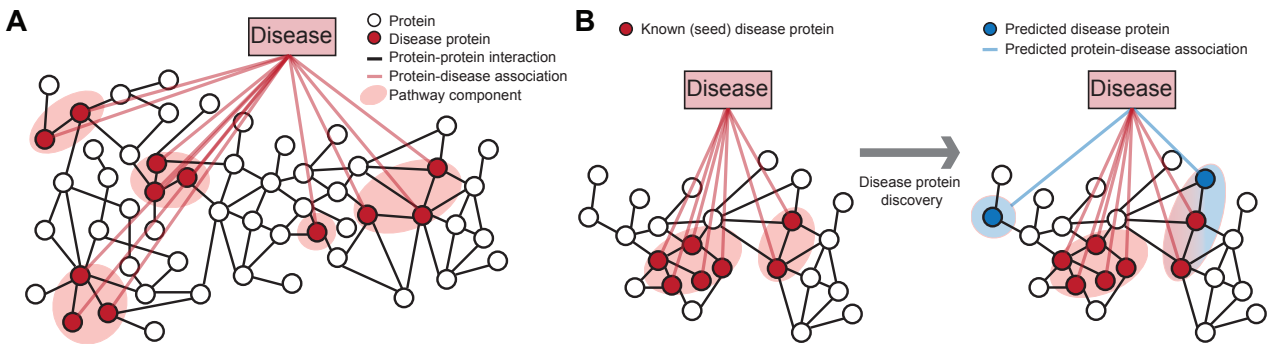


Fig. 1. **Network-based discovery of disease proteins. A** Proteins associated with a disease are projected onto the protein-protein interaction (PPI) network. In this work, *disease pathway* denotes a (undirected) subgraph of the PPI network defined by the set of disease-associated proteins. The highlighted disease pathway consists of five pathway components. **B** Methods for disease protein discovery predict candidate disease proteins using the PPI network and known proteins associated with a specific disease. Predicted disease proteins can be grouped into a disease pathway to study molecular disease mechanisms.

huge potential of these methods for the development of better strategies for disease prevention, diagnosis, and treatment, it is thus critical to identify broad conceptual and methodological limitations of current approaches.

Present work. Here we study the PPI network structure of 519 diseases. For each disease, we consider the associated disease proteins and project them onto the PPI network to obtain the disease pathway (Figure 1A).[1,4] We then investigate the network structure of these disease pathways.[a]

We show that disease pathways are fragmented in the PPI network with on average more than 16 disconnected pathway components per disease. Furthermore, we find that each component contains only a small fraction of all proteins associated with the disease. Through spatial analysis of the PPI network we find that proximity of disease-associated proteins within the PPI network is statistically insignificant for 92% (476 of 519) diseases, and that 90% of diseases are associated with proteins that tend not to significantly interact with each other, indicating that disease proteins are weakly embedded—rather than densely interconnected—in the PPI network.

We then consider state-of-the-art network-based methods for disease protein discovery (Figure 1B). These methods use the PPI network and a small set of known disease proteins to predict new proteins that are likely associated with a given disease. However, as we show here, current methods disregard loosely connected proteins when making predictions, causing many disease pathway components in the PPI network to remain unnoticed. In particular, we find that performance of present methods is better for diseases whose pathways have high edge density, are primarily contained within a single pathway component, and are proximal in the PPI network. However, our analysis shows that a vast majority of disease pathways does not display these characteristics.

The search for a solution to the better characterization of disease pathways has led us to study higher-order protein-protein interaction patterns[4,19–21] of disease proteins. Following on from earlier work[22] showing that higher-order PPI network structure around cancer proteins is different from the structure around non-cancer proteins, we find that many proteins associated with the same disease are involved in similar higher-order network patterns, even if disease proteins are not adjacent in the PPI network. In particular, we find that proteins associated with 60% (310 of 519) of diseases do exhibit

[a]All data and supplementary tables with results are at: `http://snap.stanford.edu/pathways`.

over-representation for certain higher-order network patterns, suggesting that disease proteins can take on similar structural roles, albeit located in different parts of the PPI network. We demonstrate that taking these higher-order network structures into account can shrink the gap between current and goal performance of disease protein discovery methods.

In addition to new insights into the PPI network connectivity of disease proteins, our analysis on network fragmentation of disease proteins and their distinctive higher-order PPI network structure leads to important implications for future disease protein discovery that can be summarized as:

- We move away from modeling disease pathways as highly interlinked regions in the PPI network to modeling them as loosely interlinked and multi-regional objects with two or more regions distributed throughout the PPI network.
- Higher-order connectivity structure provides a promising direction for disease pathway discovery.

2. Background and related work

Next, we give background on disease pathways and on methods for disease protein prediction.

Disease pathways. Broadly, a disease pathway in the PPI network is a system of interacting proteins whose atypical activity collectively produces some disease phenotype.[3,4,12,16] Given the PPI network $G = (V, E)$, whose nodes V represent proteins and edges E denote protein-protein interactions, the *disease pathway* for disease d is an undirected subgraph $H_d = (V_d, E_d)$ of the PPI network specified by the set of proteins V_d that are associated with d, and by the set of protein-protein interactions $E_d = \{(u, v)|(u, v) \in E \text{ and } u, v \in V_d\}$ (*e.g.*, Adrenal cortex carcinoma pathway in Figure 4). To measure the specifics of protein interactions within and outside the pathway we define pathway boundary as the set $B_d = \{(u, v)|(u, v) \in E, u \in V_d, v \in V \setminus V_d\}$ consisting of all edges that have one endpoint inside H_d and the other endpoint outside H_d.

Network-based methods for disease protein discovery. Given a specific disease, the task is to take the PPI network and the disease proteins and to predict new proteins that are likely associated with the disease. Approaches for this task are known as protein-disease association prediction or disease module detection methods (Figure 1B), and can be grouped intro three categories. (1) Neighborhood scoring and clustering methods[4,5,7,9,10,12] assume that proteins that belong to the same network cluster/community are likely involved in the same disease. In direct neighborhood scoring, each protein is assigned a score that is proportional to the percentage of its neighbors associated with the disease. To identify clusters that extend beyond direct neighbors, the methods start with a small set of disease proteins (seed proteins) and grow a cluster by expanding the seeds with the highest scoring proteins. However, few existing methods (*e.g.*, connectivity significance-based method DIAMOnD) can work with seed proteins that are not adjacent in the PPI network.[7] (2) Diffusion-based methods[5,8,23] use seed proteins to specify a random walker that starts at a particular seed protein and at every time step moves to a randomly selected neighbor protein. Upon convergence, the frequency with which the nodes in the network are visited is used to rank the corresponding proteins. (3) Representation learning methods,[6,16,17,21,24] such as matrix completion, graphlet degree signatures, and neural embeddings, construct representations for proteins (*i.e.*, latent factors, embeddings) that capture known protein-disease associations and/or proteins' network neighborhoods, and then use these representations as input to a downstream predictor. We consider a neural embedding approach[24] that first learns a vector representation for each protein using a single-layer neural network and random walks and then fits a logistic regression classifier that predicts disease proteins based on these feature vectors. We also consider a matrix completion method[6] that factorizes a protein-disease association matrix

into a set of protein and a set of disease latent factors while also incorporating the PPI network. Predictions for a new disease are obtained as a function of the feature and latent factors.

Although many methods exist for predicting disease proteins, surprisingly little is known about the PPI network structure of disease pathways and how it relates to the power of these methods.

3. Data

We continue by describing the datasets used in this study.

Human protein-protein interaction network. We use the human PPI network compiled by Menche *et al.*[18] and Chatr-Aryamontri *et al.*.[25] Culled from 15 databases, the network contains physical interactions experimentally documented in humans, such as metabolic enzyme-coupled interactions and signaling interactions. The network is unweighted and undirected with $n = 21{,}557$ proteins and $m = 342{,}353$ experimentally validated physical interactions. Proteins are mapped to genes and the largest connected component of the PPI network is used in the analysis. We also investigate two other PPI network datasets to make sure that our findings are not specific to the version of the PPI network we are using. Unless specified, results in the paper are stated with respect to the first dataset. The other two PPI networks are from the BioGRID database[25] and the STRING database.[26] Both of these networks are restricted to those edges that have been experimentally verified.

Protein-disease associations. A protein-disease association is a tuple (u, d) indicating that alteration of protein u is linked to disease d. Protein-disease associations are pulled from DisGeNET, a platform that centralized the knowledge on Mendelian and complex diseases.[2] We examine over 21,000 protein-disease associations, which are split among the 519 diseases that each has at least 10 disease proteins. The diseases range greatly in complexity and scope; the median number of associations per disease is 21, but the more complex diseases, *e.g.*, cancers, have hundreds of associations.

Disease categories. Diseases are subdivided into categories and subcategories using the Disease Ontology.[27] The diseases in the ontology are each mapped to one or more Unified Medical Language System (UMLS) codes, and of the 519 diseases pulled from DisGeNET, 290 have a UMLS code that maps to one of the codes in the ontology. For the purposes of this study, we examine the second-level of the ontology; this level consists of 10 categories, such as cancers (68 diseases), nervous system diseases (44), cardiovascular system diseases (33), and immune system diseases (21).

Altogether, we use human disease and PPI network information that is more comprehensive than in previous works,[7,18,22] which focused on smaller sets of diseases and proteins.

4. Connectivity of disease proteins in the PPI network

We start by examining the network connectivity of disease proteins. We then analyze disease protein discovery methods and contextualize their performance using disease pathway network structure.

4.1. *Proximity of disease proteins in the PPI network*

We begin by briefly describing network measures that we use to characterize connectivity of disease proteins, both within disease pathways and with respect to the rest of proteins in the PPI network.

PPI network distance and concentration measures. We consider the following measures to characterize PPI connectivity of disease proteins for each disease d and its associated pathway H_d:

- *Size of largest pathway component:* Fraction of disease proteins that lie in H_d's largest pathway component (*i.e.*, the relative size of the largest connected component (LCC) of H_d).

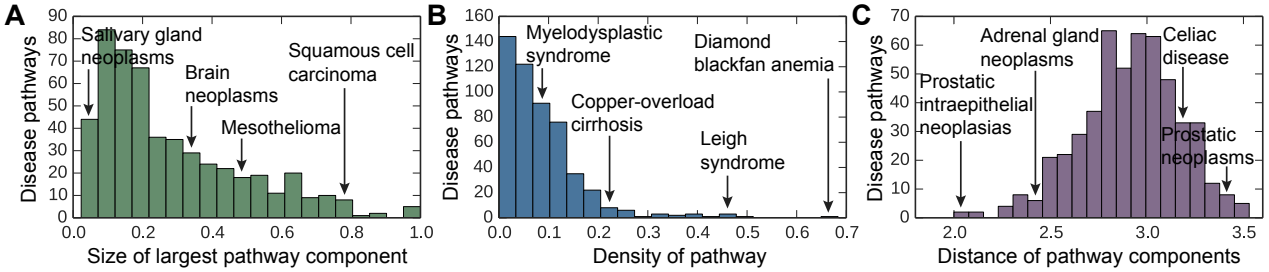


Fig. 2. **Protein interaction connectivity of disease pathways.** The distribution of (**A**) the network densities of each disease pathway, (**B**) the relative size of the largest pathway component calculated as a fraction of disease proteins that lie in the largest pathway component, and (**C**) the average shortest path length between disparate pathway components in the PPI network.

- *Density of the pathway:* It is calculated as: $2|E_d|/(|V_d|(|V_d|-1))$ and takes values in $[0,1]$. A higher density indicates a higher fraction of edges (out of all possible edges) appear between nodes in H_d.
- *Distance of pathway components:* For each pair of pathway components (Figure 1A), we calculate the average shortest path length between each set of proteins, and then, the average of this is taken over all pairs of the components.
- *Conductance:*[28] It is calculated as: $|B_d|/(|B_d|+2|E_d|)$ and takes values in $[0,1]$. A lower conductance indicates the pathway is a more well-knit community separated from the rest of the network.
- *Spatial network association:*[29,30] It measures concentration/localization of disease proteins in the PPI network by quantifying how strongly disease proteins co-cluster within the PPI network and whether this co-clustering is stronger than expected by random chance. It is calculated as: $K_d(s) = 2/(\bar{p}n)^2 \sum_i p_i \sum_j (p_j - \bar{p}) I(\ell_G(i,j) < s)$, where p_i is a binary indicator indicating if node i represents a d-associated protein, $\bar{p} = 1/n \sum_i p_i$, and $I(\ell_G(i,j) < s)$ equals 1 if the shortest path length between i and j is less than s and 0 otherwise. If all disease proteins lie in one PPI network region, most of them are found for small values of s, while for uniformly spread proteins in the PPI network $K_d(s)$ achieves larger values only for large values of s. The significance of H_d's concentration is determined by computing the area under the $K_d(s)$ curve[29] for d-associated proteins and comparing it to curves obtained by applying the same statistic to sets of random proteins.
- *Network modularity:*[31] Fraction of edges that fall within/outside the pathway minus the expected fraction if edges were randomly distributed: $Q_d = 1/(2m) \sum_{ij} (I((i,j) \in E) - \frac{k_i k_j}{2m}) \delta(p_i, p_j)$, where k_i is the degree of i, and $\delta(p_i, p_j)$ is 1 if p_i and p_j are equal and 0 otherwise.

PPI network structure of disease pathways. First, we find that disease pathways are fragmented in the PPI network, with a median of 16 connected components per disease and a median of only 21% of the proteins lying in the largest pathway component (Figure 2A). Only approximately 10% of pathways have over 60% of their proteins in the largest pathway component. We also find that disease pathways are not particularly well connected internally with only a median density of 0.07 (the overall PPI network density is 0.0015), and 90% of diseases have a density below 0.17 (Figure 2B). Furthermore, they are rather well connected externally, having a median conductance of 0.96, meaning that the disease pathway has relatively as many edges pointing outside the pathway to the rest of the PPI network as it has edges lying inside the pathway. The median distance between the pathway components is almost 2.9 (Figure 2C). These results counter expectations as they show that disease pathways do not have the PPI network structure one expects of a traditional network cluster/community, which is well connected internally and has few edges pointing outside the cluster.[14,31]

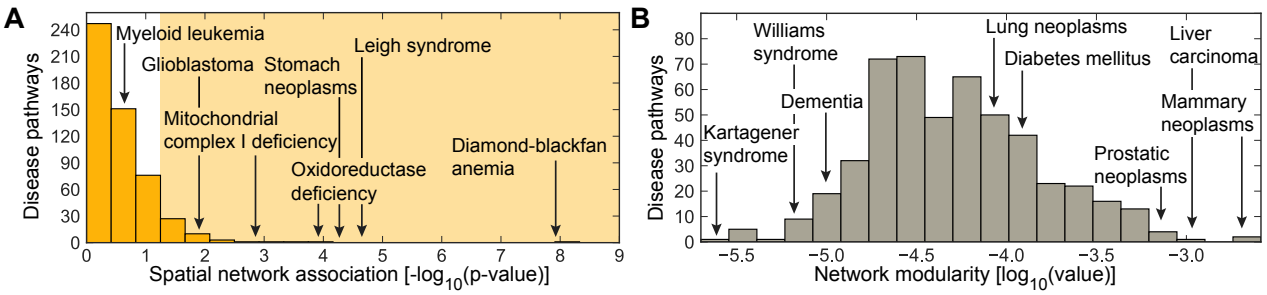


Fig. 3. **Spatial clustering and modular structure of disease pathways in the PPI network.** The distribution of (**A**) the spatial clustering calculated for each disease pathway as the strength of association[29] between the set of disease proteins and the PPI network (shaded area indicates significant spatial clustering at $\alpha = 0.05$ level), and (**B**) the modularity[31] of disease pathways in the PPI network.

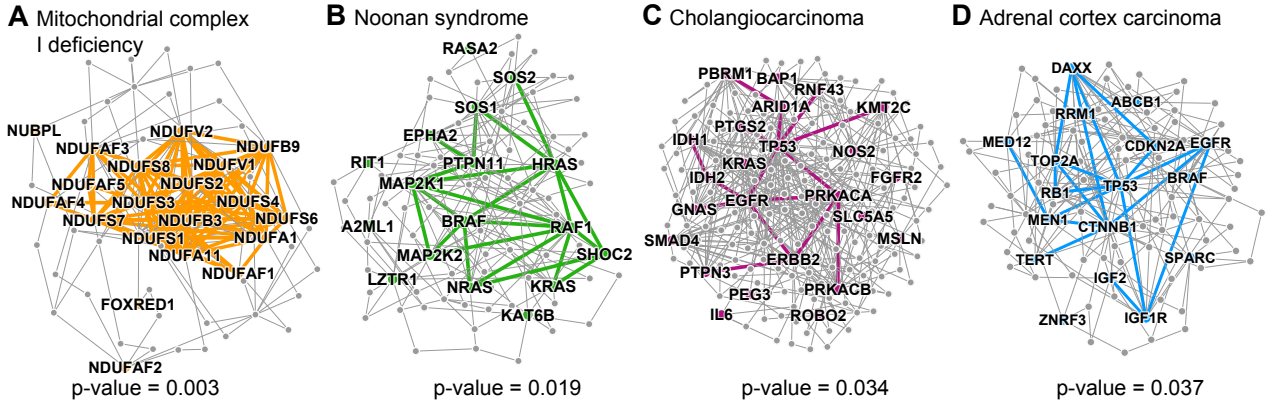


Fig. 4. **Disease pathways in the wider PPI network**. A small PPI subnetwork highlighting physical interactions between disease proteins associated with (**A**) Mitochondrial complex I deficiency, (**B**) Noonan syndrome, (**C**) Cholangiocarcinoma, and (**D**) Adrenal cortex carcinoma. Shown are selected disease pathways whose spatial clustering[29] within the PPI network is statistically significant (p-values shown; entire distribution of the p-values is shown in Figure 3A) and is also among the strongest (top-30 diseases) in the disease corpus.

To statistically test how well disease pathways are localized in the PPI network, we conduct spatial analysis of the PPI network. We find no significant pathway localization for 92% of diseases (Figure 3A), suggesting that these diseases have pathways that are multi-regional with two or more regions of disease proteins in different parts of the PPI network. The presence of multiple regions suggests that each disease might be comprised of several groups of proteins that are located in weakly connected or disconnected regions of the PPI network and thus may be functionally distinct.[1,4] We find that only 43 of 519 (8%) diseases are region-specific (examples in Figure 4), *i.e.*, they significantly associate with only one local neighborhood and can be found in a single region of the PPI network. We also observe that the number of edges within a disease pathway rarely exceeds the number expected on the basis of chance (Figure 3B). The median modularity of disease pathways is only 4.6×10^{-4}, reflecting there is no significant concentration of edges within disease pathways compared with random distribution of edges between all proteins regardless of pathways. These results suggest that integration of disconnected regions of disease proteins into a broader disease pathway will be crucial for a holistic understanding of disease mechanisms.

Finally, we note that these findings can be reproduced in three PPI network datasets (Section 3),

suggesting that our key results are robust against potential biases in the PPI network data.

4.2. *Connections between PPI network structure and disease protein discovery*

Next, we study disease protein discovery methods based on some of the most frequently used principles for identifying disease proteins,[14] and evaluate them through PPI pathway network structure.

Methods and experimental setup. We consider five methods: direct neighborhood scoring,[5] neural embeddings,[24] matrix completion,[6] network diffusion,[8,23] and connectivity significance (DIAMOnD).[7] See Section 2 for details on the methods. We use disease-centric ten-fold cross-validation. For each disease, the set of all proteins is randomly split into ten folds, with each fold containing an equal number of proteins associated with that disease. In each of the ten runs, the goal is to predict disease proteins in the test fold, assuming knowledge of disease proteins in the nine other folds. Each method assigns a score to each protein in the network representing the probability that the protein is associated with the disease. 20 diseases are set aside for hyperparameter selection, and the remaining 499 are used for testing. For evaluation, recall-at-k is measured to quantify what fraction of all the disease proteins are ranked within the first k predicted proteins (*e.g.*, $k = 25, 100$). We also calculate the mean reciprocal rank (MRR) for all of the algorithms in order to get an overall measure of method performance. Measures range between 0 and 1, and a higher score indicates better performance.

Prediction performance in the context of disease pathway structure. Figure 5 shows performance of disease protein discovery methods as a function of PPI connectivity of disease proteins. We observe that the higher the degree of agglomeration of disease proteins within the PPI network, the higher the performance of prediction methods. In particular, across all five methods, performance correlates positively with density and percent of proteins in the largest pathway component, and negatively with the distance between pathway components. These correlations are weaker for the Neural embeddings than for the other four methods, but the overall direction of the trends is the same across all of them. For example, the correlation between density and recall-at-100 is $\rho = 0.45$ for the Neural embeddings and between $\rho = 0.54$ and $\rho = 0.63$ for the other four methods (Figure 5B).

Across all diseases, random walk-based methods are the best performers, as evaluated by both mean reciprocal rank (MRR = 0.061 and MRR = 0.050 for Random walk and Neural embeddings, respectively) and recall (Recall-at-100 = 0.356 and Recall-at-100 = 0.300 for Random walk and Neural embeddings, respectively). However, we see that Random walk method is particularly dependent on the percent of disease proteins in the largest pathway component (Figure 5A). The difference in recall between the Random walk method and the Neural embeddings is positively correlated (ρ = 0.26) with that percentage. Since random walks and other diffusion-based variants are very popular, it is problematic that they are reliant on properties that are not typical of disease pathways.

Neighborhood scoring performs the worst by both metrics (MRR = 0.029, Recall-at-100 = 0.242). The superior performance of random walk-based methods indicates that the assumption that the neighborhood method makes in calculating scores based only on protein's neighbors is too restrictive when defining disease locality.[5] Though DIAMOnD does not outperform Random walk, we observe that it has a comparable recall in its higher-ranked predictions (recall-at-25 = 0.186, compared to Random walk's 0.199), but its performance lags considerably for lower-ranked predictions (Recall-at-100 = 0.300, compared to Random walk's 0.356).

We see the most complementarity between Neural embeddings, Matrix completion and the other methods, which makes sense given that the other three methods are all based on direct/indirect network neighborhood scoring, while Neural embeddings and Matrix completion more flexibly capture

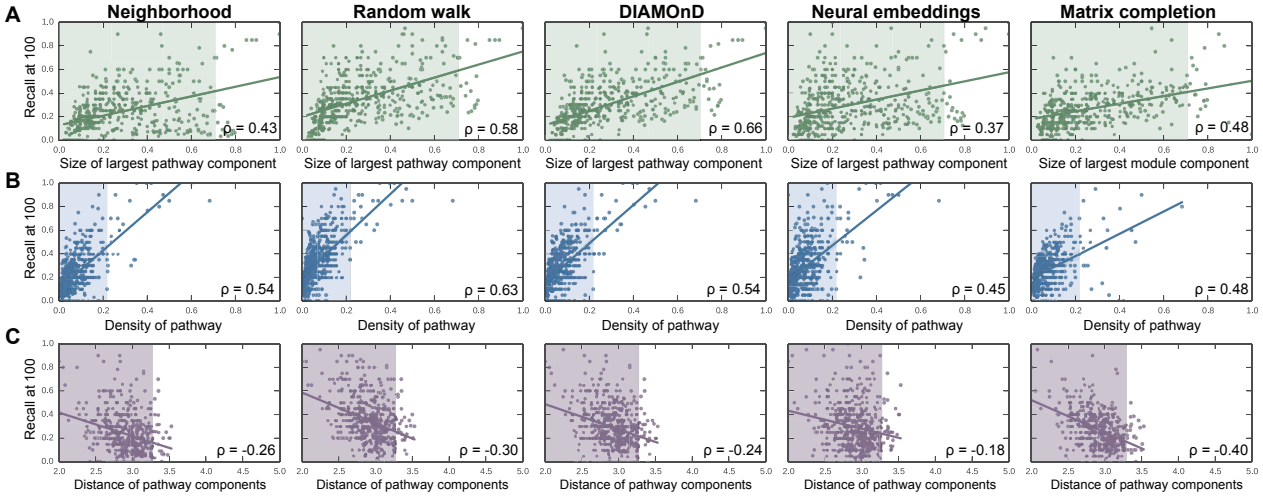


Fig. 5. **Prediction quality versus PPI connectivity of disease proteins.** Each point represents one disease; its location is determined by the quality of predicted disease proteins (y-coordinate), and by the connectivity of disease proteins in the PPI network (x-coordinate). Across all five methods, the trends uniformly indicate that (**A**) the bigger the largest pathway component, (**B**) the more densely interconnected the disease pathway, and (**C**) the lower the average shortest path length between disparate pathway components, the better the predictions. The shaded areas represent the space in which 95% (494 of 519) of all diseases reside.

network structure and neighborhoods of disease proteins.[6] For example, we can examine the disease pathway for Juvenile myelomonocytic leukemia in which Neural embeddings method performs far better than Random walk (Recall-at-100 = 0.550, compared to Random walk's 0.200). The pathway consists of nineteen nodes, but there exists only three edges within the pathway (network density = 0.008). Therefore, the Neural embeddings method is able to capture latent features about pertinent nearby nodes that Random walk struggles to find, given that Random walk is highly dependent on the edges near the seed proteins. On the other hand, the pathway for Squamous cell carcinoma is more accurately detected by Random walk than by Neural embeddings (Recall-at-100 = 0.540, compared to Neural embeddings' 0.120). This can be explained by higher interconnectivity of Squamous cell carcinoma pathway in the PPI network (network density = 0.034) suggesting that there are advantages to focusing on local edge connectivity.

Performance variation across disease categories. We observe strong differences in performance across disease categories indicating that diseases should not be considered as a monolithic category, given the very different mechanisms behind them. The same performance patterns hold across all five of the methods, suggesting that none of the assumptions each method makes about pathway structure correspond to any of the particular mechanisms that tend to be more specific to one disease category. Furthermore, because of the similar performance among methods over easy[32] (*e.g.*, median recall-at-100 = 0.720 for Mendelian diseases) and difficult[32] (*e.g.*, median recall-at-100 = 0.360 for cancer diseases) disease categories, the assumptions made by current methods do not seem to accurately reflect the uncertainty associated with a protein's true association with a disease.

5. Higher-order connectivity of disease proteins in the PPI network

We showed in Section 4 that proximity of disease proteins in the PPI network is likely insufficient for the disease protein discovery task as disease pathways have rather low PPI density and a rather high conductance. To look past just edge connectivity for the prediction of disease proteins, we in-

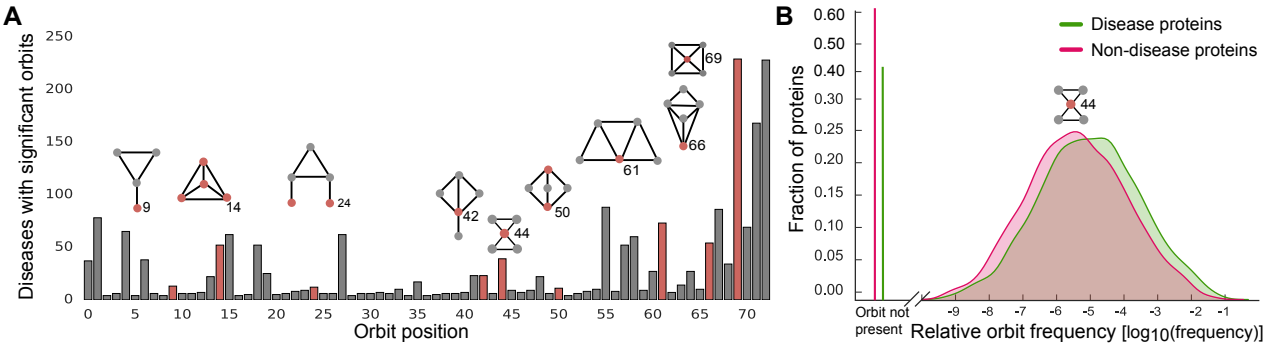


Fig. 6. **Over-representation of motifs in disease modules. A** The number of diseases (out of 519 possible) for which the associated proteins are significantly over-represented at each orbit position. A disease is deemed significant at a given orbit position if the median number of times a disease protein matching that position was significant at $\alpha = 0.01$, as compared to permutation testing over random sets of proteins of the same size. Pictured above are selected motifs (red node represents the orbit position, *i.e.*, the location where the node touches the motif). **B** The relative frequency distribution of orbit 44 for disease proteins (green) and non-disease proteins (red).

vestigate what higher-order PPI network structures disease proteins are likely to be involved in, and then incorporate this structure information to augment prediction capability of current methods.

Motif signatures of disease proteins. The analysis of higher-order PPI network structure can be formalized by counting network motifs, which are subgraphs that recur within a larger network. We here focus on graphlets,[19,20,22,33] connected non-isomorphic induced subgraphs (examples shown in Figure 6). There are 30 possible graphlets of size 2 to 5 nodes. The simplest graphlet is just two nodes connected by an edge, and the most complex graphlet is a clique of size 5. By taking into account the symmetries between nodes in a graphlet, there are 73 different positions or orbits for 2–5-node graphlets, numerated from 0 to 72. For each node in the PPI network we count the number of orbits that the node touches. Motif signature of a protein is thus a set of 73 numbers, h_i $(i = 0, 1, \ldots, 72)$ representing the number of induced subgraphs the corresponding node is in, in which the node took the i-th orbital position. We use this signature to represent protein's higher-order connectivity in the PPI network.

We conduct permutation tests, comparing the median of the orbit distribution values for proteins associated with a given disease to the medians for 5,000 random samples of sets of proteins of the same size as the disease pathway. These values are used to calculate p-values for each disease at each orbit position. An orbit position for a disease is considered significant if there is an over-representation of counts in the disease proteins compared to the 99% of random samples (*i.e.*, $\alpha = 0.01$).

Characterization of motifs around disease proteins. We find that there is a characteristic higher-order PPI network structure around disease proteins (Figure 6), indicating that disease proteins display significance in terms of the orbit positions they tend to inhabit, which could point towards the underlying mechanisms they participate in. We see that 60% (310 of 519) of diseases do show orbit signatures that differ from background proteins and are significantly greater than what one would expect at random. Therefore, even though proteins associated with these diseases may not be adjacent in the PPI network, the diseases do show overall over-representation for certain orbit positions indicating that proteins in disease pathway may take on similar structural roles, albeit in non-adjacent regions in the PPI network.

We can note that orbit position 0 corresponds to subgraphs of two nodes (only an edge), 1 through

Table 1. **Examples of disease-associated motifs.** Shown are 6 orbits (orbit position, *i.e.*, the location where the node touches the motif, is shown in red) whose over-representation is found in most diseases.

Significant orbit	# of diseases	Examples of diseases with significant orbits
	78	Neoplastic cell transformation, Celiac disease, Non-small cell lung carcinoma, Squamous cell carcinoma, Prostatic neoplasms
	62	Neoplastic cell transformation, Stomach neoplasms, Restless legs syndrome, Celiac disease, Prostatic neoplasms
	62	Neoplastic cell transformation, IGA glomerulonephritis, Precancerous conditions, Prostatic neoplasms, Liver neoplasms
	88	Peroxisome biogenesis disorders, Crohn disease, Mitochondrial encephalomyopathies, Venous thromboembolism, Myopia
	229	Amphetamine-related disorders, Mitochondrial myopathies, Cocaine-related disorders, Nuclear cataract, Polycystic ovary syndrome
	228	Amphetamine-related disorders, Leber congenital amaurosis, Craniofacial abnormalities, Hyperalgesia, Respiratory hypersensitivity

3 correspond to subgraphs of three nodes, 4 through 14 correspond to subgraphs of four nodes, and the rest correspond to induced subgraphs of five nodes. Therefore, although the distribution of smaller subgraphs such as nodes and triangles are not significant in many diseases, almost 50% of diseases have disease pathways that contain proteins that are over-represented for the most complex orbit positions. For example, orbit position 14 is statistically significantly over-represented in approximately 50 diseases, and position 72 is in over 200 of the diseases (Figure 6, examples in Table 1). We note that we also statistically test for under-representation of the orbit counts, no statistically significant results are observed.

Characterization of motifs for disease categories. We also want to investigate whether the orbit signatures are characteristic of diseases in general, or whether there are also differences that could be attributed back to the category the disease belongs to. In order to test this, for each of the 73 orbit positions and for each disease category, we find the statistical significance of the difference in distributions using a two-sample Kolmogorov-Smirnov test. The first sample consists of all the orbit counts for the proteins that are found in at least one disease in a given category, and the other sample consists of all the orbit counts for the proteins that are not associated with any disease in the category. After applying the Bonferroni correction, we then consider a p-value of $\alpha = 0.01$ to be significant.

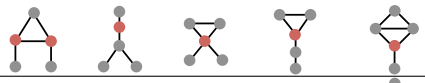
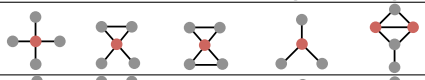
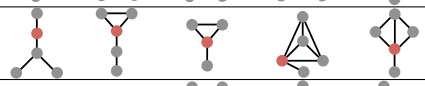
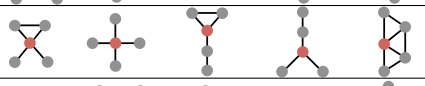
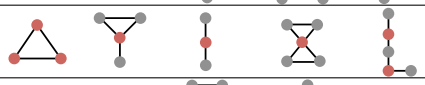
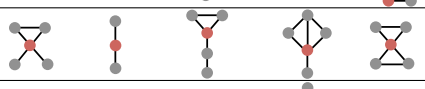
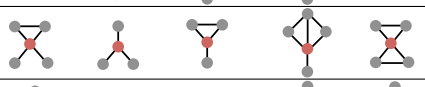
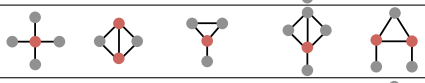
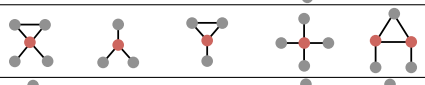
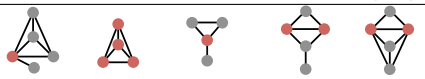
We find that the most significant differences tend to occur in the more complex motifs. Table 2 shows orbit positions that are considered most characteristic for each disease category, and the graphlets they appear in, which could indicate inherent differences in the manifestations of different classes of diseases.

6. Prediction of disease proteins using higher-order PPI network structure

Higher-order PPI network structure is generally not taken into account in current disease protein discovery, although, as showed in Section 5, this structure does encode distinct information about disease proteins. Earlier work showed that motif signatures provide useful signal for biological function prediction,[22,33] but here we want to examine whether they provide additional signal past what edge connectivity in the PPI network already contributes, and if they specifically work for disease protein discovery task.

Setup and results. We conduct a logistic regression experiment in which we augment the Neural

Table 2. **Characteristic motifs for disease categories.** Shown are 5 orbits whose over-representation is found in most diseases belonging to a disease category. The orbit position of a node is marked in red.

Disease category	Significant orbits	Orbit positions
Urinary system diseases *e.g.*, Hyperhomocysteinemia, Nephrosis		26, 20, 33, 30, 47
Acquired metabolic diseases *e.g.*, Methylmalonic acidemia, Hyperinsulinism		23, 33, 44, 7, 48
Monogenic diseases *e.g.*, Marfan syndrome, Bardet-Biedl syndrome		20, 30, 11, 42, 58
Cancer *e.g.*, Tumor of salivary gland, Papillary thyroid carcinoma		33, 23, 30, 21, 61
Gastrointestinal system diseases *e.g.*, Eosinophilic esophagitis, Oral fibrosis		3, 11, 2, 44, 16
Inherited metabolic disorders *e.g.*, Leigh disease, Mitochondrial complex deficiency		33, 2, 30, 42, 44
Immune system diseases *e.g.*, Deficiency syndromes, Hypersensitivity		33, 7, 11, 42, 44
Musculoskeletal system diseases *e.g.*, Muscular atrophy, Muscular dystrophy		23, 13, 11, 42, 26
Nervous system diseases *e.g.*, Peripheral neuropathy, Nerve degeneration		33, 7, 11, 23, 26
Cardiovascular system diseases *e.g.*, Dilated cardiomyopathy, Tachycardia		58, 14, 11, 48, 67

embeddings (Section 4.2) with information about motif signatures (Section 5). In particular, for each protein we concatenate its neural embedding with its motif signature. Instead of using the motif signature directly, we concatenate the embedding with h', where $h'_i = \max(0, \log(h_i))$, for $i = 0, 1, \ldots, 72$.

We find that neural embeddings augmented by motif signatures performed on average 11% better than neural embeddings alone (Recall-at-100 = 0.332, compared to Neural embeddings' 0.300). For example, in the case of Hearing loss, the disease that has the greatest increase in performance after the inclusion of higher-order structure, we observe that the recall-at-100 jumps from 0.03 to 0.77 (and the recall-at-100 is at most 0.10 for the four other prediction methods in Section 4.2). If we examine the signature of Hearing loss, as calculated in Section 5, we see that the Hearing loss pathway is significant across all 73 orbit positions, meaning it has a particularly unique signature compared to the background distribution. Though such improvement in performance is not typical across all diseases, this analysis identifies the opportunity to systemically identify diseases which are likely to benefit the most from the inclusion of higher-order PPI network information.

7. Conclusion

The overall goal of network biology is to develop approaches that use genomic and other network information to better understand human disease. Given the complexity of this goal, we focused on studying the PPI network structure of disease pathways, defined through sets of proteins associated with diseases. We found that disease pathways are fragmented and sparsely embedded in the PPI network, and that spatial clustering of disease pathways within the PPI network is statistically insignificant. To better understand broad caveats of current methodology for disease protein discovery we

evaluated the performance of leading methods and found that their assumptions do not fully capture PPI network structure. We showed, however, that there is detectable higher-order PPI network structure around disease proteins that can be leveraged to boost algorithm performance. These findings provide new insights into the disease pathway PPI network structure and can guide methodological advances in disease protein discovery.

Acknowledgments

This research has been supported in part by NSF IIS-1149837, NIH BD2K, DARPA SIMPLEX, Stanford Data Science Initiative, and Chan Zuckerberg Biohub.

References

1. M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim, *Nature Reviews Genetics* **16**, 85 (2015).
2. J. Piñero *et al.*, *Database* **2015** (2015).
3. M. Gustafsson *et al.*, *Genome Medicine* **6**, p. 82 (2014).
4. P. Creixell *et al.*, *Nature Methods* **12**, p. 615 (2015).
5. S. Navlakha and C. Kingsford, *Bioinformatics* **26**, 1057 (2010).
6. N. Natarajan and I. S. Dhillon, *Bioinformatics* **30**, i60 (2014).
7. S. D. Ghiassian, J. Menche and A.-L. Barabási, *PLoS Computational Biology* **11**, p. e1004120 (2015).
8. H. Zhou and J. Skolnick, *Bioinformatics* **32**, 2831 (2016).
9. Y. Silberberg, M. Kupiec and R. Sharan, *Genome Medicine* **9**, p. 48 (2017).
10. S. van Dam *et al.*, *Briefings in Bioinformatics* **bbw139**, 1 (2017).
11. A. Sharma *et al.*, *Human Molecular Genetics* (2015).
12. A. Krishnan, J. N. Taroni and C. S. Greene, *Current Genetic Medicine Reports* **4**, 155 (2016).
13. J. Loscalzo and A.-L. Barabási, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **3**, 619 (2011).
14. A.-L. Barabási, N. Gulbahce and J. Loscalzo, *Nature Reviews Genetics* **12**, 56 (2011).
15. L. I. Furlong, *Trends in Genetics* **29**, 150 (2013).
16. M. Zitnik and B. Zupan, *Bioinformatics* **32**, 90 (2016).
17. M. Zitnik and B. Zupan, *Pacific Symposium on Biocomputing* **21**, p. 81 (2016).
18. J. Menche *et al.*, *Science* **347**, p. 1257601 (2015).
19. N. Pržulj, D. G. Corneil and I. Jurisica, *Bioinformatics* **22**, 974 (2006).
20. N. Pržulj, *Bioinformatics* **23**, 177 (2007).
21. K. Sun, J. P. Gonçalves, C. Larminie and N. Pržulj, *BMC Bioinformatics* **15**, p. 304 (2014).
22. T. Milenković *et al.*, *Journal of the Royal Society Interface* **7**, 423 (2010).
23. M. D. Leiserson *et al.*, *Nature Genetics* **47**, 106 (2015).
24. A. Grover and J. Leskovec, *ACM SIGKDD* **22**, 855 (2016).
25. A. Chatr-Aryamontri *et al.*, *Nucleic Acids Research* **43**, D470 (2015).
26. D. Szklarczyk *et al.*, *Nucleic Acids Research* **43**, 447 (2015).
27. W. A. Kibbe *et al.*, *Nucleic Acids Research* **43**, D1071 (2014).
28. S. E. Schaeffer, *Computer Science Review* **1**, 27 (2007).
29. A. J. Cornish and F. Markowetz, *PLoS Computational Biology* **10**, p. e1003808 (2014).
30. A. Baryshnikova, *Cell Systems* **2**, 412 (2016).
31. M. E. Newman, *Proceedings of the National Academy of Sciences* **103**, 8577 (2006).
32. J. Loscalzo, I. Kohane and A.-L. Barabási, *Molecular Systems Biology* **3**, p. 124 (2007).
33. W. Hayes, K. Sun and N. Pržulj, *Bioinformatics* **29**, 483 (2013).

Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database*

Brett K. Beaulieu-Jones[1], Patryk Orzechowski[1,2] and Jason H. Moore[1]

[1]*Computational Genetics Lab, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia PA, 19104, United States of America*

[2] *Department of Automatics and Biomedical Engineering, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland*

Email: brettbe@med.upenn.edu

Electronic Health Records (EHRs) contain a wealth of patient data useful to biomedical researchers. At present, both the extraction of data and methods for analyses are frequently designed to work with a single snapshot of a patient's record. Health care providers often perform and record actions in small batches over time. By extracting these care events, a sequence can be formed providing a trajectory for a patient's interactions with the health care system. These care events also offer a basic heuristic for the level of attention a patient receives from health care providers. We show that is possible to learn meaningful embeddings from these care events using two deep learning techniques, unsupervised autoencoders and long short-term memory networks. We compare these methods to traditional machine learning methods which require a point in time snapshot to be extracted from an EHR.

Keywords: Electronic Health Records, Deep Learning, Patient Trajectories, Longitudinal, Unsupervised, Autoencoders, Long Short Term Memory Networks.

* This work is supported by Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health. B.K.B.-J., P.O. and J.H.M. were also supported by US National Institutes of Health grants AI116794 and LM010098.

1. Introduction

After the U.S. government mandated meaningful use of electronic health records (EHRs) by 2014, they have been widely adopted with 96% of health care providers implementing an EHR [1]. Patient interactions with the health care system are recorded in the EHR. Many research analyses treat the EHR as a static document by taking a snapshot of a patient`s EHR and using this for downstream analyses. This fails to account for the way a patient changes over time, their trajectory.

Jensen et al. [2] proposed the idea of temporal disease trajectories to model expected progression for a patient over time. This study uses billing codes as disease labels, which may introduce biases inherent to the billing process. Patients may be assigned a billing code before being diagnosed for a disease in order to receive a diagnostic test. Billing codes place also binary rules on the presence of disease. Perhaps most importantly for this work billing codes are frequently assigned after a visit and are thus not helpful for tracking patient trajectories over the course of an inpatient admission or rapid series of visits.

Interactions between patients and the health care system tend to occur in bursts, related to a specific visit or a series of visits. We label these periods of activity as care events and group these actions together. These care events represent changes over time and can capture longitudinal changes of a patient`s state.

Denny et al. [3] first showed the ability to use autoencoders to model clinical measures in an unsupervised manner. More recently, several groups have used autoencoders to learn high level features useful for classification [4,5] and imputation [6]. Tan et al. also showed the ability to extract meaningful features from gene expression data using autoencoders [7]. We use autoencoders to represent patient care events in a low dimensional vector space that is useful for visualization. Positions in this vector space represent the patient`s condition at a point in time. By connecting these positions, or care events, in order, it is possible to see how a patient`s condition changes over time and how they move through the health system. It is also possible to cluster patients in this low dimensional space and examine when patient outcomes diverge, one group having high survival and the other having high mortality.

This care event representation also provides a natural sequence of events. Recurrent neural networks have shown an impressive ability to model sequences to solve problems in many domains including object recognition in computer vision [8], image [9] and text generation [10]. Long short-term memory networks (LSTMs) [11] are a type of recurrent neural network that have recently been applied to clinical data to learn low dimension representations of medical concepts [12] and to make classifications using time series of specific clinical measures [13,14].

Trajectories have been used to model multistage dynamic decision processes (DMP) in discrete optimization problems [15]. In Algebraic Logical Meta-Model (ALMM) the state of the system in a certain time depends on the previous state, undertaken decision and transition function. This concept allows to easily describe the state of the patient at a particular time, with specific actions taken (e.g. application of medication) to manage the response to previous events within the progression of a disease.

In this work, we first demonstrate that deep learning approaches can (1) learn patient embeddings useful for both interpretable expert analysis via visualization and (2) do this we use the Medical Information Mart for Intensive Care III (MIMIC) database and apply both unsupervised deep autoencoders and LSTMs.

2. Methods

2.1. *Source Code and Analysis Availability*

Source code to reproduce the analyses in this work are provided in our repository (https://github.com/EpistasisLab/MIMIC_trajectories) under a permissive open source license. In addition, Continuous Analysis [16] was used to generate docker images matching the environment of the original analysis.

2.2. *Care Event Extraction*

2.2.1. *Medical Information Mart for Intensive Care III (MIMIC) Critical Care Database*

MIMIC [17] is a publicly available database composed of 46,297 critical care de-identified electronic health records for patients at Beth Israel Deaconess Medical Center. It includes all charted data (demographics, vital signs, medications, procedures, diagnoses, patient outputs, laboratory tests, physician notes, and treatment details) for patients from 2001 to 2012.

2.2.2. *Extracting Care Events from MIMIC*

We divided the MIMIC database into 4 groups:

1. Static data that does not change over the course of an admission (i.e. demographic data).
2. Actions performed by health care providers that have a specific time associated with them (i.e. laboratory events).
3. Actions performed by health care providers that only have a date associated with them (i.e. oral medications).
4. Streaming data measured on a per-minute basis (i.e. heart rate).

Table 1. *Categories and examples of Care Event Actions.*

Category (MIMIC Database Table)	Example
DATETIMEEVENT	Changing equipment or standard repeated treatments (i.e. dialysis).
ICUSTAYS	Transfer to or from the Intensive Care Unit.
INPUTEVENTS_CV & INPUTEVENTS_MV	Any fluids given to the patient (i.e. an IV solution, CV and MV stand for the two systems used to track these events Philips Carevue and iMDSoft Metavision).
LABEVENTS	All lab measurements for a patient (i.e. Creatinine level).
PROCEDUREEVENTS	All procedures performed on a patient (i.e. Extubation).
SERVICES	Changes in which service a patient is under (i.e. Cardiac Surgery)

To define care events, we included all actions initiated, or charted, by health care providers that have a specific time associated with them (Table 1). These actions were placed in sequential order and grouped together until there was a gap greater than the margin time (Figure 1). Because this is critical care data, the timeline between events is much smaller than typical EHR data. We found a 59 minute margin time yielded care events that had a good balance of inclusiveness while not including extended time periods. This yielded 1,566,026 total care events and an average of 26.80 care events per admission. In outpatient datasets, we expect a margin time of several days may better capture the concept of a care event.

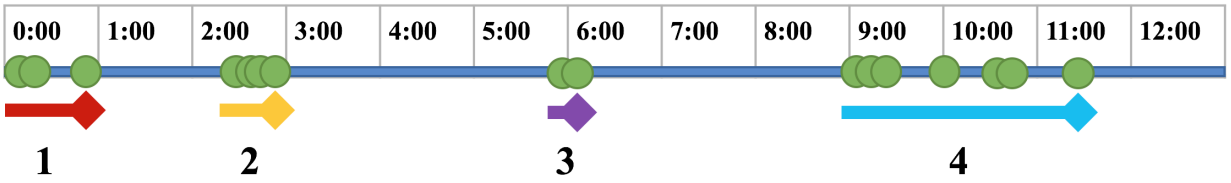


Fig. 1. Example of care event extraction. Green circles indicate actions taken by health care providers. Lines and numbers below indicate care events.

2.2.3. *Stratification of Patient Attention based on type of Insurance Provider*

Care events can provide a useful heuristic to the level of interaction between the health care provider and a patient. To evaluate attention, we compared the time spent in the hospital per admission with the number of care events per admission and the average number of care events per day. We then performed Welch's t-test between patients with private insurance and each of the other types of insurance (Medicare, Medicaid, Government, Self-Payment) to see if there were significant differences between patients with differing insurance types.

2.3. *Unsupervised learning to learn embeddings of extracted Care Events*

2.3.1. *Applying Autoencoders to Extracted Care Events to cluster in a low dimensional space.*

We used the Keras library [18] to construct autoencoders with 7 hidden layers in (1196, 512, 256, 128, 64, 128, 256, and 512 nodes per layer). We used dropout to mask 20% of the connections between the input layer and the first hidden layer. The model was trained using binary cross entropy loss with Adam [19]. The middle, hidden layer (64 nodes) was used as an output for visualization using t-Stochastic Neighbor Embedding [20]. The resulting visualizations were labeled for enrichment of 1-year patient survival. Survival data was based on the date of death variable in the MIMIC dataset, a merger between the hospital and social security data.

2.4. *Predicting Survival Using Care Events*

We evaluated how effectively different machine learning methods could predict patient survival over a 1-year period (as measured from the original admission date). The 1-year survival period began on the date of admission.

For this analysis, we performed 5-fold cross validation providing a training set of 46,751 admissions and a test set of 11,687 admissions chosen via stratified cross validation [21]. Survival was predicted using several classifiers: (1) a standard feed forward or multi-layer perceptron deep neural network [18], (2) a random forest, (3) logistic regression and (4) support vector machine (linear kernel) [21] after various numbers (*N*) of care events: 1, 3, 5, 10, 20, 30 and 50. Area under the curve of the receiver operating characteristic was used for evaluation and comparison.

2.4.1. *Traditional machine learning methods to predict survival from an EHR Snapshot.*

To build a snapshot vector useable for traditional machine learning methods. We took the mean of each value from a set of care events, up to the N^{th} care event. If the patient had less than *N* care events, we took the mean for all of their care events. Feature selection was then performed on this aggregate vector using ReliefF [22,23] to choose the top 100 features. These 100 features were then provided as input to each of the machine learning classifiers.

2.4.2. *Long Short Term Memory Networks (LSTMs) to predict survival with Care Events Sequences*.

To build the sequence vector from a set of care events we first truncated sequences longer than *N*. Sequences shorter than N were post-padded with zeros. The model was comprised of 3 types of layers, an initial embedding layer, three LSTM layers (with 100, 50 and 50 nodes respectively) and a fully connected (Dense) output layer. We trained the model using rmsprop [24] with a binary cross entropy loss function.

3. Results

The MIMIC dataset includes 58,438 admissions from 46,297 unique patients. This was extracted to form 1,566,026 care events (Table 2). Medicare patients were double the age of other patients on average. Patients using private or government insurance and Medicaid had relatively equal mortalities during the initial admission and the next 6 months. Patients using Medicare had significantly higher mortality in the 6 months after admission as their time under critical care and self-payment patients had high mortality during the admission but lower admission after leaving critical care.

3.1. *Treatment and Outcome Comparison*

Table 2. Summary statistics for MIMIC Critical Care database.

	Total	Male	Female	Private	Medicare	Medicaid	Government	Self
Patients	46,297	26,121	20,399	19,663	21,002	4,570	1,614	600
Admissions	58,438	32,950	26,026	22,250	28,103	5,713	1,767	605
Admissions per Patient	1.26	1.26	1.28	1.13	1.34	1.25	1.09	1.01
Average Age at Admission	56.01	54.95	57.34	37.82	75.95	37.91	35.15	39.11
Care Events	1,566,026	867,941	698,085	637,968	693,254	179,182	46,722	8,900
Care Events per Admission	26.80	26.34	26.82	28.67	24.67	31.36	26.44	14.71
Visit Survival	90.84%	90.38%	89.56%	95.24%	86.40%	94.60%	95.87%	85.2%
6-Month Survival	79.81%	79.63%	78.39%	89.84%	69.47%	87.61%	91.62%	83.31%
12-Month Survival	76.28%	76.19%	74.82%	87.90%	64.33%	84.75%	90.32%	82.81%

We examined the length of stay per admission by insurance type (Figure 2A) and found that patients using Medicare had the longest stays but that all groups differed significantly via an ANOVA test (p-value 5.02E-28). In addition, we compared each type of insurance against the private group using Welch's t-test. It is not surprising that patients using self-payment had the shortest stays and the least number of care events per stay (Figure 2B). Interestingly, patients with private insurance had significantly lower care events per day than the most similar (by age) other groups, government-based insurance and Medicaid (Figure 2B).

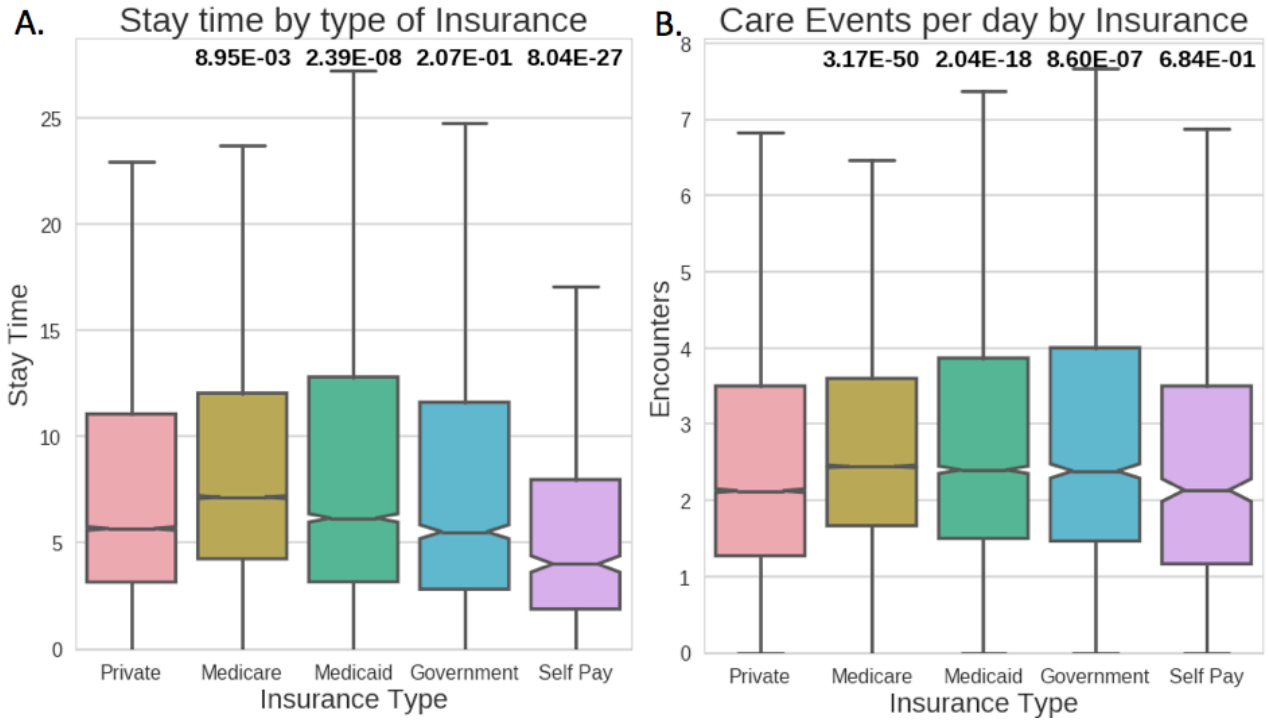


Fig. 2. Association testing between different insurance types. A.) Length of admission. C) Number of care events per day of each admission. Labels at the top indicate p-values via Welch's t-test to private group.

3.2. *Unsupervised modeling of patient care events*

To test whether unsupervised autoencoders could learn meaningful embeddings from individual care vents, we plotted the innermost hidden layer using t-Stochastic Neighbor Embedding (t-SNE) and overlaid 1-year survival labels (Figure 3). Figure 3 shows an unsupervised clustering, where the X and Y axes do not have an explicit meaning or interpretation. This clustering process yielded several clusters with high enrichment for either mortality or survival indicating the ability to learn meaningful embeddings. t-SNE does not maintain global similarity structure and as such this process is useful for visualizing single care events but not for understanding patient trajectories. To examine patient trajectories, it is necessary to look at the value of the innermost hidden layer before t-SNE was applied or to use a method designed to model sequential data. Recurrent neural networks, and specifically LSTMs are well suited at this task.

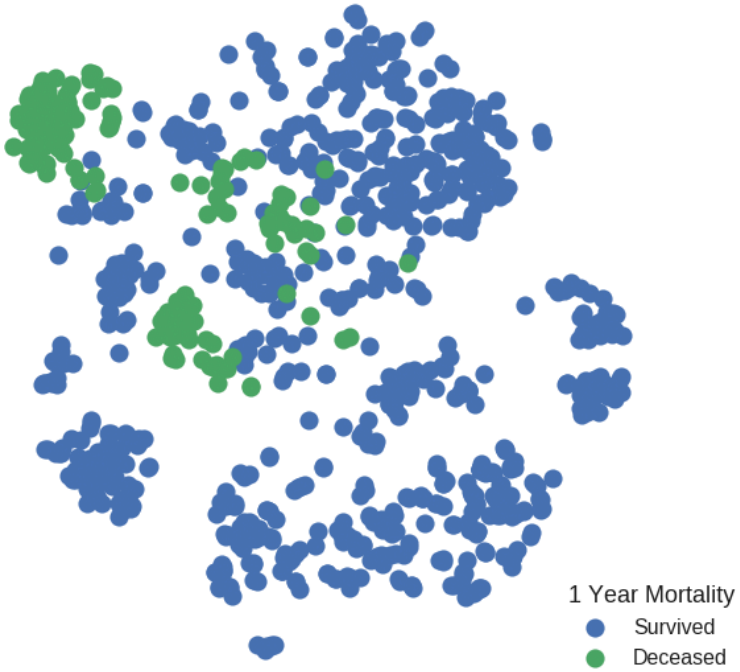


Fig. 3. Unsupervised Care Event Embedding by applying t-SNE to the innermost layer of autoencoder (1000 care events shown to prevent overplotting).

3.3. *Supervised prediction of patient survival*

Next, we preformed the supervised classification task of predicting whether a patient survived one year from the date of their admission. We measured classification accuracy with differing numbers of care events to evaluate whether the care event-based approach had advantages over traditional single point in time measurements (Figure 4). The Pearson correlation of the number of care events to 1-year mortality rate was 0.062. Of the methods predicting based on a snapshot, the random forest was by far the most effective. Despite this, it did not increase in performance as more information about an admission was added. This indicates that much of its predictive power comes from the initial presentation. Both, linear methods and a traditional feed-forward neural network barely outperformed random chance. This may have been due to the high dimensionality of the dataset. The care event-based LSTM increases in performance as more care events are provided. This is particularly evident when more than the median number of care events (26.8) are provided as input to the LSTM. Including more than 50 care events yielded weaker results for the LSTM. This is likely because most patients have fewer than 50 care events so most of the signal is captured in the first 50 care events. Going beyond 50 leads to a high level of padding to signal.

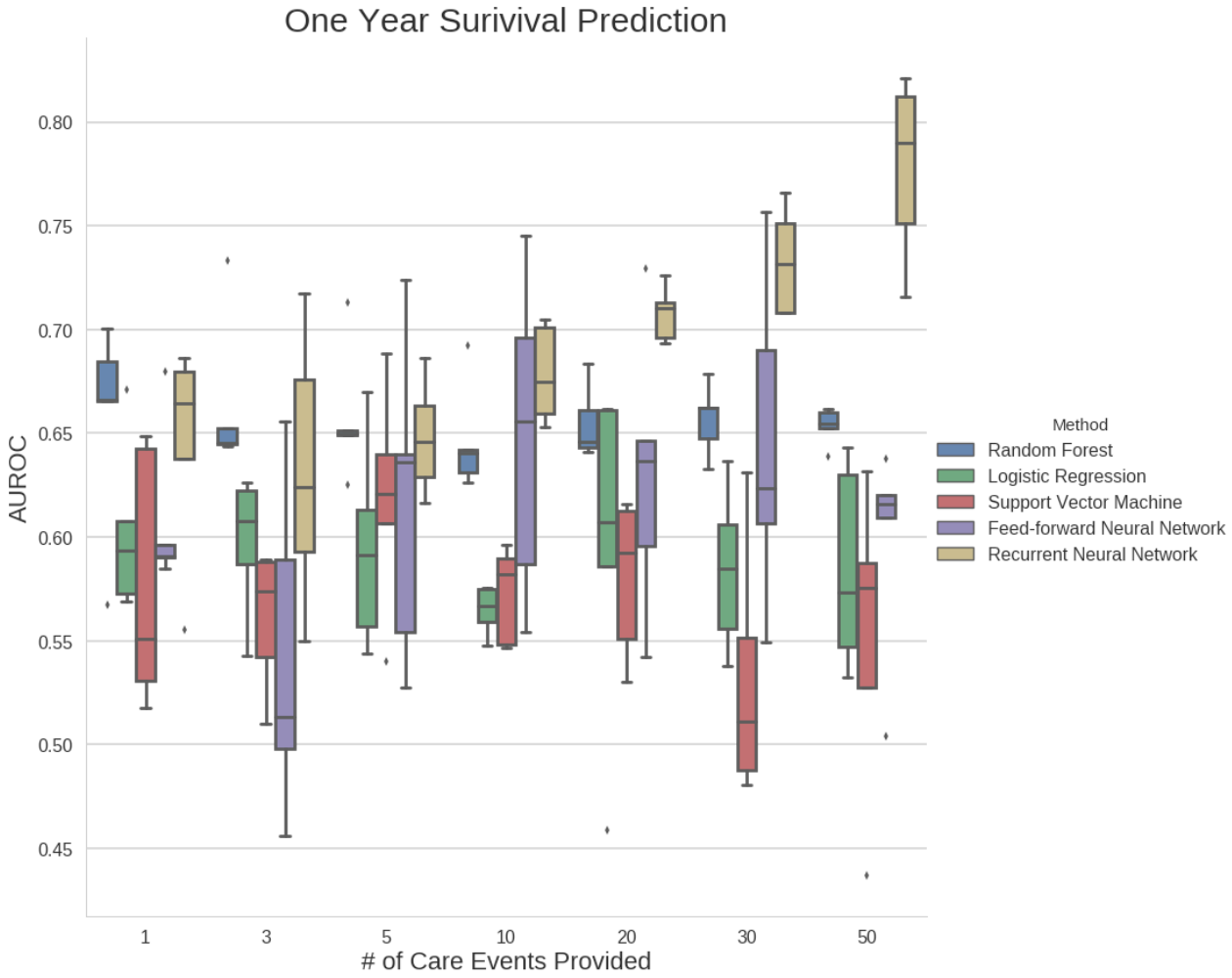


Fig. 4. Comparison of machine learning methods and the number of care events provided for 1-year survival prediction (AUROC).

4. Discussion and Conclusions

By limiting the usage of summary statistics to small time periods, we offer a granular method for modeling longitudinal clinical data. The care event extraction method provides a simple data driven approach to extracting temporal data for use in time series analyses. It allows summary statistics to be computed over short time windows as opposed to an entire patient history or arbitrary timestamps. Care events also offer a heuristic to allow comparison of the level of attention different patients receive from health care providers. We demonstrated the ability to learn embeddings enriched for different endpoints using unsupervised deep learning and were able to more accurately predict patient survival using supervised long short-term memory networks.

Though our approach showed strong performance for several tasks in this dataset, this method currently has limitations in terms of generalization. Long-short term memory networks, like many deep learning approaches, require many patients to outperform other methods. This can present a challenge when studying a single phenotype instead of a wide variety of critical care patients. The greatest benefits are likely to be seen when patients have many care events, making this approach

particularly well suited for chronic diseases like type 2 diabetes and Crohn's disease or for diseases that are hard to subtype such as multiple sclerosis. An additional challenge is if a patient with a disease like type 2 diabetes suffers an unrelated acute injury (i.e. broken rib in a vehicle accident) this acute injury may introduce too much noise to capture the type 2 diabetes trajectory.

In future work, we hope to introduce filtering techniques to exclude or deemphasize unrelated diagnoses. We also plan to increase the dimensionality of the encoders and applying additional techniques of visual clustering [25]. This includes using Shared-Nearest Neighbors (SNN) clustering to find groups of patients with similar stage of the disease in noisy data and Mukres algorithm to map groups of patients resembling a state of the disease to clusters found in the data.

Another challenge we would like to take is including streaming data in the simulation. Some measurements, e.g. heart rate or blood pressure, are performed every minute for each patient. The information about sudden changes of patient's condition is especially relevant for intensive-care patients. While our method aggregates patient data over shorter time periods than are commonly used, we plan to adapt our model by adding more detailed relevant information extracted from streaming sources.

5. Acknowledgments

We thank Casey S. Greene (University of Pennsylvania), Daniel S. Herman (University of Pennsylvania) and Andrew L. Beam (Harvard Medical School) for their helpful discussions. Funding: This work was supported by the Commonwealth Universal Research Enhancement (CURE) Program grant from the Pennsylvania Department of Health. B.K.B.-J., P.O. and J.H.M. were also supported by US National Institutes of Health grants AI116794 and LM010098 to J.H.M.. Author Contributions: B.K.B.-J. and J.H.M. conceived of the study. B.K.B.-J. and P.O. performed initial data processing. B.K.B.-J. performed analyses and wrote the manuscript. All authors revised and approved the final manuscript. Competing Interests: The authors have no competing interests to disclose. Source code availability: All source code is available via github (https://github.com/epistasislab/MIMIC_trajectory).

References

[1] J. Henry, Y. Pylypchuk, T. Searcy, V. Patel, Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2015, Coord. Heal. …. (2016). https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf (accessed July 27, 2017).

[2] A.B. Jensen, P.L. Moseley, T.I. Oprea, S.G. Ellesøe, R. Eriksson, H. Schmock, et al., Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nat. Commun. 5 (2014) 1769–1775. doi:10.1038/ncomms5022.

[3] T.A. Lasko, J.C. Denny, M.A. Levy, Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data, PLoS One. 8 (2013) e66341. doi:10.1371/journal.pone.0066341.

[4] B.K.B.K. Beaulieu-Jones, C.S. Greene, Semi-supervised learning of the electronic health record for phenotype stratification, J. Biomed. Inform. 64 (2016) 168–178. doi:10.1016/j.jbi.2016.10.007.

[5] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, Sci. Rep. 6 (2016) 26094. doi:10.1038/srep26094.

[6] B.K. Beaulieu-Jones, J.H. Moore, MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS, Pac. Symp. Biocomput. 22 (2016).
[7] J. Tan, M. Ung, C. Cheng, C.S. Greene, Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders., Pacific Symp. Biocomput. 20 (2015) 132–43. http://www.ncbi.nlm.nih.gov/pubmed/25592575.
[8] J. Ba, V. Mnih, K. Kavukcuoglu, Multiple Object Recognition with Visual Attention, (2014). http://arxiv.org/abs/1412.7755 (accessed July 28, 2017).
[9] K. Gregor, I. Danihelka, A. Graves, D. Jimenez Rezende, D. Wierstra, DRAW: A Recurrent Neural Network For Image Generation, (n.d.). https://arxiv.org/pdf/1502.04623.pdf (accessed July 28, 2017).
[10] I. Sutskever, J. Martens, G. Hinton, Generating Text with Recurrent Neural Networks, (n.d.). http://www.cs.utoronto.ca/~ilya/pubs/2011/LANG-RNN.pdf (accessed July 28, 2017).
[11] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.
[12] Y. Choi, C.Y.-I. Chiu, D. Sontag, Learning Low-Dimensional Representations of Medical Concepts., AMIA Jt. Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci. 2016 (2016) 41–50. http://www.ncbi.nlm.nih.gov/pubmed/27570647 (accessed July 28, 2017).
[13] Z.C. Lipton, D.C. Kale, R.C. Wetzel, Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks, (n.d.). https://arxiv.org/pdf/1510.07641.pdf (accessed July 28, 2017).
[14] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks, (2015). http://arxiv.org/abs/1511.03677 (accessed July 28, 2017).
[15] E. Dudek-Dyduch, Algebraic logical meta-model of decision processes-new metaheuristics, Int. Conf. Artif. Intell. (2015). http://link.springer.com/chapter/10.1007/978-3-319-19324-3_48 (accessed August 3, 2017).
[16] B.K.B. Beaulieu-Jones, C.C.S. Greene, Reproducibility of computational workflows is automated using continuous analysis, Nat Biotech. 35 (2017) 342–346. https://www.nature.com/nbt/journal/v35/n4/abs/nbt.3780.html (accessed June 18, 2017).
[17] A. Johnson, T. Pollard, L. Shen, L. Lehman, MIMIC-III, a freely accessible critical care database, Scientific. (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878278/ (accessed August 4, 2017).
[18] F. Chollet, Keras, GitHub, 2015. http://203.195.193.174/nat123CacheFolder/646F63732E626470742E6E6574/35c3a8a4cb1d4160bfd7b6e93d74ab67CD30CE37D036D032DF31CE3ACC30C533C9_e22880a46b0f1f3f3eb1e14dd5452984/media/pdf/kerascn/latest/kerascn.pdf#page=59 (accessed June 18, 2017).
[19] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014). http://arxiv.org/abs/1412.6980 (accessed July 28, 2017).
[20] L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605. doi:10.1007/s10479-011-0841-3.
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine Learning in Python, … Mach. Learn. …. 12 (2012) 2825–2830. http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490.
[22] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF, Appl. Intell. 7 (1997) 39–55. doi:10.1023/A:1008280620621.
[23] C.S. Greene, N.M. Penrod, J. Kiralis, J.H. Moore, Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions., BioData Min. 2 (2009) 5. doi:10.1186/1756-0381-2-5.
[24] T. Tieleman, G.E. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, (2012) 26–31.
[25] P. Orzechowski, K. Boryczko, Parallel approach for visual clustering of protein databases, Comput. Informatics. (2012). http://www.cai.sk/ojs/index.php/cai/article/viewArticle/140 (accessed August 4, 2017).

OWL-NETS: Transforming OWL Representations for Improved Network Inference

Tiffany J. Callahan*, William A. Baumgartner Jr., Michael Bada
Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA
**E-mail: tiffany.callahan@ucdenver.edu*

Adrianne L. Stefanski
Department of Pulmonary Sciences and Critical Care, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

Ignacio Tripodi
Interdisciplinary Quantitative Biology, University of Colorado Boulder, Boulder, CO 80309, USA

Elizabeth K. White, Lawrence E. Hunter
Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

Our knowledge of the biological mechanisms underlying complex human disease is largely incomplete. While Semantic Web technologies, such as the Web Ontology Language (OWL), provide powerful techniques for representing existing knowledge, well-established OWL reasoners are unable to account for missing or uncertain knowledge. The application of inductive inference methods, like machine learning and network inference are vital for extending our current knowledge. Therefore, robust methods which facilitate inductive inference on rich OWL-encoded knowledge are needed. Here, we propose OWL-NETS (NEtwork Transformation for Statistical learning), a novel computational method that reversibly abstracts OWL-encoded biomedical knowledge into a network representation tailored for network inference. Using several examples built with the Open Biomedical Ontologies, we show that OWL-NETS can leverage existing ontology-based knowledge representations and network inference methods to generate novel, biologically-relevant hypotheses. Further, the lossless transformation of OWL-NETS allows for seamless integration of inferred edges back into the original knowledge base, extending its coverage and completeness.

Keywords: Biological Ontologies; Knowledge Bases; Semantic Web; Machine Learning.

1. Introduction

Network representations facilitate the understanding of complex biological mechanisms, and have been used extensively in biomedical research to represent phenomena ranging from metabolism, to protein-protein interactions, and drug-drug interaction networks.[1–3] Inference over the structure of a network can provide insight and generate hypotheses regarding the functional relationships between network elements.[4]

The Web Ontology Language (OWL) is a Semantic Web standard for a network-based

knowledge representation and reasoning framework that is highly expressive and that has been used to model complex biological knowledge.[5] Inference on the Semantic Web has heavily relied on deductive and probabilistic reasoning. Deductive OWL reasoners work by inferring logical consequences from a set of explicitly asserted facts.[6] Constrained by first-order predicate logic, description logic reasoners (e.g., ELK,[7] FaCT++[8]) are unable to account for uncertainty or incomplete knowledge.[9] To account for this limitation, probabilistic methods (e.g., PR-OWL,[10] P-CLASSIC[11]) have been developed.[12] Unfortunately, these methods can only be applied in situations where the accuracy of propositions is ambiguous rather than unknown due to incomplete knowledge.[13] While inductive methods, like machine learning, are powerful tools for producing predictions that are not explicitly asserted,[12] the scalability of ontology properties can significantly limited the utility of these techniques.[14]

Link prediction, an inductive learning method, predicts unobserved connections between the nodes of a network. Most biological network representations can be assumed to be incomplete, making link prediction a potentially valuable tool for knowledge discovery. The application of such algorithms to biological networks has correctly predicted important relationships, including protein-protein interactions,[15] drug-target pairs,[16] and regulatory gene interactions.[17] Although OWL is a highly expressive representation language,[18] its use comes at the cost of a structurally complex network. We hypothesize that the expressivity of OWL reduces the power of link prediction algorithms to identify novel, biologically important insights. We further hypothesize that an abstraction of OWL networks will negate this power loss, providing a novel means to infer knowledge from available OWL resources.

In the field of biomedical informatics, abstraction networks have been used to obtain an alternative view of a terminology/ontology by reducing the complexity of its underlying structure.[19,20] Primarily developed to assess the quality of clinical terminologies and ontologies, these methods leverage the underlying terminology/ontology structure to combine subsets of nodes with similar attributes (e.g., data or object properties, and relations).[19,21–23] To the best of our knowledge, there are no existing abstraction methods designed to create network representations from complex OWL-encoded knowledge for the purpose of network inference.

We propose OWL-NETS (NEtwork Transformation for Statistical learning), a novel computational method that reversibly abstracts OWL-encoded biomedical knowledge into a network representation tailored for network inference. Using several examples built with the Open Biomedical Ontologies, we demonstrate that OWL-NETS results in networks with significantly different properties than their corresponding OWL representations. We also show that OWL-NETS can be used to leverage existing ontology-based knowledge representations and network inference methods to generate novel, biologically-relevant hypotheses. Further, the lossless transformation of OWL-NETS allows for seamless integration of inferred edges back into the knowledge base, extending its coverage and completeness.

2. Methods

OWL-NETS is implemented in Python (v2.7) and can be run from a simple GUI or from the command line. While primarily developed for use with OWL, the program can be easily extended for use with other Semantic Web technologies by modifying two primary assumptions:

(1) A knowledge source contains identifiers that directly represent biologically meaningful concepts (e.g., GO:0001525 is the identifier for the biological process of angiogenesis). Within OWL-NETS, biologically meaningful concepts are referred to as 'NETS nodes'.

(2) A knowledge source uses restrictions to specify the existence of biologically important relations between pairs of biological concepts (e.g., proteins are restricted to participate in angiogenesis). In OWL, restrictions provide a way to make the definition of a class more specific (e.g., proteins, specifically, protein kinases, participate in phosphorylation). Within OWL-NETS, biologically important relations are referred to as 'NETS edges'.

The OWL-NETS methodology is described using an example query that investigates disease-associated proteins that participate in angiogenesis (Figure 1; pseudocode for Steps 1-4 in Supplementary Material). The method takes a SPARQL query as input and outputs a directed OWL-NETS abstraction network, directed OWL representation, or both. To improve processing efficiency, the majority of the computational workload is performed on the input SPARQL query (Steps 1-3) rather than the resulting output.

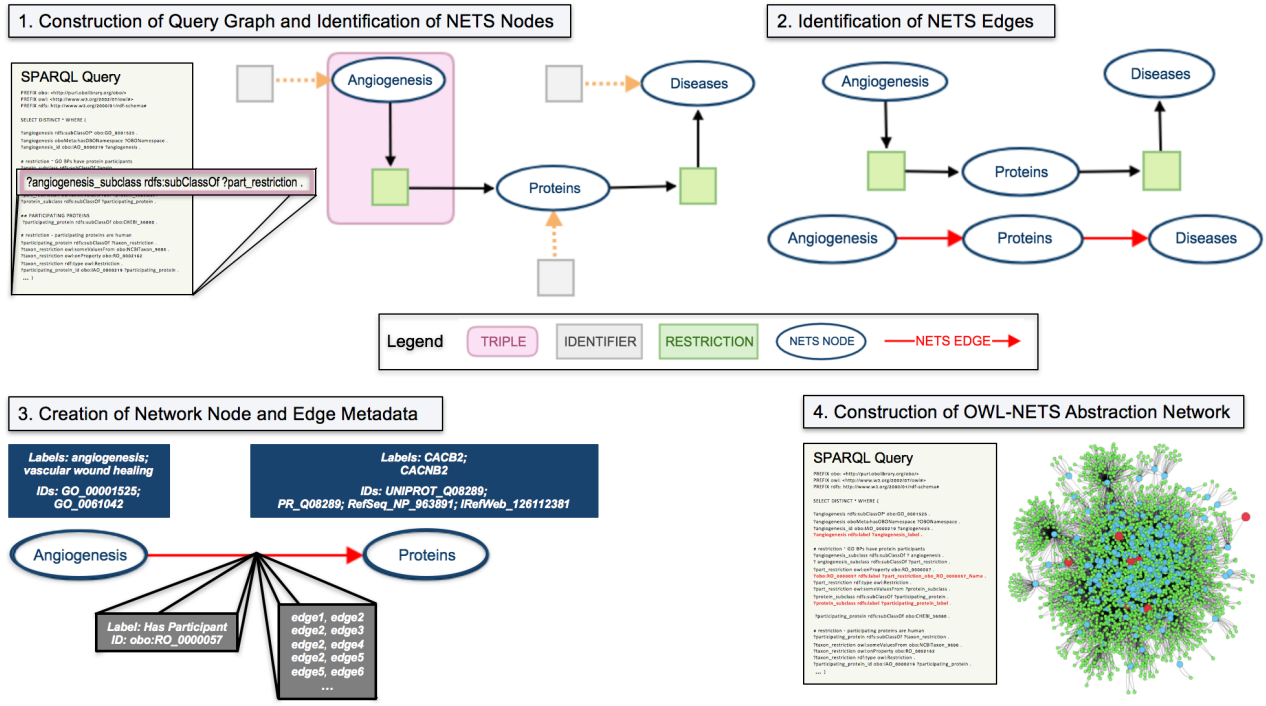


Fig. 1. Overview of the OWL-NETS Method.

(1) Construction of Query Graph and Identification of NETS Nodes: The input SPARQL query is used to create a directed query graph, where each triple (one of which is identified in Figure 1 in the query and in the graph with by a pink box) represents a directed edge in the query graph. The query graph is then searched for NETS nodes (Figure 1 shows identifier nodes pointing to NETS nodes via dashed orange arrows). NETS nodes in Figure 1 include: *Angiogenesis, Proteins, and Diseases.*

(2) Identification of NETS Edges: The query graph is searched for restrictions (shown in Figure 1 as light green nodes). The NETS node that is reachable from the restriction node's out-edges is the target of the NETS edge. In Figure 1, one of the green restriction nodes, that is pointed to by *Proteins* has an out-edge that points towards *Diseases*, thus an arrow is drawn from *Proteins* to *Diseases*. NETS edges are shown in the figure with a red arrow. When both NETS nodes can be reached from a restriction node, arrows pointing in both directions are drawn between the NETS nodes (i.e., nodeA → nodeB and nodeB → nodeA).

(3) Creation of Network Node and Edge Metadata: Identifiers and labels for each NETS node and edge (shown in Figure 1 as dark blue and gray boxes) are stored as network metadata. This metadata is needed to transform the OWL-NETS abstraction network back into the OWL representation which facilitates the seamless integration of inferred edges back into the knowledge base, extending its coverage and completeness.

(4) Construction of OWL-NETS Abstraction Network: Steps 1-3 gather information from the query graph that is needed to construct the OWL-NETS abstraction network. The final step augments the original query with this information. The red lines shown in the example SPARQL query, under Step 4, demonstrate the addition of NETS node and edge metadata to the query. The augmented SPARQL query is then run against an endpoint. The endpoint results are then used to construct the OWL-NETS abstraction network.

Supplemental material (including definitions and acronyms used throughout the paper), source code, and example data for exploring OWL-NETS can be found on GitHub (https://callahantiff.github.io/owl-nets/).

2.1. *Biomedical Use Cases*

To test the utility of OWL-NETS, we make use of the Knowledge Base Of Biology (KaBOB),[24] an open-source ontology-based semantically integrated knowledge base of biomedical data. Currently, KaBOB contains 13 sources of biomedical data on humans as well as seven model organisms in a representation grounded in 17 Open Biomedical Ontologies (OBOs). The queries that led to the development of OWL-NETS provide the use cases for the current work. The queries, along with their data sources (all data downloaded on March 2016, except Reactome which was downloaded on November 2015) are described below:

- Query 1: Human proteins localized to cellular and extracellular components and locations (Uniprot, Gene Ontology).
- Query 2: Protein targets of drugs that interact with trametinib (Uniprot, Protein Ontology, RefSeq, IRefWeb, DrugBank).
- Query 3: Protein targets of 100,000 drug-drug interactions and the pathways in which the proteins participate (Uniprot, DrugBank, Reactome).

Differences in the network properties of the OWL representation and OWL-NETS abstraction networks were explored using 100 protein localization networks (Query 1), each with 50 proteins, generated uniformly at random. Mann-Whitney U tests were then used to determine if the mean of each network property, measured across the 100 networks, significantly

differed by network representation. In addition to generating basic network properties, the power-law fit of the complementary cumulative distribution function (CCDF) was calculated for the network representations generated from Queries 2 and 3. All network properties were calculated on undirected network representations.

2.2. *Link Prediction Procedures*

Given an undirected, unweighted network $G(N, L)$, where N is the set of nodes and L is the set of observed edges between these nodes, the universal set of all possible edges U is $\frac{|N|*(|N|-1)}{2}$. The set of nonexistent edges (i.e., the set of edges that don't currently exist in the network) is U - L. From the observed network G, a uniformly random set of edges $L^{testing}$ was removed and the remaining edges $L^{training}$ were used as the training network.[25,26] Each link prediction algorithm was then run over the training network. The ability of each algorithm to recover the edges that were purposefully removed, $L^{testing}$ using only the information present in $L^{training}$, was then evaluated. The fraction of edges removed from the original network included 0.05, 0.10, 0.30, 0.50, 0.70, 0.90, and 0.95. For each removed fraction of edges, 100 iterations were run.

Ten similarity-based link prediction algorithms were run on networks resulting from Queries 2 and 3. In general, link prediction algorithms assign a similarity score to all non-observed edges in a network. The predicted edges with the highest scores are the most likely to exist.[26] Both local (i.e., node-level) and global (i.e., path-level) similarity link prediction algorithms were evaluated. The details regarding these algorithms are provided in Section 2 of Supplementary Material.

2.2.1. *Evaluation of Link Prediction Algorithm Performance*

The area under the receiver operating characteristic curve (AUC)[27] and top-L precision were used to evaluate link prediction algorithm performance.[26]

- AUC: The probability that a randomly chosen predicted edge that was purposely removed (true positive) has a higher score than a randomly chosen nonexistent predicted edge (true negative), where n' is the number of comparisons for which the randomly chosen true positive was higher than the randomly chosen true negative, n'' is the number of comparisons for which the randomly chosen true positive and true negative had the same score, and n is the total number of comparisons:

$$AUC = \frac{n' + 0.5n''}{n} \tag{1}$$

- Top-L Precision: Given a list of predicted edges sorted by score, the ratio of edges that were purposely removed L_{TP} (true positives) among all predicted edges L:

$$Precision = \frac{L_{TP}}{L} \tag{2}$$

The best performing algorithm was chosen using the highest average AUC (over 100 iterations) when removing a fraction of 0.5 edges from the original network. The time to run 100 iterations of each link prediction algorithm on each of the network representations from Query 2 was

also evaluated. Algorithms were run in parallel on a machine running macOS Sierra with 16 GB of RAM and a 2.7 GHz processor with 8 cores.

2.2.2. *Evaluation of Inferred Edges*

The best-performing link prediction algorithm was re-run (this time exposing the algorithm to all existing edges in the network) and the highest-scoring edges for each network were evaluated via expert consultation and extensive literature review by a PhD-level biologist (author ALS). Additionally, an OWL reasoner (HermiT,[28] via Protégé v5.1.1) was run on the OWL representation from Query 2 to demonstrate that deductive inference resulted in different predicted assertions than the link prediction algorithms.

3. Results

3.1. *Comparison of Network Properties*

Properties of the network representations from Query 1 are shown in Table S1 of Supplementary Material. On average, the OWL representation networks had significantly more nodes and edges, larger diameters, higher heterogeneity, a larger number of shortest paths, shorter average path lengths, more disassortative structures, and more cliques than the OWL-NETS abstraction networks. In contrast, the OWL-NETS abstraction networks had a larger average degree and a greater average clustering coefficient. Network properties are defined in Supplementary Material.

The OWL representation and OWL-NETS abstraction networks built from running Query 2 are visualized in Figure 2. The OWL representation network (left) contained 840 nodes and 1,426 directed edges. In comparison, the OWL-NETS abstraction network (middle) contained 59 nodes and 65 directed edges. The OWL-NETS abstraction network had a smaller average degree (2.00 vs. 3.42) than the OWL representation networks. As shown in the third plot

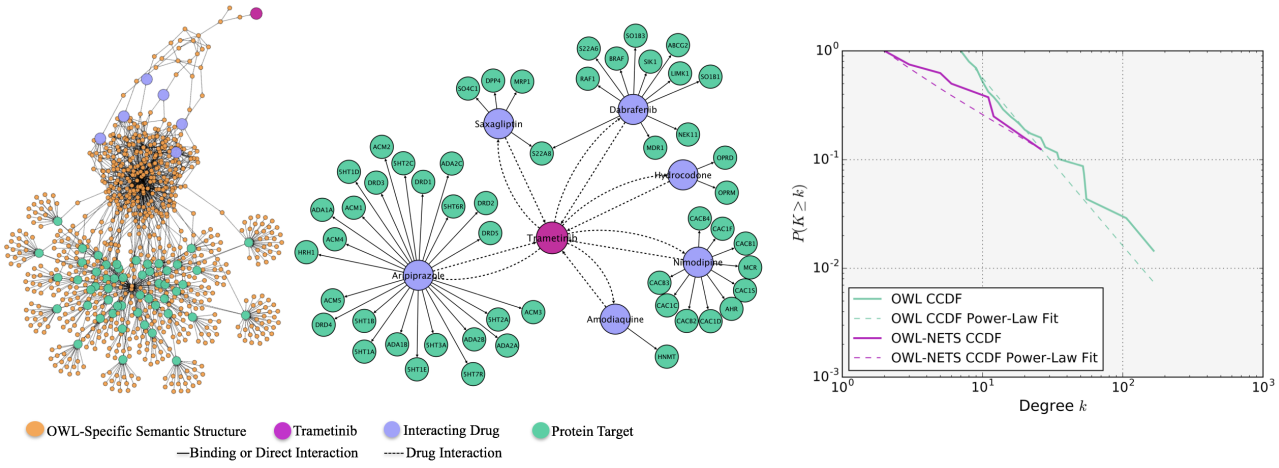


Fig. 2. Query 2 OWL representation network (Left); OWL-NETS abstraction network (middle); and CCDF power-law fit (right). NETS nodes shown in magenta, purple, and teal. Orange 'OWL-Specific Semantic Structure' nodes are needed to create valid OWL expressions and are only part of the OWL representation networks. NETS edge relations shown in solid and dashed lines.

(right), both network representations had good to moderate power-law fit. This is consistent with the fit reported in the literature for other heavy-tailed biological networks.[29,30] Additional network properties can be found in Table S2 of Supplementary Material.

The OWL representation and OWL-NETS abstraction networks built from Query 3 are visualized in Figure 3. The OWL representation network (left) contained 22,679 nodes and 33,848 directed edges. In comparison, the OWL-NETS abstraction network (middle) contained 1,783 nodes and 7,253 directed edges. The OWL-NETS abstraction network had nine connected components (as shown in the figure as one large network surrounded by eight smaller networks), with the largest connected component containing 1,702 nodes and 7,111 directed edges. The largest connected component of the OWL-NETS abstraction network had a larger average degree (4.42 vs. 2.98) and a longer average path length (6.54 vs. 4.13) than the OWL representation. As shown in the third plot (right), both networks had moderate to poor power-law fits. Similar to Query 2, the OWL representation network had a worse fit than the OWL-NETS abstraction network. See Table S2 (Supplemental Material) for additional network properties.

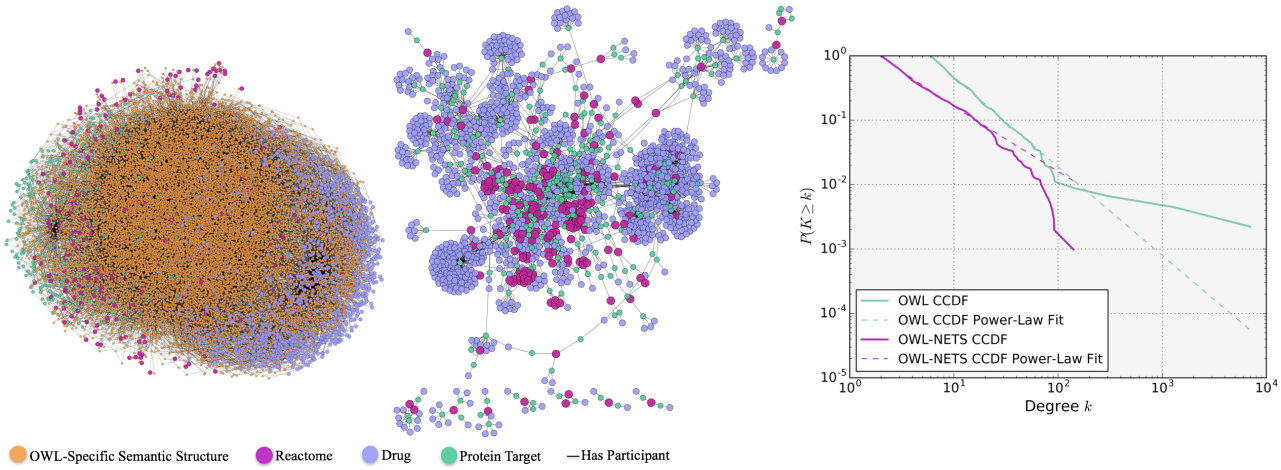


Fig. 3. Query 3 OWL representation network (Left); OWL-NETS Abstraction Network (middle); and CCDF power-law fit (right). Node size is only for visualization. NETS nodes shown in magenta, purple, and teal. Orange 'OWL-Specific Semantic Structure' nodes are needed to create valid OWL expressions and are only part of the OWL representation networks. NETS edge relations shown in solid and dashed lines.

3.2. *Link Prediction Algorithm Performance*

The results from performing link prediction on the network representations for Query 2 are shown in Figure S1 (Supplementary Material).

There were 350,944 nonexistent edges in the OWL representation network and 1,652 nonexistent edges in the OWL-NETS abstraction network. The average AUC scores for the OWL representation network, for all fractions of removed edges, ranged between 0.50 and 0.62. The highest average AUC was found for the Rooted PageRank algorithm when removing a fraction of 0.05 edges from the network (0.66). The average precision for algorithms across all fractions of removed edges was between 0.0001 and 0.001. For the OWL-NETS abstraction

network, the average AUC scores for the best performing algorithms, across all fractions of removed edges, ranked between 0.50 and 0.92. The highest average AUC was found for the Katz algorithm when removing a fraction of 0.05 edges from the network (0.92). The average precision for algorithms across all fractions of removed edges was between 0.001 and 0.04.

The OWL-NETS abstraction network from Query 3 was also explored using the link prediction algorithms (Figure S2, Supplementary Material). There were 257,123,333 nonexistent edges in the OWL representation network and 1,584,713 nonexistent edges in the OWL-NETS abstraction network. The Katz (0.51-0.78), Shortest Path (0.51-0.75), Degree Product (0.58-0.73), and Rooted PageRank (0.64-0.67) algorithms consistently had a higher average AUC across all fractions of removed edges, compared to the other algorithms. Average precision values were similar to Query 2.

The total run time of each algorithm over the 100 iterations varied widely between the network representations from Query 2 (Table S3, Supplementary Material). On average, across all algorithms the OWL-NETS abstraction networks completed 7.5-1185.7 times faster than the OWL representation networks. This is expected given the drastic difference in the number of non-existent edges.

3.2.1. *Inferred Edges*

For Query 2, the Rooted PageRank algorithm performed best on the OWL representation; however, evidence could be found for only one of the predicted edges (HRH1-ADA1A occur in the same calcium signaling KEGG pathway) when run on all edges in the network. Additionally, running HermiT over the OWL representation network resulted in no inferred axioms.

Running the top-performing Katz algorithm on the full OWL-NETS abstraction network produced dramatically better results. From extensive literature review, direct or indirect evidence of a meaningful biological association was found in support of 50% of the top 20 predicted edges (Table 2). Examples of direct evidence found to validate predicted edges include amodiaquine and nimodipine, which have been shown to experimentally to regulate the expression of voltage-dependent calcium channel L-type alpha 1 C subunits[31] and HNMT,[32] respectively. Indirect biological evidence was found to support the edge between amodiaquine and AHR, which was substantiated by their shared relationship to the CYP1a1 enzyme[33,34] and the amodiaquine-OPRM1 edge, which is supported by the reported relationship between pruritus (itching) induced by quinolones (a family of anti-malarial drugs including amodiaquine) and its pharmacologic treatment with naltrexone, an OPRM1 antagonist.[35] Interestingly, many of the top 20 edges were related to each other by common biological themes such as histamines and the opioid receptor system. Additionally, as demonstrated by the results in Table 2, DPP4 appears to link seemingly unrelated groups of drugs including narcotics (e.g., hydrocodone), those used in the treatment of malaria (e.g., amodiaquine), and diabetes (e.e., saxagliptin). Therefore, it is possible that the examination of predicted edges can reveal biological mechanisms underlying the interactions of drugs with other drugs and targets.

Evidence from the literature could be found to support 75% of the top 20 scoring edges predicted by the Katz algorithm on the OWL-NETS abstraction network generated from Query 3 (selected examples of which are presented in Table S4, Supplementary Material). The

Table 1. Top scoring edges from Query 2 OWL-NETS abstraction network (n=10 edges)

Node 1	Node 2	Description
amodiaquine[a]	DPP4[a]	Middle East Respiratory Syndrome-Coronavirus (MERS-CoV) gains entry into cells via DPP4 and amodiaquine has activity against MERS-CoV.[36]
amodiaquine[a]	CACNA1C[b]	Amodiaquine-treated mice have decreased expression of voltage-dependent calcium channel L-type alpha 1 C subunits in their livers.[31]
amodiaquine[a]	CACNA1D[b]	Amodiaquine-treated mice have decreased expression of voltage-dependent calcium channel L-type alpha 1 C subunits in their livers.[31]
nimodipine[a]	HNMT[b]	A small molecule screen demonstrated that nimodipine caused increased HNMT expression in cultured human cells.[32]
amodiaquine[a]	OPRM1[b]	Quinolone-based antimalarials can induce generalized pruritus (itch), which can be treated with the mu-opioid receptor (OPRM1) antagonist naltrexone.[35]
amodiaquine[a]	AHR[b]	Amodiaquine is metabolized by CYP1a1. CYP1a1 is induced by signaling through AhR.[33,34]
hydrocodone[a]	HNMT[b]	HNMT regulates histamine release and opiates, like hydrocodone, induce histamine release.[37]
hydrocodone[a]	DPP4[b]	DPP4 cleaves dietary gliadin into opioid peptides that can activate mu-opioid receptors. Hydrocodone activates mu-opioid receptors.[38,39]
hydrocodone[a]	MRP1[b]	MRP1 is involved in maintaining the blood-brain barrier. Down-regulation of MRP1 increases the analgesic effect of systemic morphine in mice and rats by decreasing the blood-brain barrier. Hydrocodone is a synthetic opioid.[40]
saxagliptin[a]	OPRM1[b]	Saxagliptin inhibits DPP4, which contributes to the cleavage of dietary gliadin into opioid peptides. The gliadin opioid peptide, gliadinomorphin-7, can activate the mu-opioid receptor (OPRM1).[38,39]

[a]DrugBank entity (DrugBank ID used for experimental compounds); [b]Uniprot entity (gene symbol is shown to preserve space).

predicted edge between AG-1067, a derivative of probucol (an anti-hyperlipidemic drug), and MMP2 is supported by experimental evidence that probucol decreases the expression and activity of MMP2 in a mouse model.[41] Similarly, the predicted edge between DB04513 and RAF1 is substantiated by evidence that calmodulin 1, which is the target of experimental drug N-(6-aminohexyl)-5-chloro-1-naphthalenesulfonamide (DB04513), modulates signaling through the

Ras/Raf/MEK/ERK signaling pathway.[42] Two additional edges between celiprolol and CYCS and between Reactome pathway 1454838 and transferrin were found to be biologically related through disease processes including hypertension,[43,44] and multiple myeloma.[45,46] The predicted edge between experimental drug 2-[formyl(hydroxy)amino]methyl-4-methylpentanoic acid (DB03683) and APAF1 was supported by their shared association with MMP9.[47]

4. Discussion

Networks representing biomedical mechanisms constantly evolve; the addition of new edges within a network may symbolize important interactions and provide valuable insight into its underlying biology.[48] Investigating new edges within these networks provides a methodology for generating novel hypotheses. While OWL provides powerful techniques for representing existing knowledge, well-established OWL reasoners are unable to account for missing or uncertain knowledge. Further, the structural complexity of OWL reduces the effectiveness of certain types of network inference. To address these limitations we developed OWL-NETS, a novel computational method that reversibly abstracts OWL-encoded biomedical knowledge into a network representation tailored for network inference. To the best of our knowledge, there are no existing network abstraction methods designed to create network representations from OWL-encoded knowledge sources to facilitate network inference.

Existing network abstraction methods reduce the structural complexity of a terminology/ontology by aggregating nodes with similar attributes or properties.[21–23,49] An abstraction network is considered useful if it is significantly smaller than the original terminology/ontology without losing structure and content.[20] The goal of OWL-NETS is to collapse the nodes and edges that are necessary to logically represent relationships between biological entities in OWL, but are not themselves biologically meaningful and interfere with network inference. In contrast to existing methods, the reduced size of the OWL-NETS abstraction network, relative to its original OWL representation, is not predictive of its usefulness for inference. In fact, too much aggregation may result in a network whose properties are no better for inference than the original OWL representation. More importantly, existing network abstraction methods were not designed for network inference; combining nodes having the same attributes or properties could inadvertently mask important biological relations of the resulting abstraction networks.

This work is not without limitations. Relying on literature review, even if by a domain expert, does not provide the most robust evaluation of inferred edges. Future work will include a collaboration where results can be evaluated experimentally. Additionally, the current work evaluated relatively simple, unipartite networks. Future work will explore more complex types of biological network representations, such as bipartite and multiplex networks. We are also developing methods for adding edge weight to the OWL-NETS abstraction networks to indicate the amount and/or quality of the confidence/evidence of the connection between the biological entities. Finally, exploration of more complex link prediction algorithms that can accommodate directed networks as well as alternative methods for predicting missing edges (e.g., community detection methods[50]) is also a focus of future work.

5. Conclusions

OWL-NETS is a novel abstraction network methodology that generates semantically rich network representations that are easily consumed by network inference algorithms. OWL-NETS is easy to configure and can be modified for use with other knowledge sources leveraging Semantic Web technologies. When running link prediction algorithms over OWL-NETS we provided expert-verified evidence from the literature for 50-75% of inferred edges. By leveraging many knowledge sources in a representation tailored for network inference, OWL-NETS has a unique ability to recognize existing, natural patterns in the biological world that have not yet been identified, which would be readily testable in the laboratory environment.

6. Acknowledgments

We thank Marc Daya and Laura Stevens as well as Drs. Anis Karimpour-Fard, Daniel McShan, and Carsten Goerg for their feedback on the development of OWL-NETS. We also thank Ann Cirincione and Raja Cholan for their review of the manuscript.

7. Funding

This work was supported by the National Library of Medicine Training Grant T15 LM009451, as well as R01 LM009254 and R01LM008111 to LH.

References

1. E. Ravasz, A. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabsi, *Science* **297**, 1551 (2002).
2. J. Bultinck, S. Lievens and J. Tavernier, *Curr Pharm Des* **18**, 4619 (2012).
3. M. Lee, K. Park and D. Kim, *BMC Syst Biol* **7 Suppl 3**, p. S4 (2013).
4. R. Albert, *Plant Cell* **19**, 3327 (2007).
5. O. W. Group, Web ontology language (owl) (2012), `https://www.w3.org/OWL/`.
6. K. Dentler, R. Cornet, A. Teije and N. de Keizer, *Sem Web* **1**, 1 (2011).
7. Y. Kazakov, M. Krotzsch and F. Simancik, *J Autom Reasoning* **53**, 1 (2014).
8. D. Tsarkov and I. Horrocks, Fact++ description logic reasoner: System description, in *Proceedings IJCAR-2006*, 2006.
9. R. Carvalho, K. Laskey and P. D. Costa, *Peer J Comp Sci* **2**, p. 77 (2016).
10. P. C. G. D. Costa, K. B. Laskey and K. J. Laskey, *A Bayesian ontology language for the semantic web* (Springer, Berlin, Heidlberg).
11. D. Koller, A. Levy and A. Pfeffer, P-classic: a tractable probablistic description logic, in *Proceedings AAAI-97*, 1997.
12. A. Rettinger, U. Lsch, V. Tresp, C. d'Amato and N. Fanizzi, *Data Min Knowl Disc* **24**, 613 (2012).
13. H. Mohammadhassanzadeh, W. V. Woensel, S. R. Abidi and S. S. Abidi, *BioData Mining* **10**, p. 7 (2017).
14. F. Baader, D. Calvanese, D. McGuiness, D. Nardi and P. Patel-Schneider (eds.), *The Description Logic Handbook: Theory, Implementation and Applications* (Cambridge University Press, 2003).
15. H. Wang, H. Huang, C. Ding and F. Nie, *Soft Computing Ontologies Sem Web* **20**, 344 (2013).
16. A. L. Hopkins, *Nat Chem Biol* **4**, 682 (2008).
17. J. Watkinson, K. C. Liang, X. Wang, T. Zheng and D. Anastassiou, *Ann N Y Acad Sci* **1158**, 302 (2009).
18. V. Haarslev and R. Moller, An owl reasoning agent for the semantic web, in *Proceedings of the*

International Workshop on Applications, Products and Services of Web-based Support Systems, in conjunction with 2003 IEEE/WIC International Conference on Web Intelligence,, 2003.
19. M. Halper, H. Gu, Y. Perl and C. Ochs, *Artif Intell Med* **64**, 1 (2015).
20. C. Ochs, Z. He, L. Zheng, J. Geller, Y. Perl, G. Hripcsak and M. A. Musen, *J Med Bioinfom* **61**, 63 (2016).
21. H. Gu, Y. Perl, J. Geller, M. Halper, L. M. Liu and J. J. Cimino, *J Am Med Inform Assoc* **7**, 66 (2000).
22. Y. Wang, M. Halper, H. Min, Y. Perl, Y. Chen and K. A. Spackman, *J Biomed Inform* **40**, 561 (2007).
23. Y. Wang, M. Halper, D. Wei, Y. Perl and J. Geller, *J Biomed Inform* **45**, 15 (2012).
24. K. M. Livingston, M. Bada, W. A. Baumgartner and L. E. Hunter, *BMC Bioinformatics* **16**, p. 126 (2015).
25. A. Clauset, C. Moore and M. E. J. Newman, *Nature* **453**, 98 (2008).
26. L. Lu and T. Zhou, *Physica A Journal* **390**, 1150 (2011).
27. J. A. Hanley and B. J. McNeil, *Radiology* **143**, 29 (1982).
28. B. Glimm, I. Horrocks, B. Motik, G. Stoilos and Z. Wang, *J Autom Reasoning* **53**, 245 (2014).
29. E. K. Towlson, P. E. Vertes, S. E. Ahnert, W. R. Schafer and E. T. Bullmore, *J Neurosci* **33**, 6380 (2013).
30. L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall and D. B. Chklovskii, *PLoS Comput Biol* **7**, p. e1001066 (2011).
31. S. K. Mishra, P. Singh and S. K. Rath, *Malar J* **10**, p. 109 (2011).
32. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *Science* **313**, 1929 (2006).
33. J. P. Gil, *Pharmacogenomics* **9**, 1385 (2008).
34. T. Johansson, U. Jurva, G. Grnberg, L. Weidolf and C. Masimirembwa, *Drug Metab Dispos* **37**, 571 (2009).
35. A. A. Ajayi, B. A. Kolawole and S. J. Udoh, *Int J Dermatol* **43**, 972 (2004).
36. T. Pillaiyar, M. Manickam and S. H. Jung, *Med Chem* **5**, 361 (2015).
37. B. A. Baldo and N. H. Pham, *Anaesth Intensive Care* **40**, 216 (2012).
38. L. Pruimboom and K. de Punder, *J Health Popul Nutr* **33**, p. 24 (2015).
39. M. S. Trivedi, J. S. Shah, S. Al-Mughairy, N. W. Hodgson, B. Simms, G. A. Trooskens, W. Van Criekinge and R. C. Deth, *J Nutr Biochem* **25**, 1011 (2014).
40. W. Su and G. W. Pasternak, *Synapse* **67**, 609 (2013).
41. B. J. Wu, N. D. Girolamo, K. Beck, C. G. Hanratty, K. Choy, J. Y. Hou, M. R. Ward and R. Stocker, *J Pharmacol Exp Ther* **321**, 477 (2007).
42. N. Agell, O. Bachs, N. Rocamora and P. Villalonga, *Cell Signal* **14**, 649 (2002).
43. C. P. Venditti, M. C. Harris, D. Huff, I. Peterside, D. Munson, H. S. Weber, J. Rome, E. M. Kaye, S. Shanske and S. Sacconi, *J Inherit Metab Dis* **27**, 735 (2004).
44. W. Z. Ying and P. W. Sanders, *Kidney Int* **59**, 662 (2001).
45. B. K. Arendt, D. K. Walters, X. Wu, R. C. Tschumper, P. M. Huddleston, K. J. Henderson, A. Dispenzieri and D. F. Jelinek, *Leukemia* **26**, 2286 (2012).
46. K. VanderWall, T. R. Daniels-Wells, M. Penichet and A. Lichtenstein, *J Inherit Metab Dis* **18**, 449 (2013).
47. C. S. Gondi, N. Kandhukuri, D. H. Dinh, W. C. Olivero, M. Gujrati and J. S. Rao, *Int J Oncol* **33**, 783 (2008).
48. D. Liben-Nowell and J. Kleinberg, *J. Am. Soc. Inf. Sci.* **58**, 1019 (2007).
49. H. Gu, J. J. Cimino, M. Halper, J. Geller and Y. Perl, *Proc AMIA Annu Fall Symp* **1996**, 275 (1996).
50. D. Hric, T. P. Peixoto and S. Fortunato, *Phys Rev X* **6**, 031038 (2016).

Automated disease cohort selection using word embeddings from Electronic Health Records

Benjamin S. Glicksberg[1,2*], Riccardo Miotto[1,2*], Kipp W. Johnson[1,2], Khader Shameer[1,2], Li Li[1,2], Rong Chen[1], Joel T. Dudley[1,2]

Department of Genetics and Genomic Sciences,[1] Institute for Next Generation Healthcare[2]
Icahn School of Medicine at Mount Sinai
Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.
New York, NY 10065, USA

** Authors contributed equally*
Corresponding author: joel.dudley@mssm.edu

Accurate and robust cohort definition is critical to biomedical discovery using Electronic Health Records (EHR). Similar to prospective study designs, high quality EHR-based research requires rigorous selection criteria to designate case/control status particular to each disease. Electronic phenotyping algorithms, which are manually built and validated per disease, have been successful in filling this need. However, these approaches are time-consuming, leading to only a relatively small amount of algorithms for diseases developed. Methodologies that automatically learn features from EHRs have been used for cohort selection as well. To date, however, there has been no systematic analysis of how these methods perform against current gold standards. Accordingly, this paper compares the performance of a state-of-the-art automated feature learning method to extracting research-grade cohorts for five diseases against their established electronic phenotyping algorithms. In particular, we use *word2vec* to create unsupervised embeddings of the phenotype space within an EHR system. Using medical concepts as a query, we then rank patients by their proximity in the embedding space and automatically extract putative disease cohorts via a distance threshold. Experimental evaluation shows promising results with average F-score of 0.57 and AUC-ROC of 0.98. However, we noticed that results varied considerably between diseases, thus necessitating further investigation and/or phenotype-specific refinement of the approach before being readily deployed across all diseases.

Keywords: Electronic Health Records, Automated cohort selection, Electronic phenotyping algorithms, Vector-based representations, Word embedding, Feature learning

1. Introduction

Clinical data collected from patient hospital visits are archived in electronic health records (EHR) as part of the healthcare process. These data consist of disease diagnoses, medication prescriptions, procedures performed, among others. EHR-based research enables countless opportunities for biomedical research [1-3] and precision medicine [4]. However, one of the cornerstones of EHR research is the requirement to reliably detect patients with a particular disease or phenotype for use in observational cohort studies. Accurately identifying patients with a

disease of interest in an EHR system, however, is not trivial due to input errors, coding biases, medical reporting biases, data availability, sparsity, and limitations of how the data is structured. Defining case/control disease cohorts through presence of a single clinical concept, such as an International Statistical Classification of Diseases and Related Health Problems (ICD) code, is often not sufficient to produce reliable distinctions. Furthermore, these concepts vary greatly in their performance for identifying different diseases [5]. For example, relevant medication prescriptions may phenotype patients in some disease with high precision, but not help classification in another.

Advanced EHR phenotyping of diseases is best understood as an "expert system," where researchers with advanced knowledge of a particular disease phenotype design a list of criteria which may be used to identify affected (i.e. cases) and sometimes assuredly non-affected (i.e. controls) individuals by excluding those with ambiguous clinical status or lack of enough data [6]. This typically takes the form of intricate rule-based algorithms specifying the presence and/or absence of particular billing codes, pre-defined ranges for laboratory tests, the prescription of characteristic medications, processing of clinical notes, among others. These types of rule-based algorithms, called "electronic phenotyping algorithms" when working with EHR data, perform markedly better than simpler alternatives [7]. The Electronic Medical Records and Genomics (eMERGE) [8] consortium has led the effort in defining, implementing, and validating such algorithms for a number of diseases. The Phenotype KnowledgeBase (PheKB) [9] repository contains such algorithms from eMERGE as well as from other sources. While this approach is effective, there are also two main drawbacks. First, implementing each algorithm in a new dataset (for example, a researcher wanting to use a previously published algorithm on a different set of EHR records) is a time-intensive and sometimes demanding task. This may require dealing with a variety of data formats, getting access to specific laboratory or imaging test results, implementing natural language processing pipelines, and so on. Second, biomedicine deals with an enormous amount and variety of disease – establishing criteria for each new disease is an onerous process and does not scale well. In fact, a systematic analysis of commonalities in selection criteria for 24 eMERGE algorithms found that each algorithm had variable amounts and types of design patterns [10]. Due to these restrictions, relatively few algorithms have been created. Currently, there are only 42 public phenotypes in PheKB, which altogether represent only a small fraction of human disease. As such, methodologies to expedite the design and implementation of phenotyping algorithms, or better yet avoid the process overall, would be tremendously beneficial and could lead to better/more research of this kind.

There are many well-established methodologies to automatically learn representations of data [11]. For instance, learning low-dimensional representations, such as word embeddings, is a common practice to transform high-dimensional data [12]. The use of word embeddings has been

proven to be particularly effective in NLP-related tasks, such as language modeling and information extraction [13]. Several models have been proposed for learning distribution representations of words, the most popular of which being the *skip-gram* model implemented in the *word2vec* framework [14]. The success of neural networks for computing word embeddings has motivated adaptation of these algorithms for other types of data, such as clinical data. In fact, there have been several studies that automatically learned embeddings in biomedical informatics. Choi et al. [15] learned and compared low-dimensional representations of medical concepts from medical journals (abstracted UMLS concepts from around 350,000 medical paper abstracts), medical claims (structured clinical data from an insurance company), and clinical narratives (notes from publicly available EHR data). They found that the embeddings from these different sources produced high-quality results but differed significantly based on data modality. Of particular relevance to the current study, Halpern and Horng et al. [16] used the anchor-and-learn framework to build phenotype libraries of features using emergency department EHR data. Specifically, this method identifies features as anchors for diseases that have both high positive predictive value as well as conditional independence from any other feature that could improve prediction. They built 42 phenotype definitions using this method, evaluating eight of them using physician responses to gauge performance. Rotmensch et al. [17] used three probabilistic models to automatically extract concepts and create a health knowledge graph of disease-symptom relationships from EHR data of approximately 275,000 patients. The resulting learned graphs compared encouragingly against a physician-constructed knowledge set.

Many other efforts to automatically extract phenotypes from EHRs have performed well. Miotto et al. [18] built *Deep Patient*, a three-layered stack of denoising autoencoders used to predict risk for future health states (i.e. disease risk) within EHR. Kandula et al. [19] developed an algorithm built through bootstrapping that iteratively selects data types to incorporate and tested their method for diabetes mellitus and hyperlipidemia cohort identification. Yu et al. [20] used NLP of clinical notes and machine learning to identify rheumatoid arthritis and coronary artery disease patients. Pivovarov et al. [21] developed *UPhenome*, an unsupervised, probabilistic graphical model to learn computational models of diseases. Chiu and Hripcsak [22] developed a three-tier, stacked architecture for ensemble learning and feature representations to define disease cohorts. While innovative and successful for their goals, these works have yet to benchmark their models against gold standard phenotyping algorithms to appropriately assess their utility for these types of studies. In fact, performances of these models were mainly evaluated using manual expert review of charts for a small subset of patients or from typical machine learning approaches (i.e. training and test). As such, there is still a gap to evaluate how these automated methods fare against established algorithms. Agarwal et al. [23] learned phenotype models of Type 2 Diabetes and Myocardial Infarction using a semi-automatic ("silver standard") procedure through the

manual selection of relevant terms. Their models compared favorably against electronic phenotyping algorithms, demonstrating the feasibility of these methods for research purposes. As they mention, however, this methodology was measured against only two diseases and not fully automatic.

To the best of our knowledge, there is no systematic evaluation of how fully automated disease-cohort characterization from EHR compares against research-grade, rule-based methodologies. In the current study, we build upon these previous works and compare performance of automatically derived patient cohorts using word embeddings for five diseases against established PheKB electronic phenotyping algorithms. In particular, we use the EHRs of over a million patients to learn embeddings of the medical concepts in the structured records, and use these embeddings to summarize the clinical history of the patients. For each disease of interest, we then select only a single meaningful clinical concept to use as query, and we rank patients based on the distance from the corresponding embedding of the query. From here, putative cohorts are systematically generated for the disease concept based on patient embeddings with highest similarity, and compared against the gold standards. From these comparisons, we have the unique opportunity to learn strengths and limitations of this embedding methodology on EHR data by comparing performance across diseases of different types that contain various data modalities and rule-based algorithms. Ultimately we hope to generate automated disease representations suitable for research studies.

2. Methods and Materials

We present an overall workflow of the study and methodologies in Figure 1.

2.1. *Research Cohort and Resource*

We utilized clinical data from the EHRs of the Mount Sinai Hospital (MSH). MSH is an urban, tertiary care hospital located in on the Upper East Side of Manhattan in New York City. Clinical data within the EHRs includes disease diagnoses, lab test results, vital signs, medication prescriptions, and procedures among others. For the current study, we restricted our research cohort to individuals with at least one recorded clinical feature, leaving 1,304,192 unique patients for subsequent analyses. Due to HIPAA requirements, the ages of patients within the research cohort are right censored at age 90. The mean age of the cohort is 45.24 ± 22.71 (std). The self-reported sex breakdown of the cohort is 56.7% female, 43.3% male, and 0.02% not available. The self-reported race breakdown of the cohort is: 36.9% Caucasian (White), 14.2% African American (Black), 8.88% Hispanic/Latino, 3.99% Asian, 3.1% Other, and 35.7% not available.

2.2. *Disease Phenotyping Algorithms*

For gold standard disease cohort selection, we utilized electronic phenotyping algorithms from PheKB. We selected diseases by first restricting algorithms to those that are public and are of the type "Disease or Syndrome" and then to disease only (e.g. dementia, but not peanut allergy). We did not consider algorithms that require Natural Language Processing (NLP) of clinical notes as part of the selection criteria. After filtering, we selected five of the remaining seven algorithms: Attention Deficit Hyperactivity Disorder (ADHD) [24], Dementia [25], Herpes Zoster [26], Sickle Cell disease (Sickle Cell) [27, 28], and Type 2 Diabetes (T2D) [29]. While some of these algorithms include control inclusion criteria, we attempted only case selection.

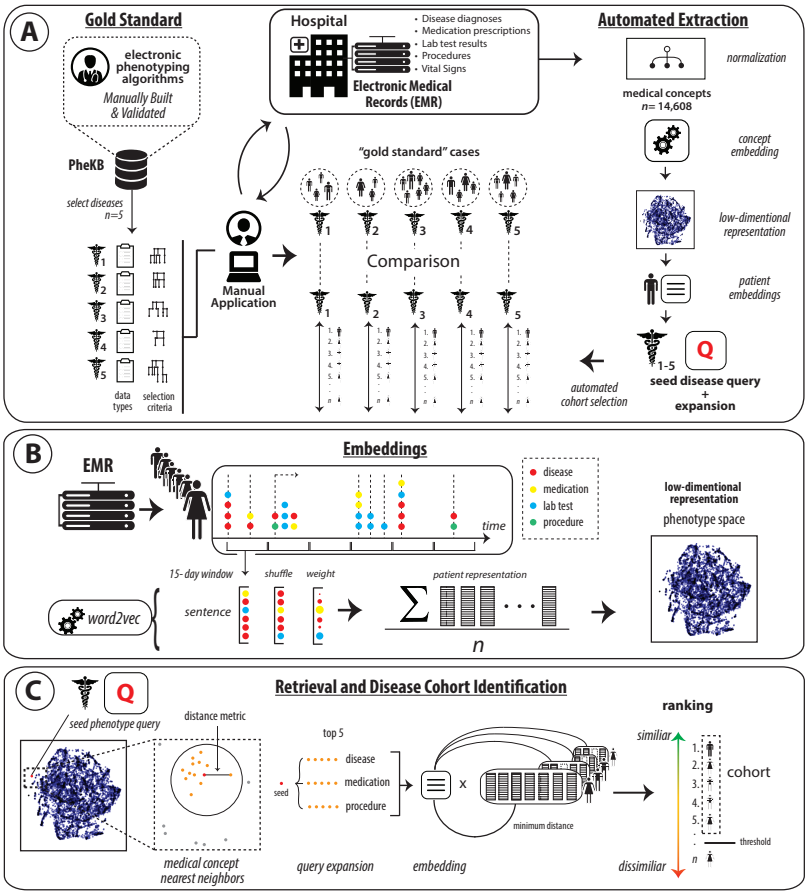


Figure 1: Workflow design of the current study. A) Framework for comparing gold standard PheKB algorithms to our automated method. B) Embeddings procedure. C) Retrieval and disease cohort identification

2.2.1. *Electronic Phenotype Algorithm Implementation*

For all algorithms, the data types included are ICD-9 for disease diagnoses; Current Procedural Terminology (CPT) and CPT-Healthcare Common Procedure Coding System (HCPCS) codes for procedures; Logical Observation Identifiers Names and Codes (LOINC) codes and descriptions for lab tests and vital signs. Unless explicitly specified otherwise, we used wildcard characters at the end of all non-five digit ICD-9 codes (e.g. 314.xx). Medications terms can include dosage and route of administration in addition to the drug name (e.g. CLONAZEPAM 0.5 MG TAB), and as such, we obtained records by querying each term surrounded with wildcard characters (e.g. "%Melipramine%"). We were able to successfully implement all algorithms with only a few minor modifications as necessary, which we describe in this section. As we do not perform association testing using the disease cohorts in the current study, we did not implement covariate-related procedures, specifically antiviral medication

minimum dosage threshold, in our application of the Herpes Zoster algorithm. In the T2D algorithm, for the glucose lab test results, we were not able to distinguish "fasting" from "non-fasting" measurements, and as such, considered all records as "non-fasting". We retrieved diabetes medical supplies information using the same search approach as for medication data, necessitated by the fact that the authors of the algorithm utilize RxNorm codes, which we did not have mappings for. For "Blood-glucose meters and sensors", we queried: 1. "glucometer" and 2. "%glucose%" [AND] ("%meter% [OR] "%monitor%" [OR] "%sensor%"), producing 6,104 records and 47 distinct terms. For "Insulin syringes", we queried "%insulin%" [AND] ("%syringe%" [OR] "%inject%" [OR] "%pen%" [OR] "%innolet%" [OR] "%flextouch%" [OR] "%solostar%" [OR] "%cart%"), resulting in 117,469 records and 356 unique items.

2.3. *Phenotype and Patient Embedding*

We learn a set of low-dimensional representations (i.e., "embeddings") of medical concepts from the structured EHR. These representations put all ICD-9 diagnosis and procedure codes, laboratory codes, and drug codes in a common metric space where similarity is inversely proportional to pairwise distance. Next, we use these embeddings to summarize the patient history by weighted average of the medical concepts over time windows. For each disease of interest, we then use a query consisting of a representative medical concept (e.g. ICD-9 code) as a seed and then expand it to other related concepts. Lastly, we use these representations to identify patients with each disease by measuring the distance of each patient from the query.

2.3.1. *Data pre-processing*

In order to systematically create embeddings, the various data types within the EHR have to be pre-processed. In particular, we normalized all ICD-9 codes to four digits resulting in 6,272 terms. We normalized medication data using Open Biomedical Annotator [30] yielding 4,022 terms. We normalized vital signs and encounter descriptions (e.g. "Outpatient") into seven and 10 terms respectively. Procedures and lab tests were normalized based upon sub-string prefixes and similarity, which generated 2,414 and 1,883 terms respectively. In total, we derived 14,608 distinct clinical concepts to be used in embedding procedures.

2.3.2. *Learning Embeddings of Medical Concepts*

We take inspiration from Choi et al. [15] and use the skip-gram algorithms to learn embeddings of the medical concepts reported in the EHRs. For each patient, we organize the normalized clinical concepts into an irregularly-sampled temporal sequence, where concepts adjacent to each other in the sequence should cluster together in the learned metric space. To this end, we first partitioned the patient data in consecutive time intervals composed by fifteen days (Figure 1B). Second, we

removed duplicates from each time interval and third, we random-shuffled the concepts in each interval. Each time interval represented as a sequence of unique medical concepts was then considered as a "sentence" to be given to the word2vec algorithm, which was trained using stochastic gradient descent and used as dynamic context the number of concepts in each sentence. At the end, every medical concept was represented as a 200-dimensional embedded vector, with all the medical concepts mapped in the same metric space. Figure 2 shows a visualization of the embeddings learned from the medical concepts in the EMR, going from the raw low-dimensional data (A) to seeding with the ADHD concept (B) and clustering using t-SNE (C, D).

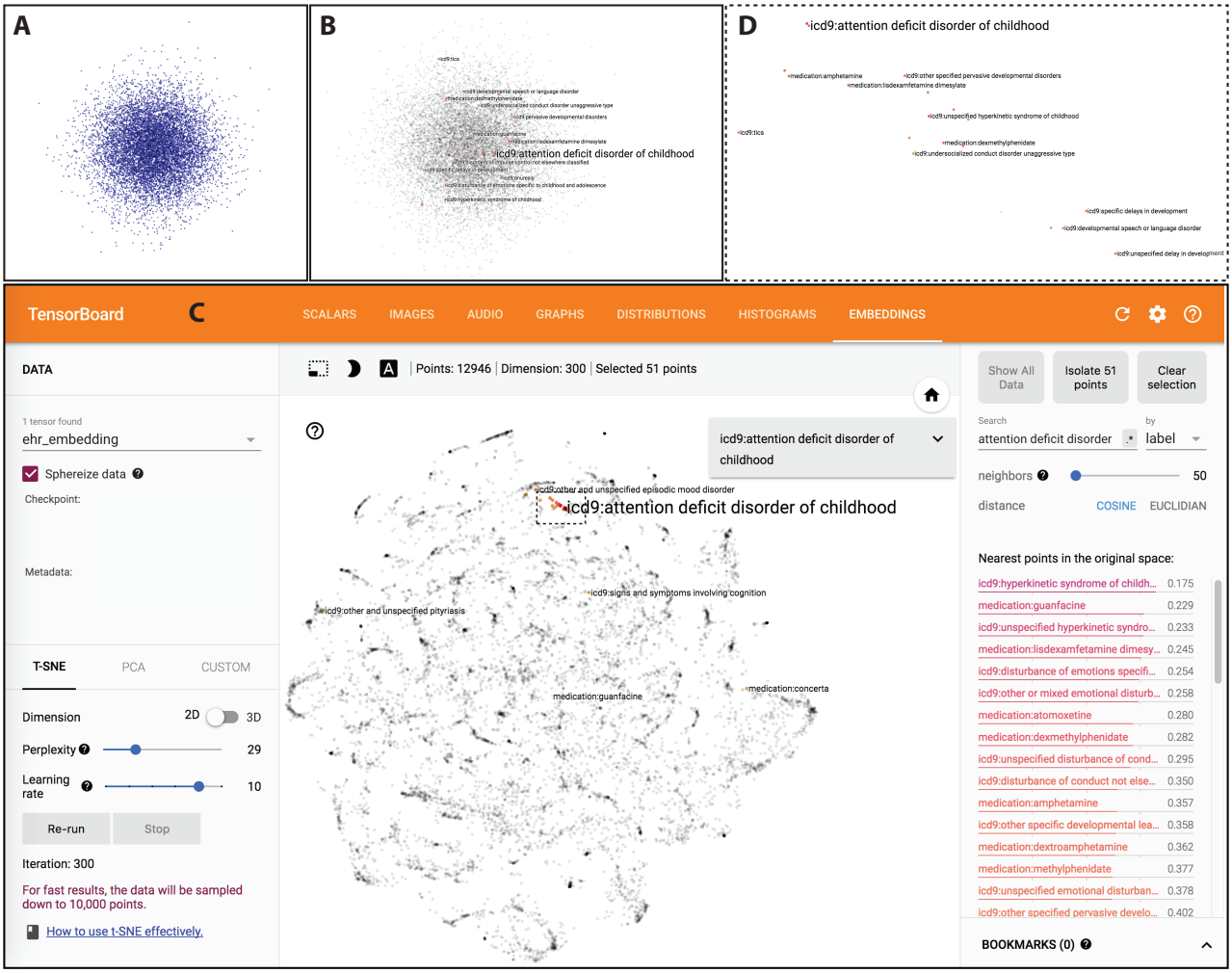


Figure 2: Disease embedding for ADHD with the top 50 closest neighboring concepts from A) the raw low-dimensional space, B) to query C), to clustering (D is zoomed in section).

2.3.3. *Deriving Patient Representations from Medical Concept Embeddings*

For every time interval considered in the patient clinical history, we used the simple sentence aggregation method proposed by Arora et al. [31]. In particular, we computed the weighted average of the medical concept embeddings and subtracted the projections of the average vectors on their first principal component. This facilitates the removal of the largely shared components from the vectors, leading to more discriminative aggregated embeddings. The weight of a phenotype w was computed as: $w = a / (a + p(w))$ with a being a parameter and $p(w)$ being the (estimated) phenotype frequency across the whole dataset. At the end of this process, every patient was characterized by a bag of clinical status embeddings, lying in the same space of the medical concepts, which are used for performing the automated phenotyping.

2.3.4. *Automatic Disease Phenotyping from the Embeddings*

For the diseases of interest, we used the following representative concepts as seed queries: ICD-9 code 314.0x for ADHD; ICD-9 code 290.xx for Dementia; ICD-9 code 053.xx for Herpes Zoster; ICD-9 code 282.6x for Sickle Cell; and ICD-9 code 250.xx for T2D. For each disease query, we sought to capture related concepts through query expansion (Figure 1C). Here, we added the top five closest ICD-9 codes, medications, and procedures to the original seed query. We used cosine distance to measure the relationship between each patient and query vectors, using the closest patient vector (i.e. sentence) as a summarized score. We then repeat this process for each vector in the expanded query pool and retained the average of all the distances as a final value. For sake of comparison, we derived disease cohorts using ICD-9 presence only, embeddings only, and embeddings with the query expansion.

2.4. *Evaluation Design*

For every disease considered, we evaluate the embeddings for annotation and retrieval and report the precision, recall, and F-score. In the annotation task, we assigned a positive label to each patient for the disease if the distance from the query was below a certain threshold. To facilitate the definition of the threshold, we mapped the distances to probabilities (ranging from 0 to 1). Precision is the number of correct positive results divided by the number of all positive results, recall is the number of correct positive results divided by the number of all true positive results, and F-score is the harmonic mean of them both. We set the threshold to 0.7, with this value optimizing the tradeoff between precision and recall for all diseases examined. In the retrieval task we sorted the patients by their distance from the query and evaluated the ranking lists obtained. As metrics, we report Precision-at-10 (Prec@10) and R-precision (Prec@R). Specfically, Prec@10 measures the ratio of relevant patients (i.e., patients with the disease in the ground truth) within the top 10 positions of the ranked embedding output list (i.e., top 10 closest patients to the query) for

each disease. Prec@R is the precision-at-*R* of the query disease, where *R* is the number of patients with that disease in the ground truth.

3. Results

3.1. *Evaluating Performance of Embeddings*

We ran the electronic phenotyping algorithms mentioned above to obtain gold standard patient cohorts for each disease of interest. The patient count for each cohort is as follows: 7,487 individuals for ADHD, 10,782 for Dementia, 1,618 for Herpes Zoster, 943 for Sickle Cell, and 56,687 for T2D.

3.1.1. *Phenotype Embedding Methodologies*

The first goal of evaluating phenotype embeddings was to assess overall performance across three different models: ICD-9 only, (Phenotype) Embedding Only, and (Phenotype) Embedding with Query Expansion. We present the evaluation metrics of each model in Table 1. For "Annotation", the ICD-9 Only method interestingly achieved highest precision (0.609) but lowest Recall and F-Score, implications of which we address in the Discussion. The Embedding with Query Expansion performed best in terms of Recall (0.795) and, more importantly, F-Score (0.569), which combines the other metrics. The Embedding Only method outperformed the ICD-9 Only method in the same metrics, but to a lower degree. The Query Expansion improved upon Embeddings in all metrics but most notably in Recall (0.795 vs. 0.489). For "Retrieval", Embedding with Query Expansion outperformed Embedding Only and ICD-9 Only in all metrics, enhancing our confidence in using this method.

Table 1: Annotation and retrieval performance for each method. All metrics are upper bounded by 1.

	Annotation			**Retrieval**		
Algorithms	*Precision*	*Recall*	*F-Score*	*Prec@10*	*Prec@R*	*AUC-ROC*
ICD-9 Only	**0.61**	0.18	0.27	0.60	0.19	0.59
Embedding Only	0.44	0.49	0.44	0.62	0.48	0.96
Embedding with Query Expansion	0.50	**0.80**	**0.57**	**0.66**	**0.56**	**0.98**

3.1.2. *Phenotype Embedding with Query Expansion at the Disease Level*

We present the evaluation metrics using the Phenotype Embedding with Query Expansion method for all five diseases of interest in Table 2. For "Annotation", it is clear that this embedding procedure exhibits variable performance depending on disease. ADHD, Sickle Cell, and T2D performed relatively well with F-scores of 0.74, 0.72, and 0.67. The Dementia query performed

the poorest with an F-score of 0.28, primarily due to low Recall (0.20). These trends mostly carried over for "Retrieval" assessment.

Table 2: Annotation and retrieval performance for each disease using the Embedding with Query Expansion Method. All metrics are upper-bounded by 1.

	Annotation			**Retrieval**		
Diseases	*Precision*	*Recall*	*F-Score*	*Prec@10*	*Prec@R*	*AUC-ROC*
ADHD	0.59	0.98	0.74	1.00	0.51	0.96
Dementia	0.53	0.20	0.28	1.00	0.53	0.96
Herpes Zoster	0.30	0.93	0.45	0.20	0.36	0.99
Sickle Cell	0.60	0.91	0.72	1.00	0.73	1.00
Type 2 Diabetes	0.50	0.97	0.67	0.30	0.54	0.97

To illustrate the utility of the query expansion, we present the concepts adopted for the embedding of Herpes Zoster in Table 3.

Table 3: Features incorporated in the Herpes Zoster model in the expanded query for each modality.

Modality	**Feature**	**Similarity**
ICD-9	Herpes zoster w/out complication (053.9)	0.788
	Herpes zoster w/ other nervous system complications (053.19)	0.724
	Herpes zoster w/ ophthalmic complications (053.29)	0.599
	Herpes zoster w/ other specified complications (053.7)	0.568
	Genital herpes (054.1)	0.538
Medication	Valacyclovir	0.607
	Famciclovir	0.587
	Valacyclovir hydrochloride	0.522
	Capsaicin	0.515
	Valtrex	0.506
Procedure	Varicella zoster	0.503
	Hiv-1 viral load	0.439
	T helper	0.431
	Tsh w/ free t4 reflex	0.410
	Virus identification	0.385

4. Discussion

For the first time, we assessed how disease cohorts automatically generated from an EHR system compare to research grade gold standard electronic phenotyping algorithms from PheKB for five diseases: ADHD, Dementia, Herpes Zoster, Sickle Cell, and T2D. As an automated method, this approach is purely data driven and requires no manual effort beyond selection of a single seed

concept. Specifically, we employed the *word2vec* algorithm to create medical concept embeddings of the phenotype space. For each disease of interest, we query the embeddings using a representative seed concept, which is automatically expanded to include highly related concepts nearby in the low-dimensional space. Overall, both embedding methods (i.e. with and without expansion) outperformed using ICD-9 codes alone; precision, however, was higher than the other methods but is most likely due to fact that the manual phenotyping algorithms themselves incorporate the code. Further, the much lower recall and poorer Retrieval outcomes indicate that using ICD codes likely miss many cases. Querying with expansion improved all metrics over using the raw embeddings alone, which is one of the strongest aspects of this work. While the performance at the disease level varied, the overall evaluation metrics are encouraging, especially instances like Sickle Cell, which performed the best.

4.1. *Limitations and Future Directions*

Many factors likely affected the performance comparison between our automated phenotyping method and the PheKB algorithms. For instance, many of the PheKB algorithms incorporate selection criteria based on amount and/or temporal length of data in a patient's record, which was not considered in the current iteration of our method. These scenarios might lead to mismatched labels due to non-phenotype related properties. Another important drawback is our seeding of the queries with ICD codes. Although we overcome many of the limitations of using ICD codes alone to electronically phenotype (i.e., low recall), it is difficult to learn across the branches of the ICD structure: for instance, it may be desirable to delineate between related phenotypes at the same hierarchical level (e.g. type 1 vs. type 2 diabetes mellitus) but since these are both branches of the major diabetes ICD code used as a seed for Type 2 Diabetes, our algorithm was not able to easily distinguish between them and lead to subpar performance. While the expanded query still performed moderately well, this caveat exemplifies that room for improvement exists. Specifically, the seed and query expansion might perform better as a learned subgraph of related concepts, such as the anchor and learn framework utilized by Halpern and Horng et al. Additionally, one of the largest limitations of the current study is that of weak labels: we could not evaluate performance of the embeddings separately and in addition to the electronic phenotyping algorithms via access to patient charts. We expect even the gold standard phenotyping algorithms to erroneously include and exclude patients. Compared to the true phenotype, we could potentially be identifying patients that are captured in the automated method but not in the phenotyping algorithms. In future work, we will also obtain clinical notes to expand our comparison to all disease-related algorithms in PheKB.

There are many extensions we wish to pursue that can address current limitations as well as strengthen performance. We hope to enhance performance through advancing the patient

embedding representation, testing other methodologies such as GloVE [32], as well as developing superior ways to summarize clinical history that keeps into account timeline. To bypass the need for data pre-processing and harmonization, we plan to standardize our raw EHR data to OMOP Common Data Model, from the Observational Health Data Sciences and Informatics (OHDSI). Further, the OHDSI framework would enable cross-validation experiments within other coordinated hospital EHR systems.

5. Acknowledgments

We would like to thank the Mount Sinai Data Warehouse for facilitating data accessibility and the Mount Sinai Scientific Computing team for infrastructural support. This study was funded by the following grants of JTD: National Institute of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) R01-DK098242-03 and the Harris Center for Precision Wellness.

References

1 P. B. Jensen, L. J. Jensen and S. Brunak, *Nat Rev Genet*, **13**. (2012)
2 J. Pathak, A. N. Kho and J. C. Denny, *J Am Med Inform Assoc*, **20**. (2013)
3 P. Yadav, M. Steinbach, V. Kumar and G. Simon, ArXiv e-prints, **1702**. (2017)
4 National Research Council, National Academies Press, (2011), ISBN: 0309222257
5 W. Q. Wei, P. L. Teixeira, H. Mo, R. M. Cronin*, et al.*, *J Am Med Inform Assoc*, **23**. (2016)
6 K. P. Liao, T. Cai, G. K. Savova, S. N. Murphy*, et al.*, *BMJ*, **350**. (2015)
7 C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi*, et al.*, *J Am Med Inform Assoc*, **21**. (2014)
8 O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett*, et al.*, *Genet Med*, **15**. (2013)
9 J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford*, et al.*, *J Am Med Inform Assoc*, **23**. (2016)
10 L. V. Rasmussen, W. K. Thompson, J. A. Pacheco, A. N. Kho*, et al.*, *J Biomed Inform*, **51**. (2014)
11 Y. Bengio, A. Courville and P. Vincent, ArXiv e-prints, **1206**. (2012)
12 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado*, et al.*, *Advances in neural information processing systems*, 3111-3119. (2013)
13 Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, *Journal of machine learning research*, **3**. (2003)
14 T. Mikolov, K. Chen, G. Corrado and J. Dean, *arXiv preprint arXiv:1301.3781*. (2013)
15 Y. Choi, C. Y. Chiu and D. Sontag, *AMIA Jt Summits Transl Sci Proc*, **2016**. (2016)
16 Y. Halpern, S. Horng, Y. Choi and D. Sontag, *J Am Med Inform Assoc*, **23**. (2016)
17 M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng*, et al.*, *Sci Rep*, **7**. (2017)
18 R. Miotto, L. Li, B. A. Kidd and J. T. Dudley, *Sci Rep*, **6**. (2016)
19 S. Kandula, Q. Zeng-Treitler, L. Chen, W. L. Salomon*, et al.*, *J Biomed Inform*, **44 Suppl 1**. (2011)
20 S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer*, et al.*, *J Am Med Inform Assoc*, **22**. (2015)
21 R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo*, et al.*, *J Biomed Inform*, **58**. (2015)
22 P. H. Chiu and G. Hripcsak, *J Biomed Inform*, **70**. (2017)
23 V. Agarwal, T. Podchiyska, J. M. Banda, V. Goel*, et al.*, *J Am Med Inform Assoc*, **23**. (2016)
24 J. Connolly, CHOP, PheKB, (2013). https://phekb.org/phenotype/179
25 C. Carlson, Group Health Cooperative, PheKB, (2012). https://phekb.org/phenotype/10
26 Group Health and University of Washington, Group Health and University of Washington, PheKB, (2012). https://phekb.org/phenotype/112
27 D. E. Michalik, B. W. Taylor and J. A. Panepinto, *Acad Pediatr*, **17**. (2017)
28 D. E. Maichalik and J. A. Panepinto, PheKB, Medical College of Wisconsin, (2017). https://phekb.org/phenotype/615
29 J. Pacheco and W. Thompson, Northwestern University, PheKB, (2012). https://phekb.org/phenotype/18
30 C. Jonquet, N. H. Shah and M. A. Musen, *Summit Transl Bioinform*, **2009**. (2009)
31 S. Arora, Y. Liang and T. Ma, *International Conference on Learning Representations*. (2016)
32 J. Pennington, R. Socher and C. D. Manning, *EMNLP*, **14**, 1532. (2014)

Functional network community detection can disaggregate and filter multiple underlying pathways in enrichment analyses

Lia X. Harrington
Department of Biomedical Data Science,
Geisel School of Medicine at Dartmouth College, Hanover 03784, USA
Email: lia.harrington.gr@dartmouth.edu

Gregory P. Way
Department of Systems Pharmacology and Translational Therapeutics,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: gregway@mail.med.upenn.edu

Jennifer A. Doherty
Huntsman Cancer Institute, Population Health Sciences, University of Utah,
Salt Lake City, UT 84112-5550, USA
Email: jen.doherty@hci.utah.edu

Casey S. Greene
Department of Systems Pharmacology and Translational Therapeutics,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: csgreene@mail.med.upenn.edu

Differential expression experiments or other analyses often end in a list of genes. Pathway enrichment analysis is one method to discern important biological signals and patterns from noisy expression data. However, pathway enrichment analysis may perform suboptimally in situations where there are multiple implicated pathways – such as in the case of genes that define subtypes of complex diseases. Our simulation study shows that in this setting, standard overrepresentation analysis identifies many false positive pathways along with the true positives. These false positives hamper investigators' attempts to glean biological insights from enrichment analysis. We develop and evaluate an approach that combines community detection over functional networks with pathway enrichment to reduce false positives. Our simulation study demonstrates that a large reduction in false positives can be obtained with a small decrease in power. Though we hypothesized that multiple communities might underlie previously described subtypes of high-grade serous ovarian cancer and applied this approach, our results do not support this hypothesis. In summary, applying community detection before enrichment analysis may ease interpretation for complex gene sets that represent multiple distinct pathways.

Keywords: gene enrichment analysis, community detection, power, false positives, pathway analysis

1. Introduction

Researchers' experiments that include high-throughput data generation often lead to a set of genes. These genes may be genes that are over- or under-expressed in a disease subtype, are upregulated in response to a drug, or contain variants associated with a disease. After potentially interesting genes are identified, the next challenge is to interpret the biological processes or pathways that underlie the set. Overrepresentation-based methods are commonly used to identify pathways that have more members in the identified set than would be expected by chance[1]. Typically, pathways or similar groups of genes are obtained from structured vocabularies outlined in curated ontologies such as KEGG, PID, GO, or Reactome[2–5]. Recently, computational researchers have sought to improve the power of such analyses by considering network interactions among pathway members[6,7]. We sought to evaluate overrepresentation analysis in a different setting: one where multiple pathways underlie a set of associated genes. In this situation, applying standard overrepresentation analysis to gene sets constructed by randomly selecting members of multiple pathways identifies many false positive pathways. We hypothesized that reducing the noise of the gene list input via community detection might decrease the number of false positive pathways.

Functional networks are a type of network where genes are connected if they have a high probability of working together in the same pathway or process[8–11]. To address the challenge posed by multi-pathway gene sets, we developed an approach that incorporates information from functional networks to first partition gene sets into subsets, or communities, which are then analyzed for overrepresented pathways. To accomplish this, enrichment analysis is applied to each extracted community resulting from community detection preprocessing[12,13] of the original gene set. Community detection has been applied to financial data, social media, and biological data[12,14]. To our knowledge, this is its first application to disambiguate the pathways associated with complex gene sets. We evaluate four community detection methods in this context: Fastgreedy, Walktrap, Multilevel, and Infomap. These algorithms all aim to identify groups/communities within a network:

- Fastgreedy – This algorithm starts from a completely unclustered set of nodes and iteratively adds communities such that the modularity (score maximizing within edges and minimizing between edges) is maximized until no additional improvement can be made[15].
- Walktrap – This algorithm performs random walks using a specified step size. Where densely connected areas occur, the random walk becomes "trapped" in local regions that then define communities[16].
- Multilevel – This algorithm is similar to fastgreedy, but it merges communities to optimize modularity based upon only the neighboring communities as opposed to all communities[17]. The algorithm terminates when only a single node is left, or when the improvement in modularity cannot result from the simple merge of two neighboring communities.
- Infomap – This algorithm uses the probability flow of information in random walks, which occurs more readily in groups of heavily connected nodes. Thus, information about network structure can be compressed in maps of modules (nodes where information travels quickly)[18].

Outside of the multi-pathway gene set challenge, there are a number of R packages that implement algorithms for network interpretation of experimental results including WGCNA[19], EnrichNet[20], pathDIP[21], and CePa[22,23]. In this work, community detection algorithms are used to partition multi-pathway gene sets before overrepresentation analysis. By detecting these gene communities, we aim to provide cleaner inputs for overrepresentation analyses in the case of multiple underlying pathways – thereby reducing the number of identified false positives. In contrast with other methods that use network information as priors or as post-analysis visualization aides, we group genes before enrichment analysis. While we use the Integrative Multi-species Prediction (IMP) networks, our approach can be applied to a gene set from any source[11,24]. For example, a user may wish to use tissue-specific networks from the GIANT webserver[9] if tissue specificity is important. Finally, our approach makes no assumptions about the covariance structure of the networks[25] and is thus potentially more useful in real world applications where certain assumptions may not apply.

In summary, we propose an alternative gene enrichment approach for cases when multiple pathways are suspected to be implicated in a gene list. In this approach, candidate genes are overlaid onto a functional network and separated into communities of related genes via community detection. Communities are then subjected to an overrepresentation analysis independently and multiple testing corrections are applied. We compare four community detection approaches in simulated experiments and then apply the approach to identifying enriched pathways across high grade serous ovarian cancer (HGSC) subtypes.

2. Methods

We conducted an experiment that contained a control and an experimental arm. The control arm was an overrepresentation analysis without community detection, and the experimental arm was an overrepresentation analysis with various community detection methods applied as a preprocessing step.

2.1. *General Approach*

From the KEGG ontology, m randomly chosen pathways were selected to form a list of candidate genes. To help evaluate the impact of incomplete pathway discovery, only p percent of the genes in each pathway were randomly selected for inclusion in the final gene list. Finally, a percent of additional random genes selected without replacement from the ontology were added to the gene list to create noise. As to only consider genes that influence pathway analysis, genes that were not in both IMP and KEGG were excluded for a resulting set of 5195 genes. This procedure was performed for both control and experimental arms so that differences in results could be attributed to community detection preprocessing.

We performed one hundred iterations for each parameter level combination of number of pathways (m = 2-8), percentage of genes included from each pathway (p = 30%, 47.5%, 65%, 82.5%, and 100%), and percentage additional random genes from IMP (a = 10%, 32.5%, 55%, 77.5%, and 100%) for a total of 105,000 individual runs. Over the 100 iterations of the specific parameter combination, we measured the number of seeded pathways correctly detected (true

positives), incorrectly detected (false positives), correctly missed (true negatives), and incorrectly missed (false negatives). The false positive proportion, false negative proportion, precision, recall, and F1 score were calculated for each parameter combination over the 100 iterations. The F1 score is the weighted average of precision and recall where precision is the number of true positives divided by all positives and recall is the number of true positives divided by the sum of true positives and false negatives.

2.2. *Control Arm*

The control arm followed the steps outlined in General Approach.

2.2.1. *Control All (CtrAll)*

For this method, we determined true positives, false positives, true negatives, and false negatives using all significantly enriched pathways and complete gene lists of seeded pathways. For example, if a gene list was seeded with three pathways and the enrichment analysis identified ten pathways (including correctly identifying the original three), then all ten pathways would be counted as positives with the seven unseeded pathways considered false positive.

2.2.2. *Control M (CtrM)*

For this method, true positives, false positives, true negatives, and false negatives were determined using only the top *m* significant pathways where *m* is the number of seeded pathways. For example, if three pathways were seeded and there were ten significant pathways, then only the top three pathways in the significant enrichment results would be considered. Thus, if all three seeded pathways were in the top three significant results, the true positive would be three and false positive would be zero. If, however, only two of the three seeded pathways were in the top three significantly enriched pathways, then true positive would be two and false positive would be one. CtrM provides provides an upper bound on possible performance as it is unrealistic in practice for investigators to know *a priori* the correct number of pathways.

2.3. *Experimental Arm*

For the experimental arm, the subgraph associated with each gene list described in the General Approach was extracted from IMP and subjected to community detection to provide community-level gene sets before the overrepresentation analysis. Fastgreedy, Walktrap, Infomap, and Multilevel community detection algorithms were applied in the community detection step. The communities of genes detected by the algorithm were then used as separate candidate gene lists for overrepresentation analysis. True positive, false positive, true negative, and false negative were calculated for all pathways that remained statistically significant after Bonferroni multiple testing correction at $\alpha = .05$ was applied. This correction was applied for each community if multiple were found.

All simulation analyses were performed using Python 2.7.6 with the iGraph package (version 0.71). Figures were produced using ggplot in R 3.3.1. Open source software to reproduce the

results of this paper is provided at https://github.com/greenelab/GEA_Community_Detection. Figure 1 provides an overview of both the control and experimental arms.

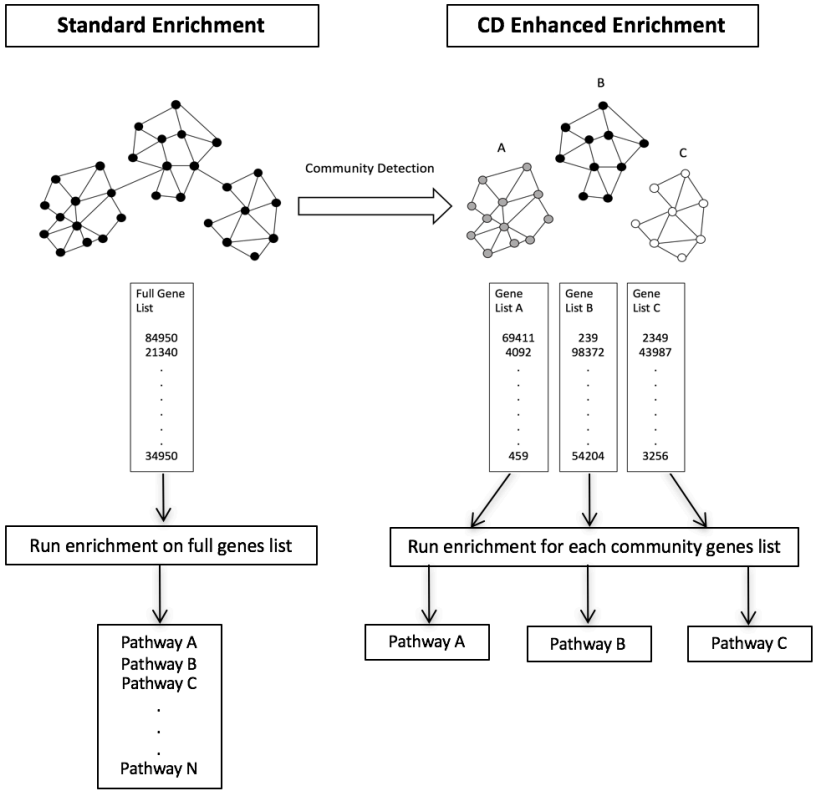


Fig. 1. In standard enrichment analysis, the full gene list is subjected to enrichment analysis and all significantly enriched pathways are returned. In the proposed experimental community detection enhanced method, the full gene list is first subjected to community detection to parse the gene list into sub-gene lists. Enrichment analysis is then performed for each gene list associated with each "discovered" community. Only the most significant pathway is returned for each community.

HGSC Application

Based on the results of the simulation study, we applied the top performing community detection algorithms to lists of genes characterizing high-grade serous ovarian cancer (HGSC) subtypes. The gene lists were previously identified by a one cluster versus all differential expression analysis[26] of cluster specific genes in common to four HGSC datasets[27–30]. While previous reports have described four HGSC subtypes, the multi-population study suggested that the number was three or fewer[26]. Given these conflicting results, we applied community detection to HGSC subtype-specific gene lists previously derived from results classifying 2, 3, and 4 subtypes[26]. Because this is an analysis of cancer genomics data, we used cancer pathways from the Pathway Interaction Database (PID)[5].

3. Results and Discussion

3.1. *Simulation Study*

In general, community detection methods reduced the number of false positive associations in the multi-pathway setting. When seeding a gene list with four random pathways, all community detection methods had higher F1 scores than the standard enrichment analysis, CtrAll (Figure 2). In cases where pathways were incompletely seeded, the community detection methods often outperformed CtrM, which only considers the top *m* pathways as statistically significant (Figure 2). These findings are consistent when using the top 2-8 pathways (pathway numbers 2, 3, 5, 6, 7,

and 8 are Supplementary Figures S1-6). Performance was robust to the number of genes taken from each seeded pathway over a broad range of values, and the relative performance of methods was largely unaffected by the proportion of genes sampled from the seeded pathways (i.e, 30% or all 100%) to make the gene lists. Thus, our approach may be more useful than standard enrichment techniques in situations where one is presented with a long, heterogeneous, and incomplete gene list and one wishes to find a set of robust pathways for further investigation. The Walktrap and Multilevel methods demonstrated the most success in this context as they resulted in high F1 scores and relatively low false negative and false positive proportions. Compared to other community detection methods, Fastgreedy appeared to have a broader range of performance values, with higher variability and increased outliers. The performance of community detection algorithms may be network-specific; users may wish to apply our open source code to perform a new simulation study if different networks are selected.

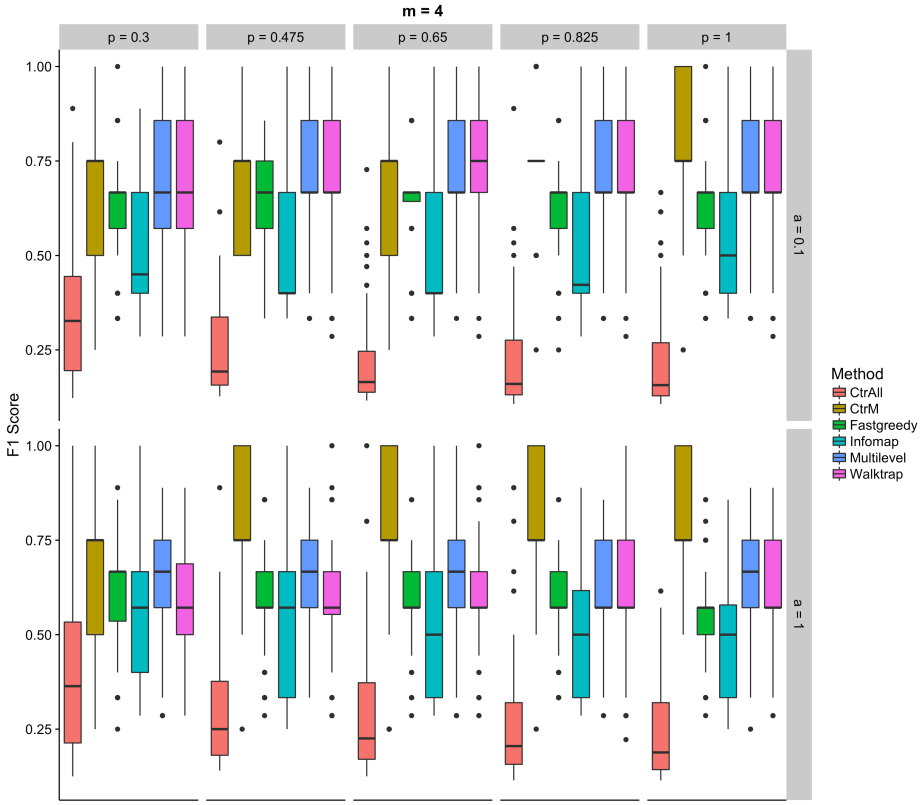


Fig. 2. F1 scores for the controls (using all (CtrAll), or only the top 4 (CtrM), statistically significant pathways) and the community detection methods: Fastgreedy, Infomap, Multilevel, and Walktrap for various percentages of genes in each pathway (top axis) and percentages of additional genes (right side axis) for simulations using 4 random pathways. The percentage of genes indicates the percentage of random genes selected from each pathway. The percentage of additional genes indicates how many unrelated genes are randomly added to the analysis to represent increasing amounts of noise. Each comparison includes 100 iterations.

The combination of community detection and enrichment was designed to filter false positives in the multi-pathway setting. When we evaluated the proportion of false positives, we observed that the F1 score improvements were driven by successful filtration. In particular, all community detection methods outperformed standard enrichment analyses for false positive proportions (Figure 3). As expected, when the number of seeded pathways increased, the proportions of false positives steadily increased for control runs that included all statistically significant pathways. The standard enrichment analysis approach was well suited to identifying a single pathway. The more pathways that were present in a single genelist, the worse standard enrichment-based methods performed.

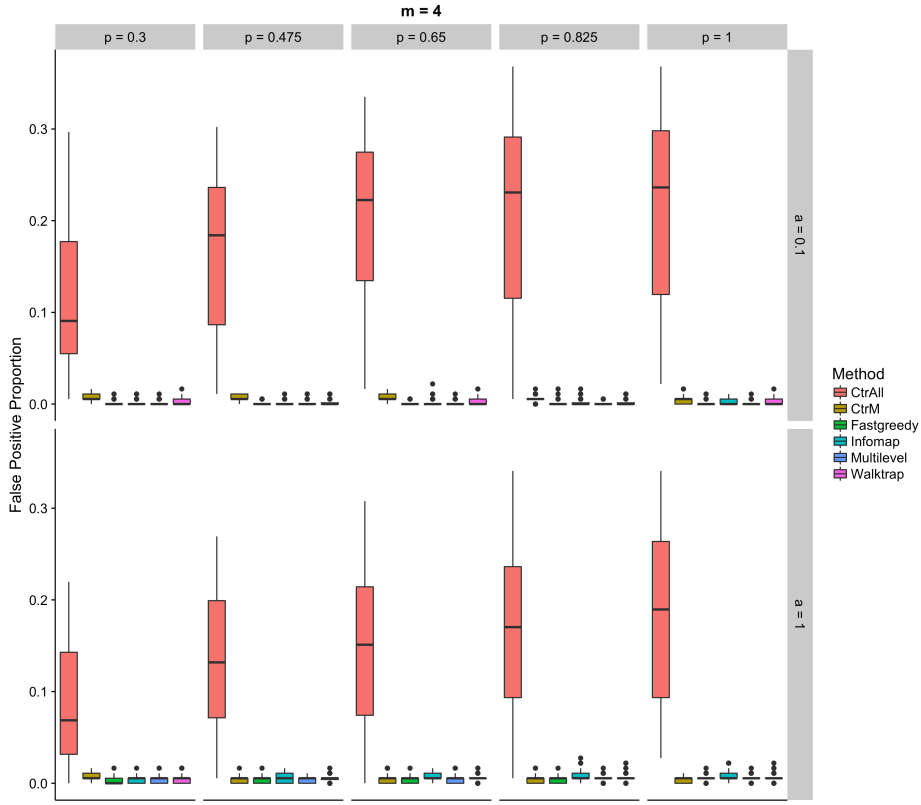


Fig. 3. Proportions of false positives for the controls (using all (CtrAll), or only the top 4 (CtrM), statistically significant pathways) and the community detection methods: Fastgreedy, Infomap, Walktrap, and Multilevel for various percentages of genes in each pathway (top axis) and percentage of additional genes (right side axis) for simulations using 4 random pathways.

All community detection methods other than CtrAll usually miss some portion of the true positives using 4 seeded pathways (Figure 4). In general, Walktrap, Infomap, and Multilevel tend to have greater variability in the number of pathways missed compared to CtrAll and Fastgreedy. It is not surprising that the community detection and CtrM methods have higher proportions of false negatives than CtrAll since they were designed to reduce false positives. Thus, a traditional

enrichment approach may be more appropriate in sitatuions where false negatives are more of a concern, such as when investigating a relatively small gene list or conducting an exploratory analysis.

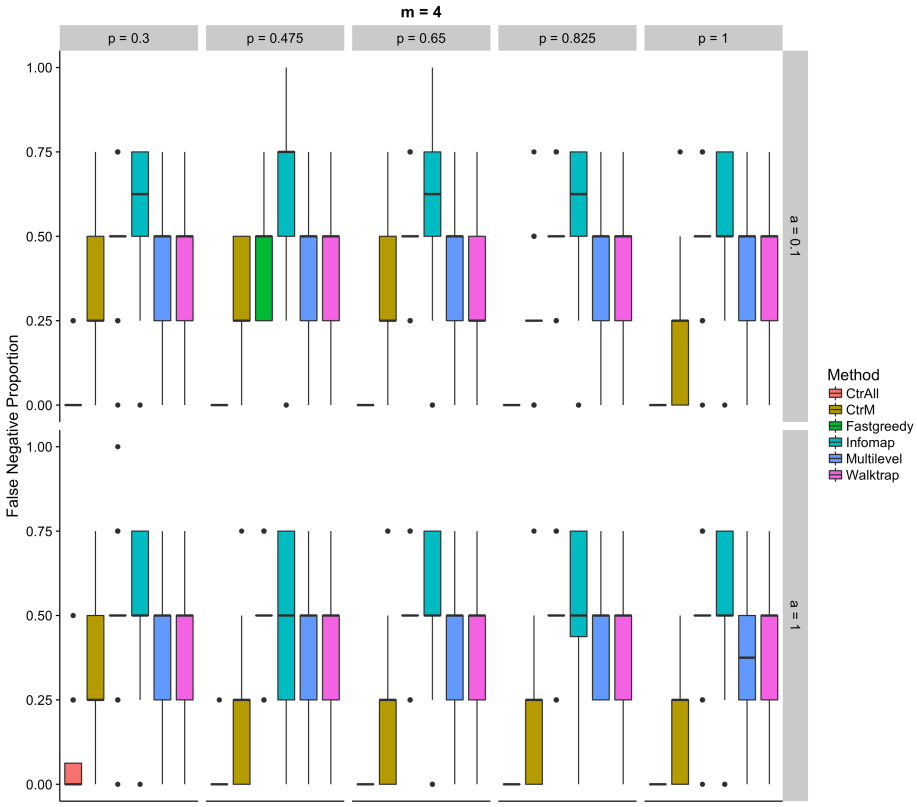


Fig. 4. Proportions of false negatives in the controls (using all (CtrAll), or only the top 4 (CtrM), statistically significant pathways) and the community detection methods: Fastgreedy, Infomap, Walktrap, and Multilevel for various percentage of genes in each pathway (top axis) and percentage of additional genes (right side axis) for simulations using 4 random pathways.

3.2. *HGSC Results*

To examine the biological applicability of community detection, we independently applied the community detection approach to previously defined, HGSC subtype-specific gene lists for when 2, 3, and 4 subtypes are assigned. We previously derived these gene lists from a differential expression analysis across HGSC subtypes that were concordant across different populations[26]. We selected only the top performing algorithms from our simulation study, Walktrap and Multilevel. Applying these methods to PID pathways, we found that most clusters mapped to either Beta1 integrin cell surface interactions or IL12-mediated signaling events (Table 1). Community detection methods was able to separate upregulated and downregulated genes coming from the same pathway into different communities (Table 1).

While many pathways were implicated in the original pathway analysis (see Supplementary Table S6 of Way et al. 2016[26]), our community detection approach only implicated two distinct pathways consistently, for 2-4 subtypes. This did not support our hypothesis that HGSC subtypes are driven by differences across multiple pathways that are captured in differentially expressed gene lists. HGSC subtypes are known to be primarily characterized by a mesenchymal gene signature and immunoreactivity. Our analysis suggested that up- and down-regulation of beta 1 integrin signaling, and down-regulation of IL12 signaling, primarily define the subtype-specific signatures. However, the lack of PID pathway enrchiment in the presence of community structure may indicate novel biological pathways driving subtype separation. Beta 1 integrin signaling is a well characterized indicator of metastasis[31] and its high expression is associated with poor survival in ovarian cancer patients[32]. IL12 is an important immune system process with many coordinated functions[33]. Importantly, administration of intraperitoneal IL12 is being explored as a therapeutic agent in ovarian cancer[34]. The community detection approach pointed to specific HGSC subtypes that were aligned with this characterization, but did not identify multiple pathways for any specific subtype. We often observed that pathways that were highly expressed for one subtype would be underexpressed for another, which was consistent with a model that HGSC subtypes exist along a continuum of underlying pathway or cell type content. These results are also generally consistent with those found previously[27, 28,35 ,36].

Table 1. The statistically significantly enriched pathways found by Walktrap and Multilevel community detection methods and the number of genes in each pathway that are either upregulated (more highly expressed) or downregulated (less expressed) in HGSC[26]. We identified statistically significant pathways in communities defined by only k = 4 in cluster 1 (k4c1), cluster 2 (k4c2), and cluster 4 (k4c4). The id number of the enriched community is also provided. Clusters 1, 2, 3 and 4 correspond to mesenchymal, proliferative, immunoreactive, and differentiated subtypes as previously defined by TCGA[27].

Cluster	Community	Method	Pathway Name	p-value	Downregulated	Upregulated
k4c1	0	Walktrap	Beta1 integrin cell surface interactions integrin1_pathway	8.01E-06	0	23
k4c4	1	Walktrap	Beta1 integrin cell surface interactions integrin1_pathway	2.87E-06	12	0
k4c1	2	Multilevel	Beta1 integrin cell surface interactions integrin1_pathway	8.47E-05	0	23
k4c2	0	Multilevel	IL12-mediated signaling events il12_2pathway	1.99E-04	22	0
k4c4	1	Multilevel	Beta1 integrin cell surface interactions integrin1_pathway	2.30E-05	12	0

4. Conclusion

In summary, we developed an alternative enrichment method that uses community detection to group genes based on network connectivity prior to enrichment analyses. This approach is designed for situations where a researcher hypothesizes that multiple pathways contribute to a gene set. It trades an increase in false negatives for a dramatic reduction in false positives. The standard enrichment approach may be more appropriate in exploratory stages of research when high power is more desired than false positive control. Applying this method to gene sets that characterize HGSC subtypes did not reveal multiple pathways underlying any of the previously described subtypes. These results are consistent with a model where factors other than the activity of multiple pathways are responsible for the difficult to discern HGSC subtypes.

5. Acknowledgments

We thank James Rudd for several thoughtful conversations about the developed approach. This work was funded in part by grants from the Gordon and Betty Moore Foundation (GBMF 4552) to CSG and from the National Institutes of Health (R01 CA200854) to JAD and CSG.

6. Supplementary Material

Supplementary figures can be found at http://doi.org/10.5281/zenodo.830568[37].

References

1. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology* (2012). doi:10.1371/journal.pcbi.1002375
2. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* (1999). doi:10.1093/nar/27.1.29
3. Gene Ontology Consortium, T. *et al.* Gene Ontology: tool for the unification of biology.
4. Joshi-Tope, G. *et al.* Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki072
5. Schaefer, C. F. *et al.* PID : the Pathway Interaction Database. **37,** 674–679 (2009).
6. Kong, B. *et al.* Protein – protein interaction network analysis and gene set enrichment analysis in epilepsy patients with brain cancer. *J. Clin. Neurosci.* **21,** 316–319 (2014).
7. Ma, J., Shojaie, A. & Michailidis, G. Systems biology Network-based pathway enrichment analysis with incomplete network information. **32,** 3165–3174 (2016).
8. Szklarczyk, D. *et al.* The STRING database in 2017 : quality-controlled protein – protein association networks , made broadly accessible. **45,** 362–368 (2017).
9. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** 569–576 (2015).
10. Goya, J. *et al.* FNTM: A server for predicting functional networks of tissues in mouse. *Nucleic Acids Res.* **43,** W182–W187 (2015).
11. Wong, A. K. *et al.* IMP : a multi-species functional genomics portal for integration , visualization and prediction of protein functions and networks. **40,** 484–490 (2012).
12. Malliaros, F. D. & Vazirgiannis, M. Clustering and community detection in directed networks: A survey. *Phys. Rep.* **533,** 95–142 (2013).
13. Fortunato, S. Community detection in graphs. *Phys. Rep.* **486,** 75–174 (2010).
14. Harenberg, S. *et al.* Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdiscip. Rev. Comput. Stat.* **6,** 426–439 (2014).
15. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Physics (College. Park. Md).* 1–6 (2004). doi:10.1103/PhysRevE.70.066111
16. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10,** 191–218 (2006).
17. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10008,** 6 (2008).
18. Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 7327–31

(2007).
19. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
20. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: Network-based gene set enrichment analysis. *Bioinformatics* **28,** 451–457 (2012).
21. Rahmati, S., Abovsky, M., Pastrello, C. & Jurisica, I. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic Acids Res.* gkw1082 (2016). doi:10.1093/NAR/GKW1082
22. Gu, Z. & Wang, J. CePa: An R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* **29,** 658–660 (2013).
23. Gu, Z., Liu, J., Cao, K., Zhang, J. & Wang, J. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. doi:10.1186/1752-0509-6-56
24. Wong, A. K., Krishnan, A., Yao, V., Tadych, A. & Troyanskaya, G. IMP 2 . 0 : a multi-species functional genomics portal for integration , visualization and prediction of protein functions and networks. **43,** 128–133 (2015).
25. MacMahon, M. & Garlaschelli, D. Community detection for correlation matrices. *Phys. Rev. X* **5,** 1–34 (2015).
26. Way, G. P. *et al.* Comprehensive Cross-Population Analysis of High-Grade Serous Ovarian Cancer Supports No More Than Three Subtypes. **6,** 4097–4103 (2016).
27. Cancer, T. & Atlas, G. Integrated genomic analyses of ovarian carcinoma. 0–7 (2012). doi:10.1038/nature10166
28. Konecny, G. E. *et al.* Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* **106,** (2014).
29. Yoshihara, K., Tsunoda, T., Shigemizu, D., Fujiwara, H. & Hatae, M. High-Risk Ovarian Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by Downregulation of Antigen Presentation Pathway. **18,** 1374–1386 (2012).
30. Tothill, R. W. *et al.* Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. **14,** 5198–5209 (2008).
31. Kato, H. *et al.* The Primacy of β1 Integrin Activation in the Metastatic Cascade. *PLoS One* **7,** 1–11 (2012).
32. Watanabe, T. *et al.* Production of IL1-beta by ovarian cancer cells induces mesothelial cell beta1-integrin expression facilitating peritoneal dissemination. *J. Ovarian Res.* **5,** 7 (2012).
33. Liu, J. *et al.* Interleukin-12: an update on its immunological activities, signaling and regulation of gene expression.
34. Anwer, K., Barnes, M. N., Fewell, J., Lewis, D. H. & Alvarez, R. D. Phase-I clinical trial of IL-12 plasmid/lipopolymer complexes for the treatment of recurrent ovarian cancer. *Gene Ther.* (2010). doi:10.1038/gt.2009.159
35. Verhaak, R. & Tamayo, P. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. …* **123,** 1–9 (2012).
36. Wang, C. *et al.* Pooled Clustering of High-Grade Serous Ovarian Cancer Gene Expression Leads to Novel Consensus Subtypes Associated with Survival and Surgical Outcomes. *Clin. Cancer Res.* **123,** 1–9 (2017).
37. Harrington, L. X., Way, G. P., Doherty, J. A. & Greene, C. S. Gene Enrichment Analysis via Community Detection. *Zenodo* (2017). doi:10.5281/zenodo.830568

An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data

Hyun-Hwan Jeong[1,2], Hari Krishna Yalamanchili[1,2], Caiwei Guo[2,3],
Joshua M. Shulman[1,2,3,4], Zhandong Liu[2,5,†]

[1]*Department of Molecular and Human Genetics, Baylor College of Medicine,*
[2]*Jan and Dan Duncan Neurological Research Institute, Texas Childrens Hospital,*
[3]*Department of Neuroscience, Baylor College of Medicine,*
[4]*Department of Neurology, Baylor College of Medicine,*
[5]*Department of Pediatrics, Baylor College of Medicine,*
Houston, Texas 77030, USA
[†]*E-mail: zhandonl@bcm.edu*

Transposable elements (TEs) are DNA sequences which are capable of moving from one location to another and represent a large proportion (45%) of the human genome. TEs have functional roles in a variety of biological phenomena such as cancer, neurodegenerative disease, and aging. Rapid development in RNA-sequencing technology has enabled us, for the first time, to study the activity of TE at the systems level.

However, efficient TE analysis tools are not yet developed. In this work, we developed `SalmonTE`, a fast and reliable pipeline for the quantification of TEs from RNA-seq data. We benchmarked our tool against `TEtranscripts`, a widely used TE quantification method, and three other quantification methods using several RNA-seq datasets from Drosophila melanogaster and human cell-line. We achieved 20 times faster execution speed without compromising the accuracy. This pipeline will enable the biomedical research community to quantify and analyze TEs from large amounts of data and lead to novel TE centric discoveries.

Keywords: Transposable Element; Quasi-Mapping; RNA-seq; Next Generation Sequencing; Large Scale Genome Analysis

1. Introduction

Transposable elements (TEs) are DNA elements which can be mobilized or inserted into the genome and represent a significant proportion of most eukaryotic genomes.[1] Most of the TEs in the genome are not functional and had been considered as 'junk DNA,' except for a few that retain intact functions such as transcription and mobilization.[2] Furthermore, the mobilization of TEs can disrupt normal gene structure in the genome, sometimes leading to disease such as cancer[3,4] neurodegenerative diseases,[1] and aging.[5]

Recent development of high-throughput Next Generation Sequencing (NGS) technologies, like RNA-seq, enables genome-wide study for TEs.[6–9] Toward this end, several algorithms and pipelines were proposed to analyze reads files from TE studies.[10–16] However, most of the

tools share some common limitations: 1) discordant read mapping due to increased chance of multiple mapping in repetitive elements from TEs in the same clade, 2) limited scalability for large-scale analysis, and 3) small coverage for the entire TEs defined in the human genome, i.e., a tool used in [16] only considered LINE 1 (Long Interspersed Nuclear Element 1) elements.[17]

Among the existing tools, `TEtranscripts` has performed well on various datasets.[14] Nonetheless, The scalability of `TEtranscripts` is a critical limiting factor for large systems biology studies because it cannot handle `FASTQ` files directly and needs `SAM` (Sequence Alignment Map)/`BAM` (Binary Sequence Alignment Map) files generated from raw `FASTQ` files. Since there are many tuning parameters on handling repetitive sequence among different RNA-seq mapping algorithms, this step will be highly variable depending on the mapping parameters and sometimes even generate artifactual results if a unique mapping parameter is superimposed by a previous analyst who handled the mapping.

Although `TEtranscripts` is the fastest tool for TE quantification,[14] the interval tree algorithm,[18] which is used to find the interval of genes or TEs on the reference genome, performed poorly in terms of running time in practice, making `TEtranscripts` suboptimal for large-scale TE analysis.

In recent studies, many large-scale analysis of public meta RNA-seq datasets offered new insight and findings that cannot be discovered in each dataset alone.[19] However, a meta-study on TE without using a large number of high-performance computing cluster is not yet feasible given the time complexity of current algorithms. Toward this end, we developed a new pipeline called `SalmonTE`. It deploys a low time-complexity quantification method, `Salmon`,[20] and contains various statistical models for TEs quantification. Moreover, `SalmonTE` provides a rich set of built-in functions for data pre-processing from raw `FASTQ` files. In the results section, we demonstrate the running speed of `SalmonTE` outperforms all other methods including `TEtranscripts` and delivers a reliable quantification result as well.

2. Methods

The proposed pipeline consists of three parts: library preparation, quantification, and statistical analysis (Figure 1). To increase the usability and to enable parallel processing for multiple RNA-seq reads files, we adopted the `Snakemake` workflow system and wrote a script based on the execution rule of `Snakemake` for the TE quantification.[21] In contrast to TEtranscripts, SalmonTE starts with raw RNA-seq files, and does not need any additional pre-processing for a given sequence file. Moreover, TEtranscripts requires a modified GTF files based on RepeatMasker database.[22] SalmonTE only needs the FASTA file of cDNA (complementary DNA) sequences of each TE. The entire source code and executable scripts are available at `https://github.com/hyunhwaj/SalmonTE`.

2.1. *Transposable Element Library Preparation*

To build the index library for the quasi-mapping, `SalmonTE` takes the FASTA file of cDNA sequences from TE databases such as Repbase (version 22.06).[23] In the current version, the index files for *Homo sapiens* and *Drosophila melanogaster* are available. We reasoned that it is hard to estimate TEs which replicate without an RNA intermediate from RNA-seq sample.

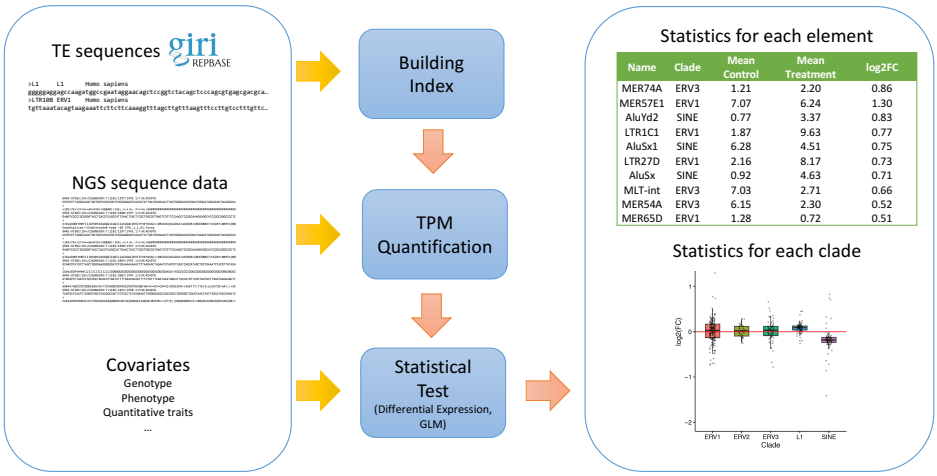


Fig. 1. An illustration of the `SalmonTE` pipeline. Left Panel: Input from Repbase to build the mapping index, raw `FASTQ` file, and covariates for statistical testing. Middle Panel: The workflow of `SalmonTE` consists of three parts: building the index based on Repbase or user-input cDNA sequences of TEs, quantification based on `FASTQ` file, and statical test through the generalized linear model or differential expression analysis. Right Panel: Example output including the statistical report and box plot on estimated log_2 fold-change.

Therefore, we excluded the following elements: simple repeats and multi-copy genes, and DNA transposable. After collecting the cDNA sequences, we manually curated clades of each TE based on the repeat class annotation from Repbase.
As a result, the generated TE library index database contains 687 TEs for *Homo sapiens* and 163 TEs for *Drosophila melanogaster*.

2.2. *Salmon quantification algorithm*

We adopted the `Salmon` [20] algorithm to estimate the relative TE abundance from a given RNA-seq sample. `Salmon` enables a fast and accurate quantification of TE expression from RNA-seq reads with a light-weight mapping, online initial expression estimation phase, and offline inference for the estimation refinement.[20,24–26] Salmon quantifies the relative abundance of each TE given a set of TE sequences T and a set of sequenced fragments (reads) F. Suppose that we have M TEs and the set of underlying true TE counts are given as $T = \{(t_1, \ldots, t_M), (c_1, \ldots, c_M)\}$, where t_i is the nucleotide sequence of i-th TE in the set and c_i is the true count of the corresponding TE. If T contains a complete count, we can calculate the nucleotide fraction η_i of each t_i from (1),

$$\eta_i = \frac{c_i \cdot \widetilde{l_i}}{\sum_{j=1}^{M} c_j \cdot \widetilde{l_j}} \tag{1}$$

where $\widetilde{l_i}$ is the effective transcript length of t_i.[27]

We can also calculate the Transcripts Per Million (TPM) using (2),

$$TPM_i = \frac{\frac{\eta_i}{\tilde{l}_i}}{\sum_{j=1}^{M} \frac{\eta_j}{\tilde{l}_j}} \times 10^6 \tag{2}$$

where TPM_i is used as a relative abundance of each transposable element in a given sample.

It is difficult to directly estimate the η and TPM given T and F, so `Salmon` performs the following processes. First, `Salmon` runs a quasi-mapping procedure which is initially proposed in [24]. A quasi-mapping specifies the target of each given read and also determines the position and the orientation of the read concerning the target by computing the Maximum Mappable Prefix (MMP) [28] and Next Informative Position (NIP) [24] of the read. This mapping procedure uses a generalized suffix array[29] and enables a fast and accurate mapping as compared to other mapping tools, such as `Bowtie 2`, `STAR`, and `Kalisto`.[24] The mapping also provides a possible mapping locations for each read.

The maximum-likelihood objective model for a set of reads F is defined as follows:

$$Pr\{F|\eta, Z, T\} = \prod_{j=1}^{N} \sum_{i=1}^{M} Pr\{t_i|\eta\} \cdot Pr\{f_i|t_i, z_{ij} = 1\} \tag{3}$$

where $z_{ij} = 1$ if j-th read in F is derived from i-th TE. Since $Pr\{f_i|t_i, z_{ij} = 1\}$ is unknown, `Salmon` uses the following auxiliary terms to define conditional model to estimate the probability:

$$Pr\{f_j|t_i\} = Pr\{l|t_i\} \cdot Pr\{p|t_i, l\} \cdot Pr\{o|t_i\} \tag{4}$$

where $Pr\{l|t_i\}$ is the probability of drawing a read of the inferred length l given t_i, $Pr\{p|t_i, l\}$ is the probability of the read starting at position p on t_i, $Pr\{o|t_i\}$ is the probability of obtaining a read alignment with the given orientation o to t_i, and this model accounts for sample-specific parameters and biases.

With these probabilistic models, `Salmon` performs online inference to estimate read counts α and nucleotide fraction η using a variant of stochastic collapsed variational Bayesian inference (See Supplementary Algorithm in [20]).[26] In addition to the inference algorithm, `Salmon` constructs equivalence classes for a given F. We assign any pair of reads mapped to same set of target TEs in the same equivalence class. This construction shrink the representation of the sequencing experiment and greatly reduce the running time of offline phase.[20]

Next, `Salmon` starts the offline phase. Given the set of equivalence classes of F, an EM algorithm was used to refine the previous estimation for each equivalence class with following objective function L:

$$L\{\alpha|F, Z, T\} = \prod_{j=1}^{N} \sum_{i=1}^{M} \hat{\eta}_i Pr\{f_j|t_i\} \tag{5}$$

, where $\hat{\eta}_i = \frac{\alpha_i}{\sum_j \alpha_j}$. Once the offline phase is done, `Salmon` outputs the estimation of each TE abundance for F.

2.3. *Statistical tests*

We provide a statistical analysis function to identify differentially expressed TEs from the counts table as the last step of the pipeline. Differential analysis using DESeq2 can handle binary covariates such as binary genotype: phenotype and gender.[30] To handle quantitative covariates such as age, we apply the General Linear Model (GLM).[31] The statistical analysis will produce two statistics to represent associations between the TEs and the covariates: the first one is the test statistics for each TE, and the second one is the summary of the statistics for each clade. The output files are provided with various file formats, such as tab-separated values file (TSV), XML spreadsheet file format (XLS, XLSX), R object file (Rdata), and Portable Document Format (PDF) file.

3. Results

3.1. *Datasets*

Two datasets were used for our comparison to other methods. The first dataset is the RNA-seq data from Gene Expression Omnibus (accession no. GSE47006) which includes wild-type and *Piwi* (P-element Induced WImpy testis) knockdown flies. This dataset was used as a benchmark dataset in the `TEtranscripts` paper as well.[6] We compared the performance in terms of running time and quantification accuracy between our proposed pipeline and other tools, including `TEtranscripts`, `HTSeq-count`, `Cuffdiff` and `RepEnrich`.[14,32–34]

In the second dataset, we seek to identify new TEs that are associated with Amyotrophic Lateral Sclerosis (ALS). We applied our pipeline to a K562 cell-line RNA-seq dataset from ENCODE (Encyclopedia of DNA Elements, `http://encodeproject.org`) Consortium (accession ID: ENCBS555BYH).[35] The dataset consists of two biological replicates of shRNA (short hairpin RNA) knockdown (KD) targeting *TARDBP* (TAR DNA Binding Protein, as known as TDP-43) gene and two biological replicates of controls (a shRNA inserted but targets no genes). It has been reported that loss of *TDP-43* function causes ALS.[7,36] To measure scalability with the dataset. we also ran `TEtranscripts` to compare running time of both methods. We also performed an integrative analysis for highly differentially expressed TEs for further understanding of any new mechanism of ALS.

3.2. *Computational experiment setup*

Generating BAM files from `FASTQ` files are mandatory to `TEtranscripts`, `HTSeq-count`, `Cuffdiff`, and `RepEnrich`, we applied `STAR` [37] to generate the files with the following parameters: `--outFilterMultimapNmax 100` and `--winAnchorMultimapNmax 100`. Sixteen threads were used for both `SalmonTE` and `STAR` . We also used the same parameter setup for each quantification tool similar to the `TEtranscripts` paper.

All of the computational experiments were done in a workstation with `Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz` (10 cores and maximum 40 threads) and `128GBytes` RAM.

3.3. *SalmonTE guarantees a reliable TE expression estimation*

For the quantification accuracy comparison, we first took estimated abundance of 8 TEs from each quantification tool. To validate the results, Reverse Transcription-quantitative Polymerase Chain Reaction (RT-qPCR) was done on these 8 TEs [6]. We observed `SalmonTE` outperformed all other tools ($r^2 = 0.98$, Figure 2 and Table 1). We also found that `SalmonTE` identified a weak down-regulation of DM1731_I and HETA which was missed by `TEtranscripts`.

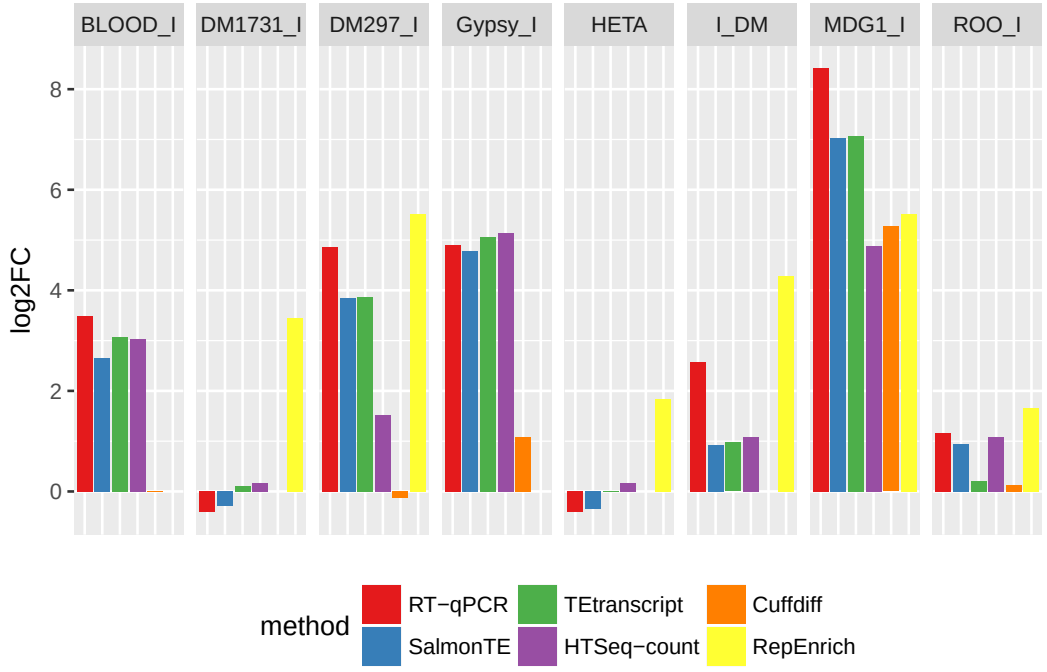


Fig. 2. Comparison of Drosophila TE expression estimation. Four computational methods were compared to `SalmonTE`. RT-qPCR was used to validate the expression levels of the 8 TEs in the Drosophila samples.

Table 1. Pearson Correlation between RT-qPCR and computational TE quantification methods.

Method	`SalmonTE`	`TEtranscripts`	HTSeq-count	Cuffdiff	RepEnrich
r^2	0.98	0.97	0.85	NA	NA

Next, we compared the estimated log_2FC of `SalmonTE` to `TEtranscripts`on each transposable element for a deeper investigation. Our data shows that the estimated TE abundance of both methods are highly correlated ($r^2 = 0.98$), and we also observed there is a high concordance in the direction of fold-changes between `SalmonTE` and `TEtranscripts` (Figure 3). We also measured the correlations of normalized read counts between `SalmonTE` and `TEtranscripts`, and we observed that the calculated read counts from those methods are highly correlated in each sample as well ($r^2 = 0.92$ for wild-type (WT) sample and $r^2 = 0.91$ for

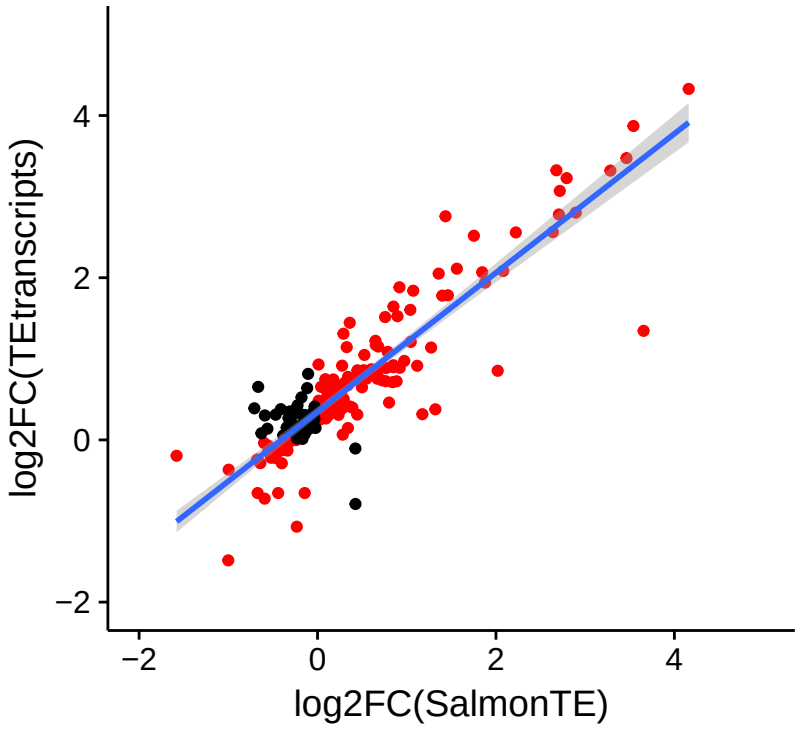


Fig. 3. Correlation of log_2FC ($\frac{Piwi}{WT}$) for each transposable element between `SalmonTE` and `TEtranscripts`. Red points represent TEs with the same fold change direction between `SalmonTE` and `TEtranscripts`.

Piwi KD sample). From this observation, we conclude that both tools generate a similar estimation result. It is not a surprising result because `TEtranscripts` deploys `RSEM` algorithm,[38] and previous studies have demonstrated that transcripts count estimations from `RSEM` and `Salmon` are very correlated.[39,40]

3.4. *SalmonTE shows a better scalability in the speed benchmark dataset*

We measured the speed of `SalmonTE` and`TEtranscripts` on two different datasets (Table 2). Compared to `TEtranscripts`, `SalmonTE` showed a 19x to 27x fold increase in speed. In this analysis, we demonstrate that `SalmonTE` outperformed `TEtranscripts` in processing speed. Our pipeline finishes in less than 5 minutes, while `TEtranscripts` needs about 2 hours to process a single sample. Moreover, our benchmark shows that estimated cost of our pipeline in the cloud computing environment is for the thousands of samples 22 times cheaper than `TEtranscripts` in the computing environment (Table 3).

3.5. *Discover differentially expressed TEs in ALS cell line*

Next, we applied `SalmonTE` pipeline to the *TDP-43* knockdown dataset. We identified 23 transposable elements that are differential expressed between TARDBP knockdown and control cell lines (Table 4) with the threshold of $|log_2FC| \geq 0.5$. No statistical test were performed because the number of replicates in the dataset are small.

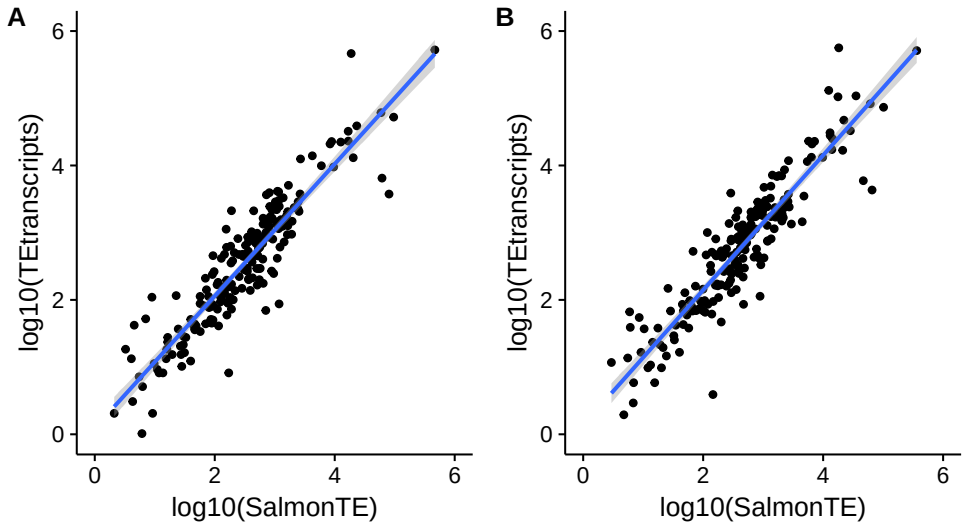


Fig. 4. Sample correlation of count for each transposable element between `SalmonTE` and `TEtranscripts`. **A**. WT sample, **B**. Piwi KD sample.

Table 2. Running speed comparison between `SalmonTE` and `TEtranscripts`.

Dataset	Piwi KD [6]	K562 *TDP-43*
Total number of samples	2	4
RNA-seq file type	Single end	Paired ends
Total number of reads	90,411,467	309,701,182
`SalmonTE` runtime (hh:mm:ss)	0:05:33	0:17:13
`TEtranscripts` runtime (hh:mm:ss)	1:45:26	7:49:40
Speedup	19.00x	27.28x

Table 3. Price estimation of `SalmonTE` and `TEtranscripts` in cloud computing environment (Amazon Elastic Compute Cloud (EC2), and Amazon Elastic Block Store (EBS)). We assume that the size of a `FASTQ` file for a sample is 20GB for the calculations.

Methods	`SalmonTE`	`TEtranscripts`
Estimated using 1000 samples	90 hours	2,000 hours
The price of Amazon EC2 (m4.10xlarge, US Oregon region) [41]	$180	$ 4,000
The price of Amazon EBS (gp2 40TB, US Oregon region) [42]	$500	$ 11,111
Total price	$680	$ 15,111

We can see that most of the differentially expressed features are Endogenous Retrovirus (15 of 23) in *TDP-43* cell-line sample, and we hypothesize that some of the differentially Endogenous Retrovirus TEs are associated with ALS. *TDP-43* is an established and well-studied DNA and RNA binding protein, and could potentially regulate transposable elements at multiple levels.[43] To facilitate a mechanistic understanding of the underlying regulatory mechanism of *TDP-43* and to substantiate the identified differentially expressed transposable, we performed an integrative analysis by combining RNA-seq and *TDP-43* binding data. We obtained DNA binding (ChIP-Seq [44] data) and RNA binding (CLIP-Seq [45] data) datasets

of *TDP-43* in the same K562 cell line from the ENCODE consortium. For illustration, we choose MER74A and AluJo elements that are highly up and down regulated respectively and are also found in Dfam database.[46] We quantified the number of overlapping *TDP-43* ChIP/CLIP peaks with MER74A and AluJo annotations from Dfam. We observed that AluJo element which is down regulated in *TDP-43* knockdown samples is enriched for *TDP-43* ChIP and CLIP peaks as shown in Figure 5, which might indicate that *TDP-43* positively regulate AluJo elements. On the other hand, we did not find any enrichment of *TDP-43* binding for MER74A elements. This preferential binding of *TDP-43* substantiates the differentially expressed transposable elements by our pipeline.

Table 4. 23 Differentially expressed transposable elements in the ENCODE TARDBP data

Name	Clade	log2FC
MER74A	ERV3	1.68
MER57E1	ERV1	1.30
AluYd2	SINE	0.83
LTR1C1	ERV1	0.77
AluSx1	SINE	0.75
LTR27D	ERV1	0.73
AluSx	SINE	0.71
MLT-int	ERV3	0.66
MER54A	ERV3	0.52
MER65D	ERV1	0.51
LTR28	ERV1	-0.59
LTR1F	ERV1	-0.63
FLAM	SINE	-0.64
MER21	ERV3	-0.68
MER101	ERV1	-0.69
LTR26B	ERV1	-0.70
MER83C	ERV1	-0.71
AluJo	SINE	-0.72
LTR06	ERV1	-0.73
MLT2D	ERV3	-0.78
AluYf5	SINE	-0.86
AluYd3	SINE	-1.41
THER2	SINE	-2.03

To identify if there is any general differential expression trend on subfamilies of TEs, we grouped all the TEs based on their clade information. We excluded all of the CR1 (Chicken Repeat 1) since the number of such elements in the clade is small. We found that SINE (Short Interspersed Nuclear Elements) are mostly down expressed, and elements in L1 (Long interspersed nuclear element 1) are generally over expressed in *TDP-43* knockdown samples. This result provides a working hypothesis that knocking-down of *TDP-43* repress the expression of SINE elements and induce the expression of L1 elements.

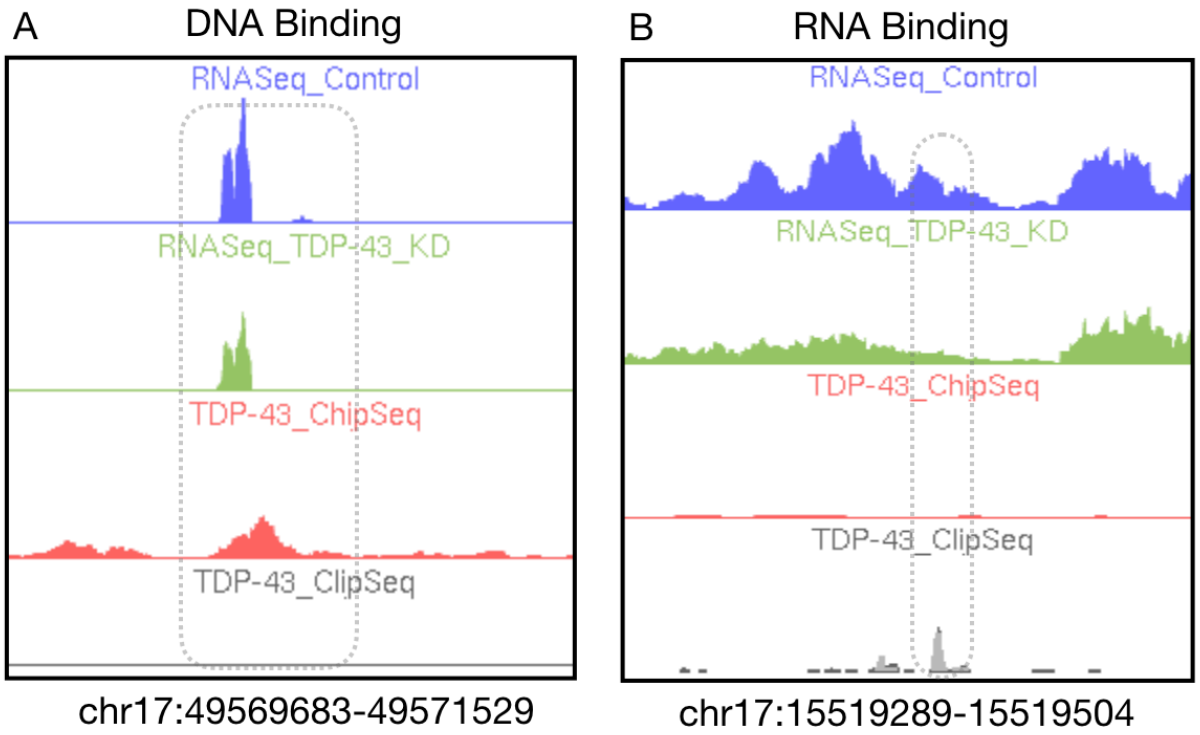


Fig. 5. **A**. Showing down-regulation of AluJo with *TDP-43* ChIP-seq peak, **B**. Showing down-regulation of AluJo with *TDP-43* CLIP-seq peak.

4. Conclusion

In this work, we developed `SalmonTE`, a fast and reliable pipeline for quantification of TEs from NGS data. Our results of `SalmonTE` on the various datasets have shown a large speed-up in computing time relative to `TEtranscripts`, while preserving an accurate quantification on TEs. Therefore, we expect this pipeline will enable the biomedical research community to rapidly quantify and analyze TEs from large amounts of data generated over the past years that are otherwise lost due to genome-masking. Our tool could help the research community to discovery many TE associated with diseases.

There are still several remaining features that could be implemented in the future to improve the usability of `SalmonTE`. For example, prediction of genomic locations, which contain the differentially expressed TEs, is useful in many TE studies. Several methods were developed toward this end,[15,47] but these tools share the scalability issue and require massive computing power for a large-scale TE study. Moreover, alignment free algorithms are intrinsically limited to addressing this question. Therefore, we foresee a novel algorithm which extends and improves the current alignment-free methods.

Acknowledgments

This work has been supported by National Institute of General Medical Sciences R01-GM120033, National Science Foundation - Division of Mathematical Sciences DMS-1263932,

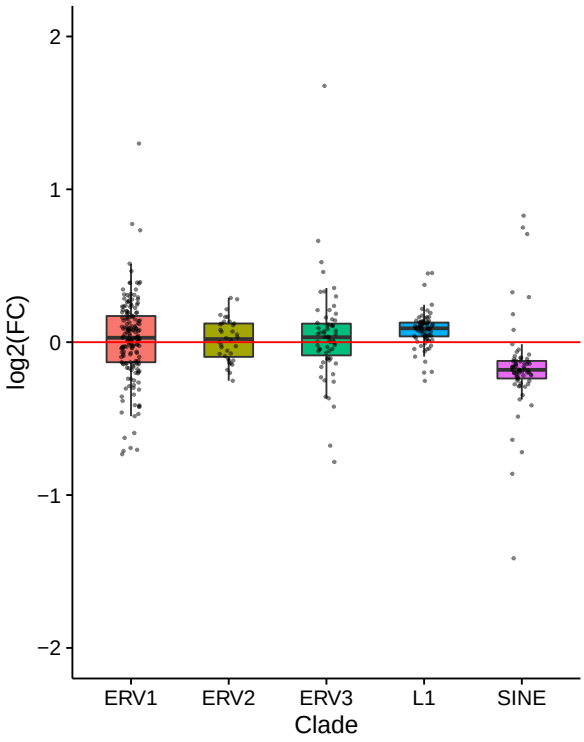


Fig. 6. A boxplot of log_2FC for each clade in the ENCODE *TDP-43* data

Cancer Prevention Research Institute of Texas RP170387, Houston Endowment (Z.L.), and the Alzheimer's Association (J.M.S.). We thank Kala Pham and Rami Al-Ouran for comments that greatly improved this manuscript.

References

1. J. A. Erwin, M. C. Marchetto and F. H. Gage, *Nature Reviews Neuroscience* **15**, 497 (2014).
2. C. Biémont and C. Vieira, *Nature* **443**, 521 (2006).
3. V. P. Belancio, D. J. Hedges and P. Deininger, *Genome research* **18**, 343 (2008).
4. R. L. Jirtle and M. K. Skinner, *Nature reviews. Genetics* **8**, p. 253 (2007).
5. J. G. Wood and S. L. Helfand, *Frontiers in genetics* **4** (2013).
6. H. Ohtani, Y. W. Iwasaki, A. Shibuya, H. Siomi, M. C. Siomi and K. Saito, *Genes & development* **27**, 1656 (2013).
7. S. P. Mihevc, M. Baralle, E. Buratti and B. Rogelj, *Scientific reports* **6**, p. 33996 (2016).
8. W. Li, Y. Jin, L. Prazak, M. Hammell and J. Dubnau, *PloS one* **7**, p. e44099 (2012).
9. L. Krug, N. Chatterjee, R. Borges-Monroy, S. Hearn, W.-W. Liao, K. Morrill, L. Prazak, N. Rozhkov, D. Theodorou, M. Hammell *et al.*, *PLoS genetics* **13**, p. e1006635 (2017).
10. E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson *et al.*, *Science* **337**, 967 (2012).
11. A. Platzer, V. Nizhynska and Q. Long, *Biology* **1**, 395 (2012).
12. E. Helman, M. S. Lawrence, C. Stewart, C. Sougnez, G. Getz and M. Meyerson, *Genome research* **24**, 1053 (2014).
13. E. Hénaff, L. Zapata, J. M. Casacuberta and S. Ossowski, *BMC genomics* **16**, p. 768 (2015).
14. Y. Jin, O. H. Tam, E. Paniagua and M. Hammell, *Bioinformatics* **31**, 3593 (2015).

15. J. R. de Ruiter, S. M. Kas, E. Schut, D. J. Adams, M. J. Koudijs, L. F. Wessels and J. Jonkers, *Nucleic Acids Research* (2017).
16. Z. Tang, J. P. Steranka, S. Ma, M. Grivainis, N. Rodić, C. R. L. Huang, I.-M. Shih, T.-L. Wang, J. D. Boeke, D. Fenyö *et al.*, *Proceedings of the National Academy of Sciences* **114**, E733 (2017).
17. A. D. Ewing, *Mobile DNA* **6**, p. 24 (2015).
18. H. Samet, *The design and analysis of spatial data structures* (Addison-Wesley Reading, MA, 1990).
19. A. Nellore, A. E. Jaffe, J.-P. Fortin, J. Alquicira-Hernández, L. Collado-Torres, S. Wang, R. A. Phillips III, N. Karbhari, K. D. Hansen, B. Langmead *et al.*, *Genome biology* **17**, p. 266 (2016).
20. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford, *Nature Methods* **14**, 417 (2017).
21. J. Köster and S. Rahmann, *Bioinformatics* **28**, 2520 (2012).
22. Amazon, Amazon ebs pricing (2017), `http://www.repeatmasker.org/`.
23. G. I. R. Institute, Repbase (2017), `http://www.girinst.org/repbase/update/browse.php`.
24. A. Srivastava, H. Sarkar, N. Gupta and R. Patro, *Bioinformatics* **32**, i192 (2016).
25. C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).
26. J. Foulds, L. Boyles, C. DuBois, P. Smyth and M. Welling, Stochastic collapsed variational bayesian inference for latent dirichlet allocation, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.
27. B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson and C. N. Dewey, *Bioinformatics* **26**, 493 (2009).
28. H. Li, *Bioinformatics* **28**, 1838 (2012).
29. U. Manber and G. Myers, *siam Journal on Computing* **22**, 935 (1993).
30. M. I. Love, W. Huber and S. Anders, *Genome biology* **15**, p. 550 (2014).
31. R. Johnston, *Multivariate statistical analysis in geography; a primer on the general linear model*, tech. rep. (1980).
32. S. Anders, P. T. Pyl and W. Huber, *Bioinformatics* **31**, 166 (2015).
33. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter, *Nature biotechnology* **31**, 46 (2013).
34. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy and N. Neretti, *BMC genomics* **15**, p. 583 (2014).
35. E. P. Consortium *et al.*, *Nature* **489**, p. 57 (2012).
36. C. Yang, H. Wang, T. Qiao, B. Yang, L. Aliaga, L. Qiu, W. Tan, J. Salameh, D. M. McKenna-Yasek, T. Smith *et al.*, *Proceedings of the National Academy of Sciences* **111**, E1121 (2014).
37. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras, *Bioinformatics* **29**, 15 (2013).
38. B. Li and C. N. Dewey, *BMC bioinformatics* **12**, p. 323 (2011).
39. H. Jin, Y.-W. Wan and Z. Liu, *BMC bioinformatics* **18**, p. 117 (2017).
40. C. Zhang, B. Zhang, L.-L. Lin and S. Zhao, *BMC genomics* **18**, p. 583 (2017).
41. Amazon, Amazon ec2 instance pricing (2017), `https://aws.amazon.com/ec2/pricing/on-demand/`.
42. Amazon, Amazon ebs pricing (2017), `https://aws.amazon.com/ebs/pricing/`.
43. Q. Tan, H. K. Yalamanchili, J. Park, A. De Maio, H.-C. Lu, Y.-W. Wan, J. J. White, V. V. Bondar, L. S. Sayegh, X. Liu *et al.*, *Human molecular genetics* **25**, 5083 (2016).
44. D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, *Science* **316**, 1497 (2007).
45. R. B. Darnell, *Wiley Interdisciplinary Reviews: RNA* **1**, 266 (2010).
46. R. Hubley, R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. Smit and T. J. Wheeler, *Nucleic acids research* **44**, D81 (2015).
47. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy and N. Neretti, *BMC genomics* **15**, p. 583 (2014).

Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula

Kipp W. Johnson[1], Benjamin S. Glicksberg[1], Rachel A. Hodos[1,2], Khader Shameer[1], and Joel T. Dudley[1†]

1. Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10065

2. Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012

Corresponding author email: joel.dudley@mssm.edu

Hypertension is a major risk factor for ischemic cardiovascular disease and cerebrovascular disease, which are respectively the primary and secondary most common causes of morbidity and mortality across the globe. To alleviate the risks of hypertension, there are a number of effective antihypertensive drugs available. However, the optimal treatment blood pressure goal for antihypertensive therapy remains an area of controversy. The results of the recent Systolic Blood Pressure Intervention Trial (SPRINT) trial, which found benefits for intensive lowering of systolic blood pressure, have been debated for several reasons. We aimed to assess the benefits of treating to four different blood pressure targets and to compare our results to those of SPRINT using a method for causal inference called the parametric g formula. We applied this method to blood pressure measurements obtained from the electronic health records of approximately 200,000 patients who visited the Mount Sinai Hospital in New York, NY. We simulated the effect of four clinically relevant dynamic treatment regimes, assessing the effectiveness of treating to four different blood pressure targets: 150 mmHg, 140 mmHg, 130 mmHg, and 120 mmHg. In contrast to current American Heart Association guidelines and in concordance with SPRINT, we find that targeting 120 mmHg systolic blood pressure is significantly associated with decreased incidence of major adverse cardiovascular events. Causal inference methods applied to electronic methods are a powerful and flexible technique and medicine may benefit from their increased usage.

Keywords: Causal inference; blood pressure; electronic health records; parametric g formula; preventative medicine.

Acknowledgements: We gratefully acknowledge Yi Zhang (yz@mttpi.org) for clarification of the parametric G formula and her previous work on the method over email correspondence.

1. Introduction

1.1. *Global Burden of Hypertension*

Ischemic cardiovascular disease and cerebrovascular disease are the primary and secondary causes of global disease burden respectively, both in the United States and the rest of the world (1). Hypertension (HTN), also known as elevated blood pressure (BP), is a primary risk factor for both diseases (2). Additional evidence continues to accumulate for the role of HTN as a central risk factor for a wide variety of chronic diseases such as dementia (3) and type 2 diabetes mellitus (4). As both developed and developing countries continue to experience greater chronic disease burden, treatment and prevention of HTN is one of the most important issues in medicine.

1.2. *Challenges in Previous Efforts to Discover Optimal Target Blood Pressures*

There are up to 69 different drugs from 15 different classes available to manage HTN, demonstrating the widely understood importance of managing this condition. Even with this plethora of available medications, current estimates indicate that up to 65% of patients with HTN have difficulty controlling their blood pressure (5). One issue confounding this matter is a lack of consensus on the appropriate target blood pressure. In 2014, the Eighth Joint National Committee on Hypertension (JNC 8) released the latest guidelines for hypertension treatment which recommend that patients over 60 years old be medicated to a target BP of 150 mmHg systolic (SBP) and 90 mmHg diastolic (DBP) (6). Additionally, they recommended that patients 30-59 years old be medicated to a DBP target of 90 mmHg. Notably, the JNC 8 could not establish an evidence-based SBP goal for individuals within this age range. In 2015, investigators for the Systolic Blood Pressure Intervention Trial (SPRINT) released highly anticipated findings in the New England Journal of Medicine for the benefit of intensive HTN treatment measured against the standard hypertension treatment regime (7). The SPRINT investigators defined intensive therapy as medical therapy intended to reduce BP to 120/80 mmHg or below, instead of the standard goal therapy of 140/90 mmHg. This large trial recruited 9,361 patients from 102 different clinical sites (from 5 clinical networks) across the country and followed them for adverse outcomes for a median of 3.26 years.

SPRINT was intended to provide a definitive answer for the benefits of intensive antihypertensive therapy. The trial results seemed to demonstrate that intensive antihypertensive therapy was strongly associated with lowered risk for the study's primary study endpoint of myocardial infarction, acute coronary syndrome, stroke, congestive heart failure, or cardiovascular death ($HR=0.75$, $p<0.001$) (7). However, this conclusion provoked a firestorm of controversy in the cardiovascular medicine community. Blood pressure is a difficult phenotype to measure, since readings vary from minute to minute, many patients suffer from "white-coat hypertension" in the presence of a physician, and measurements are usually taken with a manual sphygmomanometer and stethoscope (8,9). In contrast, the SPRINT trial broke with decades of precedence by using an automated electronic BP measurement device to capture a series of six measurements (8,9). Three

measurements were spaced one minute apart in the presence of a researcher and three more were recorded outside of the presence of the researcher after a five-minute break (10,11). It has been suggested that this difference in BP measurement technique may make their results impossible to apply to the clinic (10). This is because the SPRINT blood pressure measurement method is known to produce systematically different blood pressure readings when compared to the standard technique used in many hospitals and in previous trials (9). Typical systolic blood pressure measurements are 14 mmHg lower using the SPRINT method compared to the standard method (9). For example, this implies a SPRINT target of 120mmHg may be equivalent to a real-world target of 134mmHg. Thus, the external validity of the SPRINT findings is unclear–SPRINT targets result from BP measurements that may not be comparable to normal BP measurements made outside of the clinical trial's setting. Due to this controversy, we believe that additional complementary evidence supporting aggressive antihypertensive treatment could provide insight on treatment decisions and outcomes.

1.3. *Causal Inference from Electronic Health Records As a Tool to Answer Difficult Clinical Questions*

Electronic health records (EHR) present an excellent potential data source to analyze and determine optimal blood pressure treatment goal. EHR contain longitudinal information captured during the routine care process such as visit dates, patient demographics such as age; sex; self-declared race/ethnicity; medication prescription orders, disease and procedure billing codes, and most importantly, blood pressure measurements. Of note, the BP measurements contained within the EHR will reflect what is routinely measured clinically, instead of BP as is measured within a contained clinical trial setting such as SPRINT.

One challenge in the use of EHR for BP analysis is the clinical situation in which hypertension is managed. Hypertension is a chronic disease in which BP measurements are longitudinally manipulated by the administration of a variety of drugs over an extended period of time. This is in contrast to a clinical trial, where other exogenous factors are explicitly modeled and a drug may be consistently administered throughout the study. Confounding post-baseline time-varying dependent relationships such as these cannot be modeled using conventional statistical methods such as regression or survival analysis without many potentially unrealistic assumptions (11). While observational analyses are restricted to a framework where one can only test interventions that have been explicitly carried out in the data, the *g-formula* approach enables us to simulate dynamic treatment strategies and estimate their effects, even if those strategies have not been fully carried out in the data used to construct the model (12). *G methods* may be used to estimate the effect of different interventions on an outcome in the presence of time-varying confounders. For example, an extension of g methods called the *parametric G formula* has been used to measure the effect of different treatment regimes for highly active retroviral therapy (HAART) in AIDS (13); to estimate the effect of different governmental policies on radon and lung cancer (14); and to decide upon optimal anemia management strategies (15). However, these causal inference methods have never been applied to real, hospital-derived EHR. Here, we use the parametric g

formula to model the effect of different BP treatment targets on major adverse cardiovascular outcomes (MACE).

2. Methods

2.1. *Data Acquisition from the Mount Sinai Hospital EHR*

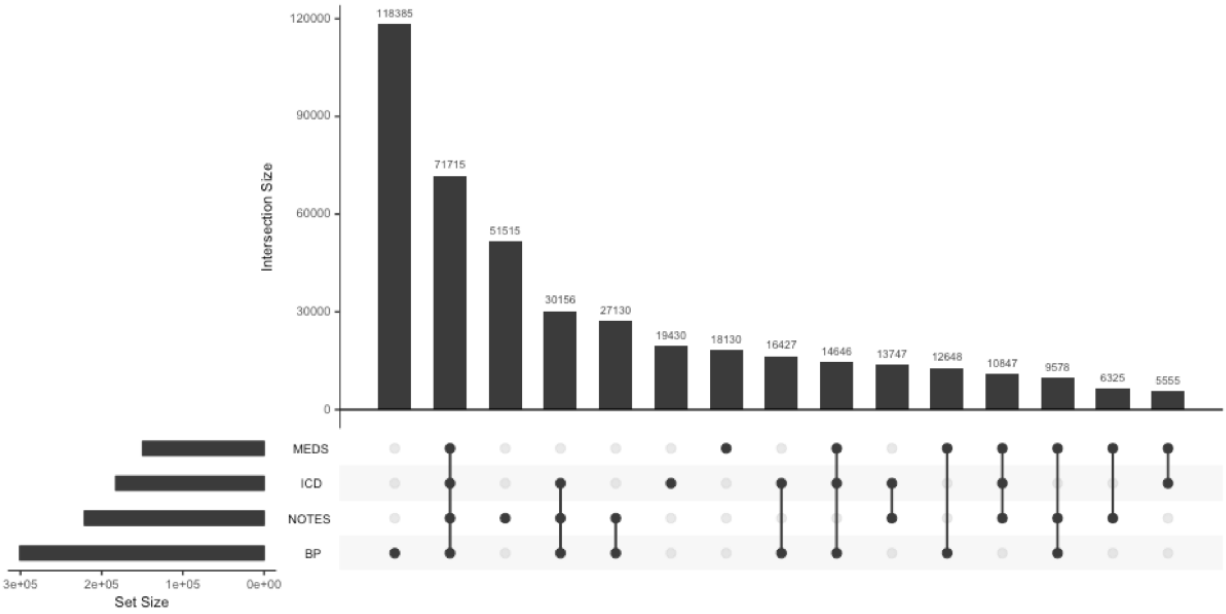


Figure 1: Hypertension phenotyping algorithm patient counts

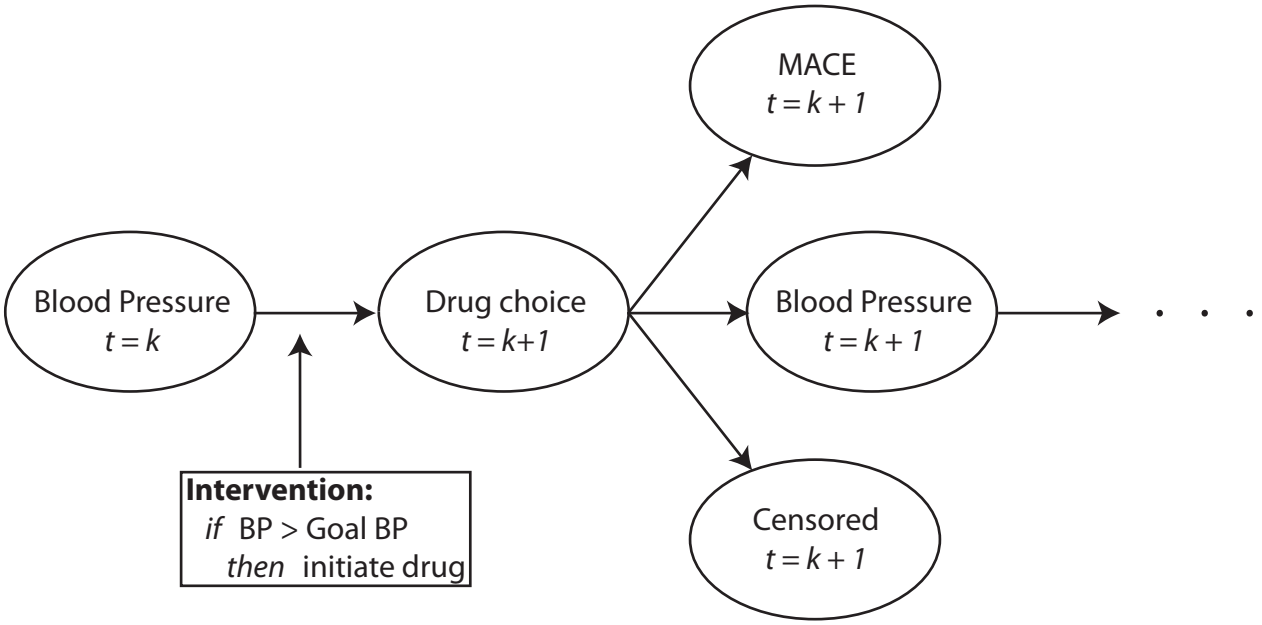
The Mount Sinai Hospital (MSH) is a tertiary-care urban hospital located on the Upper East Side of Manhattan in New York City. MSH's EMR contains longitudinal information for more than two million patients (800,00 with at least 1 prescription) collected from 2008-2016. We identified hypertensive patients using a phenotyping algorithm combining International Classification of Disease (ICD-9) billing codes, blood pressure measurements, natural language processed physician notes, and the prescription of antihypertensive medications (Figure 1). Phenotyped patients must have had at least two elevated blood pressure measurements (SBP>140 or DBP>90) and two instances (each on different days) of one of the following: hypertensive medications, ICD hypertension billing codes, or mentions of hypertension in the physician notes. Medications in the EMR were normalized to Anatomical Therapeutic Chemical (ATC) Classification System drug classes using RxNorm (16). Antihypertensive medications were defined as those belonging to ATC classes C02 (*Antihypertensives*), C03 (*Diuretics*), C04 (*Peripheral vasodilators*), C07 (*Beta blocking agents*), C08 (*Calcium channel blockers*), and C09 (*Agents acting on the renin-angiotensin system*) and then further filtered by intersection with JNC 8 recommended antihypertensive medications. Essential hypertension billing codes were identified as those starting with 401.xx. We assessed cardiovascular adverse outcomes with the commonly used major adverse cardiovascular event (MACE) composite endpoint, which is often composed of

myocardial infarction, stroke, and heart failure events. We identified patients with a MACE using the corresponding ICD9 codes 410.xx, 411.xx, 428.xx, 433.xx, 434.xx, and 436.xx.

2.2. *Problem setup*

We evaluated the effect of treating patients to four different SBP targets: 150 mmHg, 140 mmHg, 130 mmHg, and 120 mmHg. We chose these four targets because they have each been suggested as SBP blood pressure targets. We assume the causal pathway demonstrated by the directed acyclic graph in Figure 2, where, blood pressure measurements are related to drug choice selection as well as MACE outcomes. Additionally, at each time step an individual may be censored (i.e., lost from the EHR). Finally, antihypertensive medications act to reduce blood pressure measurements subsequent to their administration.

Figure 2: Directed acyclic graph for relationship between blood pressure measurement, drug administration, MACE, censoring, and modeled intervention policy



2.3. *Parametric g formula*

The g formula directly models probabilities for a given outcome conditional upon covariates and exposures. For real-world datasets, modeling all conditional probabilities directly is not feasible, especially in the presence of continuous covariates such as BP. The parametric g formula is an extension of the g formula where parametric models are used to model probabilities instead of direct calculations. Parametric methods have the advantage of being computationally feasible, able to handle continuous covariates, and more efficient given the data. Analysis using the parametric g formula requires a three-step algorithm (Figure 3).

1. Model Conditional Probabilities

First, we model effect sizes and conditional probabilities for all person-times for (a) all covariates and (b) outcomes in our dataset using Bayesian logistic regression. Bayesian prior regression models are implemented in the R package *arm* (17). Models are restricted to those who survive and remain uncensored to time t = k + 1 and include the effects of baseline covariates as well as time-varying covariates at time k.

2. Monte Carlo Simulation

Second, we perform Monte Carlo simulation for 10,000 individuals for each treatment target using the probabilities from step 1, intervening as required with drug treatment if BP exceeds our target BP. Each individual's baseline covariates are sampled from the dataset, and time-varying covariates are simulated using the predicted probability as the conditional mean in a random draw from the Bernoulli distribution in the case of binary variables. In the case of BP, we used the predicted blood pressure as the mean for a normal distribution with standard deviation equal to the mean standard deviation for individuals with matched baseline covariates.

3. Risk computation under different treatment goals

We fit Cox proportional hazards models to evaluate the relative efficacy to the results from each of the different simulated treatment policies and estimate their efficacy.

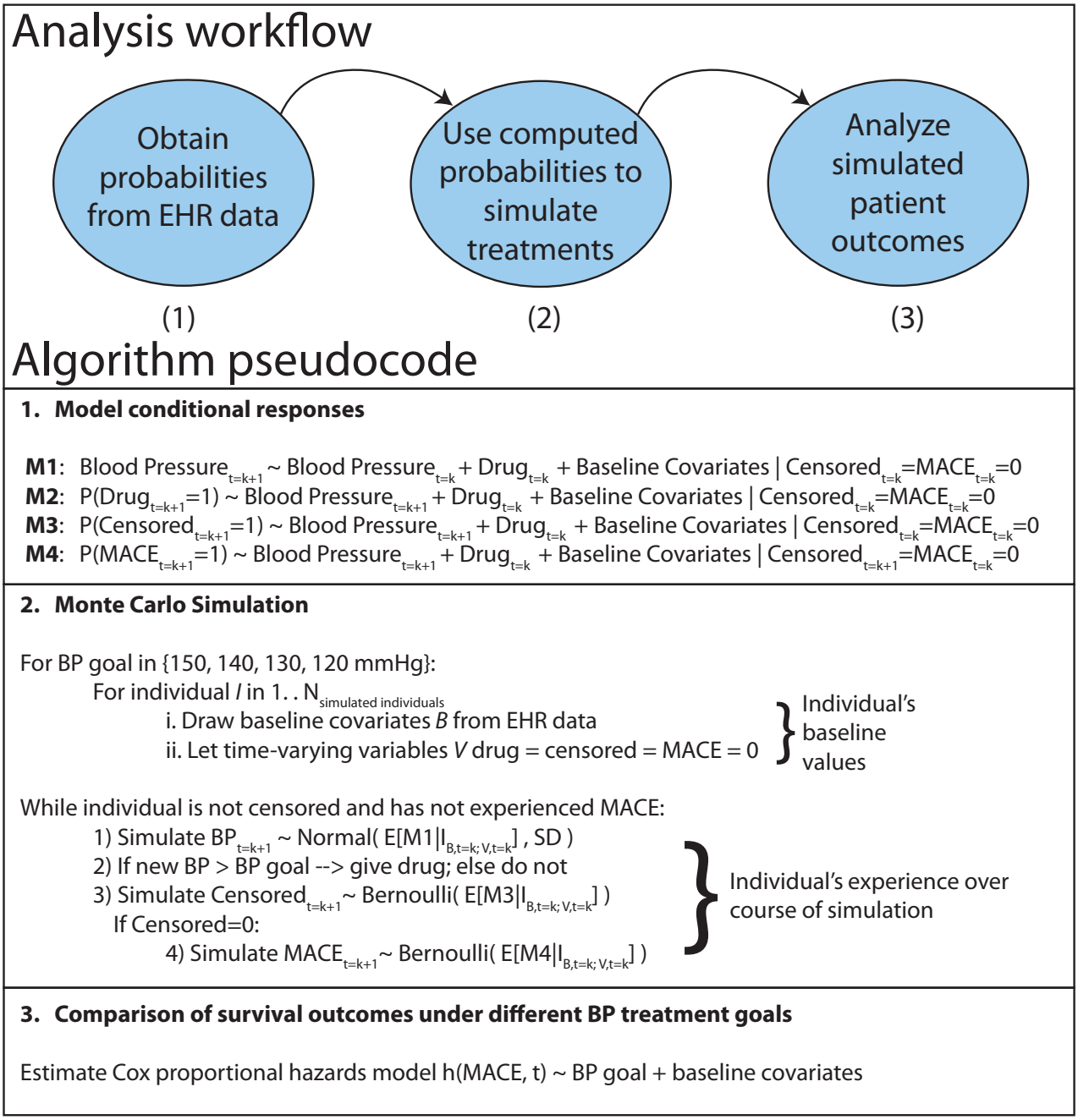


Figure 3: Explanation of parametric g formula analysis for BP target

3. Results

3.1. *Electronic Health Records Data*

We obtained health records from the Mount Sinai Hospital for 218,221 different hypertensive patients. 104,514 individuals were female, 98,048 individuals were male, and 4 individuals were of indeterminate sex and dropped from the analysis due to insufficient power. The remaining individuals did not have sex coded. 81,054 individuals were Caucasian, 38,134 were African

American; 37,008 listed "Other;" 24,435 were Unknown; 12,108 were Hispanic/Latino; 7329 were Asian; and the rest were composed of known but rare ethnicities (Native American, Pacific Islander, etc., – *sample counts not shown to preserve patient confidentiality*) and subsequently combined into "Other." These individuals were collectively prescribed medications 3,678,597 times (16.9 medication prescriptions per person). The EHR contained 31,088,598 measurements of SBP and 31,039,040 measurements of DBP, although the count of BP measurements per individual was significantly right-skewed likely due to frequent measurements of hospitalized critical-care unit patients.

3.2. *Survival time by goal blood pressure target*

We simulated 10,000 patients for each BP target. The 40,000 simulated patients experienced a total of 14,501 major adverse cardiovascular events. The number of MACE was highest in the 150 mmHg target group (3853 events), followed by 140 mmHg group (3722 events), followed by 130 mmHg group (3559 events), followed finally by the 120 mmHg target group (3367 events). Median MACE-free survival times were similarly ordered: 150 mmHg, 31 encounters (95% CI: 29-33); 140 mmHg, 33 encounters (95% CI: 31-34); 130 mmHg, 34 encounters (95% CI: 33-35); 120 mmHg, 35 encounters (95% CI: 34-37).

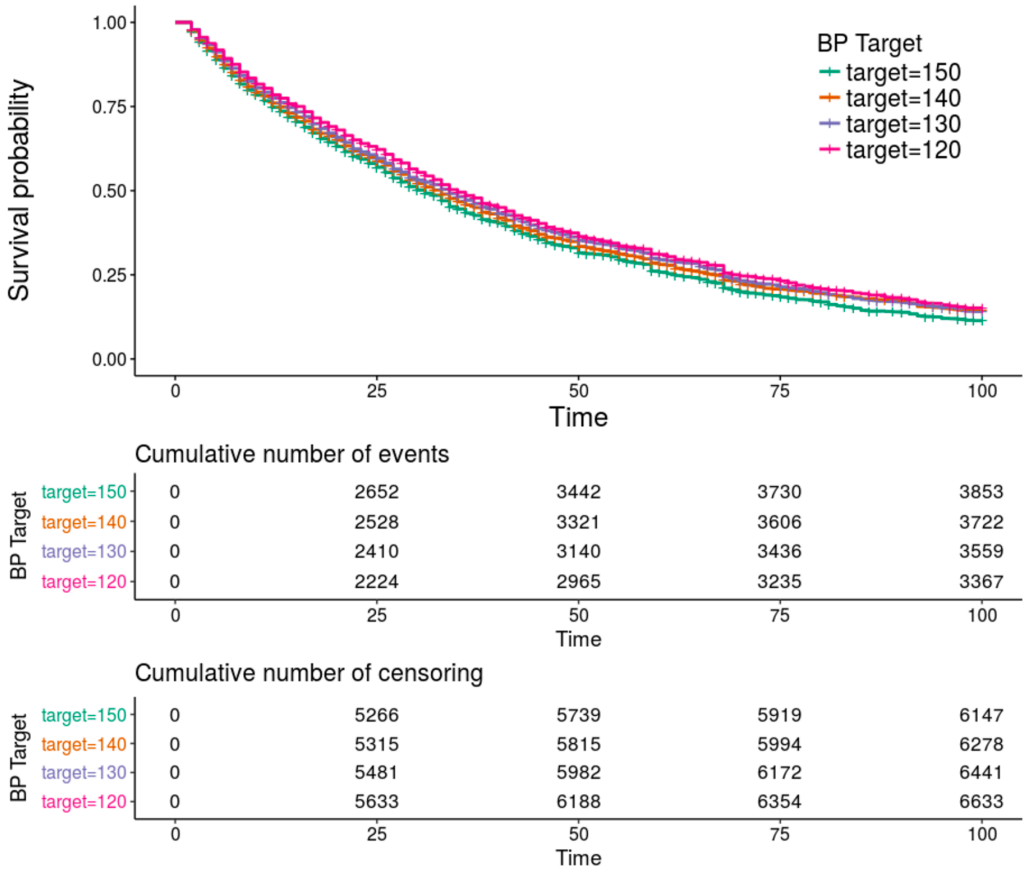
Survival was significantly associated with BP target (Figure 4), with survival highest at the lowest target goal (120mmHg) and less at 130 mmHg and 140 mmHg. Setting the reference target BP as 150 mmHg, we found hazard ratios of 0.934 (Target goal of 140 mmHg, p=0.003); 0.895 (Target goal of 130 mmHg, $p=1.94\text{x}10^{-6}$); and 0.846 (Target goal of 120 mmHg, $p=1.46\text{x}10^{-12}$). Male sex was significantly associated with elevated hazard compared to females (HR=1.27, $p<10^{-15}$). Compared to Caucasian individuals, those of self-declared Hispanic/Latino and Unkown ancestry were more likely to have a MACE (HR=1.364, $p<2\text{x}10^{-16}$ and HR=1.134, p=0.0003 respectively). Interestingly, those of Native American or African American ancestry actually had a slightly lower hazard ratio for MACE outcomes compared to Caucasian individuals (HR=0.932, p=0.023 and HR=0.922, p=0.016 respectively). Those of Asian or Other ancestry did not have a significantly different hazard ratio than Caucasian ancestry individuals (p=0.33 and p=0.279, respectively).

Table 1: Results of Survival Analysis

Covariate	*Reference Level*	*Hazard Ratio*	*H.R. 95% CI*	*P-Value*
140 mmHg Target	150 mmHg Target	0.934	(0.8928, 0.9770)	0.002975
130 mmHg Target	150 mmHg Target	0.895	(0.8553, 0.9370)	$1.94\text{x}10^{-6}$
120 mmHg Target	150 mmHg Target	0.846	(0.8079, 0.8862)	$1.46\text{x}10^{-12}$
Male	Female	1.270	(1.2294, 1.3124)	$<2\text{x}10^{-16}$

African American	Caucasian	0.922	(0.8631, 0.9849)	0.015937
Asian	Caucasian	0.970	(0.9101, 1.0326)	0.334363
Hispanic/Latino	Caucasian	1.364	(1.2773, 1.4566)	$<2x10^{-16}$
Native American	Caucasian	0.932	(0.8775, 0.9904)	0.023023
Other Race	Caucasian	0.934	(0.8394, 1.0518)	0.278997
Unknown Race	Caucasian	1.134	1.0584 1.2138	0.000341

Figure 4: Kaplan-Meier survival plot for observed survival times stratified by target blood pressure. Survival was greatest at a target blood pressure of 120 mmHg, second best at 130 mmHg, third best at 140 mmHg, and worst at a BP target of 150 mmHg. The overall P value for the model was $<10^{-16}$.



Interestingly, examination of the blood pressures of survivors revealed that the distribution of survivors' blood pressures tended to shrink as time advanced. This can be explained by the fact

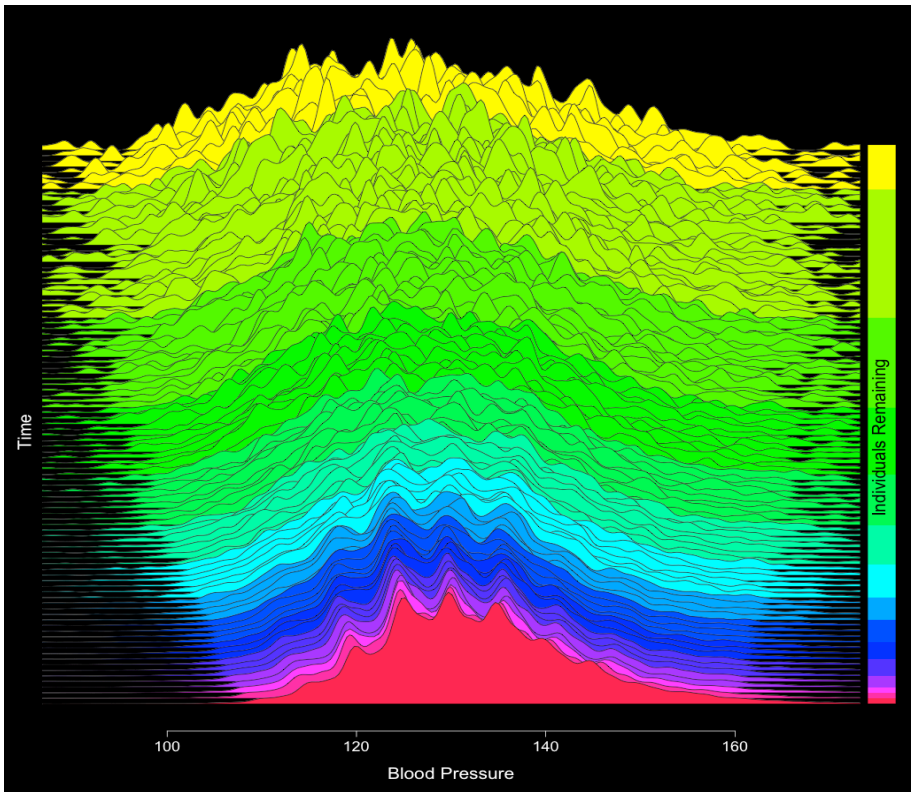


Figure 5: Plots of overall conditional predicted blood pressures over time (Beginning time at top of plot, final predictive models at the bottom of the plot). Distribution of conditional blood pressure

that those with very high or very low blood pressures tended to either have their blood pressure managed toward a goal target or else experienced MACE (Figure 5).

4. Conclusion

In this study, we applied an established causal inference technique to a large EHR database to simulate the effects of four different antihypertensive treatment regimes on MACE outcomes. Our findings are concordant with the recent conclusions from SPRINT, where a target systolic blood pressure goal of 120 mmHg was found to dramatically lower the incidence of adverse cardiovascular outcomes. Our estimated hazard ratio of 0.85 for 120 mm target compared to 150 mmHg compares to the SPRINT hazard ratio estimate of 0.75 for the same comparison, although our estimated hazard ratio is less optimistic than the SPRINT hazard ratio.

This contrasts with the current recommendations from the American Heart Association and the American Association of Family Physicians, who both do not recommend targeting 120 mmHg. Interestingly, BP measurements obtained at the Mount Sinai Hospital are generally measured

using a manual sphygmomanometer and stethoscope, instead of with an automated device as in the SPRINT trial. Despite this difference, we arrived at similar conclusions for the efficacy of intensive antihypertensive therapy. This may suggest that SPRINT BP estimates are not as discordant with past literature as is believed. Certainly, as hypertension is a vitally important topic, we believe our finding points out the need for additional studies. Future studies for BP goal measurements will likely benefit greatly from the inclusion of more personalized drug treatments strategies and a more “precision medicine” approach.

There are several limitations to our study. First, the typical definition of major adverse cardiovascular events (MACE) often includes death from a cardiovascular etiology. Due to the limitations of EHR, we were not able to include death as a MACE outcome in our study. This limitation may be partially mitigated by the fact that many of those who die from a cardiovascular etiology would have one of the MACE ICD-9 billing codes preceding their death. Second, our treatment dynamic algorithm currently does not yet include changes in drug dosages, which is one technique physicians can use to adjust the degree of antihypertensive medication efficacy. Similarly, we do not consider patient BMIs when modeling drug efficacy per patient, which is related to drug dosages. Finally, as with all observational studies, there may be unmeasured confounders which may influence our results to some degree. To this end, we are comforted to some extent by the fact that our study conclusions more closely resembled those from a randomized clinical trial results than those from other observational analyses.

To the best of our knowledge, the e results represent a first application of the parametric g formula to a hospital-based electronic health record system, and almost certainly the first application of such a model in the context of cardiovascular preventative medicine. Causal inference methods allow for the retrospective analysis of treatment regimes which could not easily be performed, or in some cases would not even be ethical to be performed in a randomized clinical trial setting. Taken altogether, we believe our study (1) demonstrates the utility of the parametric g-formula for analysis of treatment interventions in EHR data; and (2) presents complementary and concordant evidence to the SPRINT trial in support of intensive antihypertensive therapy.

References

1. Mortality GBD, Causes of Death C. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388:1459-1544.
2. Blacher J, Levy BI, Mourad JJ, Safar ME, Bakris G. From epidemiological transition to modern cardiovascular epidemiology: hypertension in the 21st century. Lancet 2016;388:530-2.
3. Livingston G, Sommerlad A, Orgeta V et al. Dementia prevention, intervention, and care. Lancet 2017.
4. Kim MJ, Lim NK, Choi SJ, Park HY. Hypertension is an independent risk factor for type 2 diabetes: the Korean genome and epidemiology study. Hypertens Res 2015;38:783-9.
5. Oparil S, Schmieder RE. New approaches in the treatment of hypertension. Circ Res 2015;116:1074-95.
6. James PA, Oparil S, Carter BL et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA 2014;311:507-20.
7. Group SR, Wright JT, Jr., Williamson JD et al. A Randomized Trial of Intensive versus Standard Blood-Pressure Control. N Engl J Med 2015;373:2103-16.
8. Cuspidi C, Sala C, Grassi G, Mancia G. White Coat Hypertension: to Treat or Not to Treat? Curr Hypertens Rep 2016;18:80.
9. Flack JM. Method of Blood Pressure Measurement, Interpretation of SPRINT, and the Atlantic Divide. Curr Hypertens Rep 2017;19:19.
10. Bakris GL. The Implications of Blood Pressure Measurement Methods on Treatment Targets for Blood Pressure. Circulation 2016;134:904-5.
11. Myers MG, Cloutier L, Gelfer M, Padwal RS, Kaczorowski J. Blood Pressure Measurement in the Post-SPRINT Era: A Canadian Perspective. Hypertension 2016;68:e1-3.
12. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiology 2014;25:889-97.
13. Edwards JK, McGrath LJ, Buckley JP, Schubauer-Berigan MK, Cole SR, Richardson DB. Occupational radon exposure and lung cancer mortality: estimating intervention effects using the parametric g-formula. Epidemiology 2014;25:829-34.
14. Westreich D, Cole SR, Young JG et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med 2012;31:2000-9.
15. Zhang Y, Young JG, Thamer M, Hernan MA. Comparing the Effectiveness of Dynamic Treatment Strategies Using Electronic Health Records: An Application of the Parametric g-Formula to Anemia Management Strategies. Health Serv Res 2017.
16. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011;18:441-8.
17. Su AGaY-S. arm: Data Analysis Using Regression and Multilevel/Hierarchical. Models. R package version 1.9-3, 2016.

Data-driven advice for applying machine learning to bioinformatics problems

Randal S. Olson[†*], William La Cava[†*], Zairah Mustahsan, Akshay Varik, and Jason H. Moore[†]

Institute for Biomedical Informatics, University of Pennsylvania
Philadelphia, PA 19104, USA
[†]*E-mails: rso@randalolson.com, lacava@upenn.edu and jhmoore@upenn.edu*

As the bioinformatics field grows, it must keep pace not only with new data but with new algorithms. Here we contribute a thorough analysis of 13 state-of-the-art, commonly used machine learning algorithms on a set of 165 publicly available classification problems in order to provide data-driven algorithm recommendations to current researchers. We present a number of statistical and visual comparisons of algorithm performance and quantify the effect of model selection and algorithm tuning for each algorithm and dataset. The analysis culminates in the recommendation of five algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems.

Keywords: machine learning; data science; best practices; benchmarking; bioinformatics

1. Introduction

The bioinformatics field is increasingly relying on machine learning (ML) algorithms to conduct predictive analytics and gain greater insights into the complex biological processes of the human body.[1] For example, ML algorithms have been applied to great success in GWAS, and have proven effective at detecting patterns of epistasis within the human genome.[2] Recently, deep learning algorithms were used to detect cancer metastases on high-resolution pathology images[3] at levels comparable to human pathologists. These results, among others, indicate heavy interest in ML development and analysis for bioinformatics applications.

Owing to the development of open source ML packages and active research in the ML field, researchers can easily choose from dozens of ML algorithm implementations to build predictive models of complex data. Although having several readily-available ML algorithm implementations is advantageous to bioinformatics researchers seeking to move beyond simple statistics, many researchers experience "choice overload" and find difficulty in selecting the right ML algorithm for their problem at hand. As a result, some ML-oriented bioinformatics projects could be improved simply through the use of a better ML algorithm.

ML researchers are aware of the challenges that algorithm selection presents to ML practitioners. As a result, there have been some efforts to empirically assesses different algorithms across sets of problems, beginning in the mid 1990s with the StatLog project.[4] Early work in this field also emphasized bioinformatics applications.[5] More recently, Caruana *et al.*[6] and

[*] Contributed equally

Fernández-Delgado *et al.*[7] analyzed several supervised learning algorithms, coupled with some parameter tuning. The aforementioned literature often compared many algorithms but on relatively few example problems (between 4 and 12), with only [7] using upwards of 112 example problems. In the time since these assessments, researchers have moved towards standardized, open source implementations of ML algorithms (e.g. scikit-learn[8] and Weka[9]), and the number of publicly available datasets that can be used for comparison have skyrocketed, leading to the creation of decentralized, collaboration-based analyses such as the OpenML project.[10] However, the value of focused, reproducible ML experiments is still paramount. These observations motivated our work, in which we conduct a contemporary, open source, and thorough comparison of ML algorithms across a large set of publicly available problems, including several bioinformatics problems.

In this paper, we take a detailed look at 13 popular open source ML algorithms and analyze their performance across a set of 165 supervised classification problems in order to provide data-driven advice to practitioners who wish to apply ML to their datasets. A key part of this comparison is a full hyperparameter optimization of each algorithm. The results highlight the importance of selecting the right ML algorithm for each problem, which can improve prediction accuracy significantly on some problems. Further, we empirically quantify the effect of hyperparameter (i.e. algorithm parameter) tuning for each ML algorithm, demonstrating marked improvements in the predictive accuracy of nearly all ML algorithms. We show that the underlying behaviors of various ML algorithms cluster in terms of performance, as might be expected. Finally, based on the results of the experiments, we provide a refined set of recommendations for ML algorithms and parameters as a starting point for future researchers.

2. Methods

In this study, we compared 13 popular ML algorithms from scikit-learn,[8] a widely used ML library implemented in Python. Each algorithm and its hyperparameters are described in Table 1. The algorithms include Naïve Bayes algorithms, common linear classifiers, tree-based algorithms, distance-based classifiers, ensemble algorithms, and non-linear, kernel-based strategies. The goal was to represent the most common classes of algorithms used in literature, as well as recent state-of-the-art algorithms such as Gradient Tree Boosting.[11]

For each algorithm, the hyperparameters were tuned using a fixed grid search with 10-fold cross-validation. In our results, we compare the average balanced accuracy[12] over the 10 folds in order to account for class imbalance. We used expert knowledge about the reasonable hyperparameters to specify the ranges of values to tune for each algorithm. It is worth noting that we did not attempt to control for the *number* of total hyperparameter combinations budgeted to each algorithm. As a result, algorithms with more parameters have an advantage in the sense that they have more training attempts on each dataset. However, it is our goal to report as close to the best performance as possible for each algorithm on each dataset, and for this reason we chose to optimize each algorithm as thoroughly as possible.

The algorithms were compared on 165 supervised classification datasets from the Penn Machine Learning Benchmark (PMLB).[13] PMLB is a collection of publicly available classification problems that have been standardized to the same format and collected in a central location

with easy access via Python[a]. Although not limited to problems in biology and medicine, PMLB includes many biomedical classification problems, including tasks such as disease diagnosis, post-operative decision making, and exon boundary identification in DNA, among others. A sample of the biomedical classification tasks contained in PMLB is listed in Table 2.

Prior to evaluating each ML algorithm, we scaled the features of every dataset by subtracting the mean and scaling the features to unit variance. This scaling step was necessitated by some ML algorithms, such as the distance-based classifiers, which assume that the features of the datasets will be scaled appropriately beforehand.

The entire experimental design consisted of over 5.5 million ML algorithm and parameter evaluations in total, resulting in a rich set of data that is analyzed from several viewpoints in Section 3. As an additional contribution of this work, we have provided the complete code required both to conduct the algorithm and hyperparameter optimization study, as well as access to the analysis and results[b]. Doing so allows researchers to easily compare algorithm performance on the datasets that are most similar to their own, and to conduct further analysis pertaining to their research.

3. Results

In this section, we analyze the algorithm performance results through several lenses. First we compare the performance of each algorithm across all datasets in terms of best balanced accuracy in Section 3.1. We then look at the effect of hyperparameter tuning and model selection in Section 3.2. Finally, we analyze how algorithms cluster across the tested problems, and present a set of algorithms that maximize performance across the datasets in Section 3.3.

3.1. *Algorithm Performance*

As a simple bulk measure to compare the performance of the 13 ML algorithms, we plot the mean rankings of the algorithms across all datasets in Figure 1. Ranking is determined by the 10-fold CV balanced accuracy of each algorithm on a given dataset, with a lower ranking indicating higher accuracy. The rankings show the strength of ensemble-based tree algorithms in generating accurate models: The first, second, and fourth-ranked algorithms belong to this class of algorithms. The three worst-ranked algorithms also belong to the same class of Naïve Bayes algorithms.

In order to assess the statistical significance of the observed differences in algorithm performance across all problems, we use the non-parametric Friedman test.[14] The complete set of experiments indicate statistically significant differences according to this test ($p < 2.2e^{-16}$), and so we present a pairwise post-hoc analysis in Table 3. The post-hoc test underlines the impressive performance of Gradient Tree Boosting, which significantly outperforms every algorithm except Random Forest at the $p < 0.01$ level. At the other end of the spectrum, Multinomial NB is significantly outperformed by every algorithm except for Gaussian NB. These strong statistical results are interesting given the large set of problems and algorithms

[a]URL: `https://github.com/EpistasisLab/penn-ml-benchmarks`
[b]URL: `https://github.com/rhiever/sklearn-benchmarks`

Table 1. ML algorithms and hyperparameters tuned in the experiments.

Algorithm	**Hyperparameters**
Gaussian Naïve Bayes (GNB)	No parameters.
Bernoulli Naïve Bayes (BNB)	**alpha**: Additive smoothing parameter. **binarize**: Threshold for binarizing the features. **fit_prior**: Whether or not to learn class prior probabilities.
Multinomial Naïve Bayes (MNB)	**alpha**: Additive smoothing parameter. **fit_prior**: Whether or not to learn class prior probabilities.
Logistic Regression (LR)	**C**: Regularization strength. **penalty**: Whether to use Lasso or Ridge regularization. **fit_intercept**: Whether or not the intercept of the linear classifier should be computed.
Stochastic Gradient Descent (SGD)	**loss**: Loss function to be optimized. **penalty**: Whether to use Lasso, Ridge, or ElasticNet regularization. **alpha**: Regularization strength. **learning_rate**: Shrinks the contribution of each successive training update. **fit_intercept**: Whether or not the intercept of the linear classifier should be computed. **l1_ratio**: Ratio of Lasso vs. Ridge reguarlization to use. Only used when the 'penalty' is ElasticNet. **eta0**: Initial learning rate. **power_t**: Exponent for inverse scaling of the learning rate.
Passive Aggressive Classifier (PAC)	**loss**: Loss function to be optimized. **C**: Maximum step size for regularization. **fit_intercept**: Whether or not the intercept of the linear classifier should be computed.
Support Vector Classifier (SVC)	**kernel**: 'linear', 'poly', 'sigmoid', or 'rbf'. **C**: Penalty parameter for regularization. **gamma**: Kernel coef. for 'rbf', 'poly' & 'sigmoid' kernels. **degree**: Degree for the 'poly' kernel. **coef0**: Independent term in the 'poly' and 'sigmoid' kernels.
K-Nearest Neighbor (KNN)	**n_neighbors**: Number of neighbors to use. **weights**: Function to weight the neighbors' votes.
Decision Tree (DT)	**min_weight_fraction_leaf**: The minimum number of (weighted) samples for a node to be considered a leaf. Controls the depth and complexity of the decision tree. **max_features**: Number of features to consider when computing the best node split. **criterion**: Function used to measure the quality of a split.
Random Forest (RF) & Extra Trees Classifier (ERF)	**n_estimators**: Number of decision trees in the ensemble. **min_weight_fraction_leaf**: The minimum number of (weighted) samples for a node to be considered a leaf. Controls the depth and complexity of the decision trees. **max_features**: Number of features to consider when computing the best node split. **criterion**: Function used to measure the quality of a split.
AdaBoost (AB)	**n_estimators**: Number of decision trees in the ensemble. **learning_rate**: Shrinks the contribution of each successive decision tree in the ensemble.
Gradient Tree Boosting (GTB)	**n_estimators**: Number of decision trees in the ensemble. **learning_rate**: Shrinks the contribution of each successive decision tree in the ensemble. **loss**: Loss function to be optimized via gradient boosting. **max_depth**: Maximum depth of the decision trees. Controls the complexity of the decision trees. **max_features**: Number of features to consider when computing the best node split.

compared here. Because the No Free Lunch theorem[15] guarantees that all algorithms perform the same on average over all possible classes of problems, the differentiated results imply that the problems in the PMLB belong to a related subset of classes. The initial PMLB study[13] also noted the similarity in properties of several publicly available datasets, which could lead

Table 2. A non-exhaustive sample of datasets included in the PMLB archive that pertain to biomedical classification.

Data Set	Classes	Samples	Dimensions	Description
allbp	3	3772	29	Diagnosis
allhyper	4	3771	29	Diagnosis
allhypo	3	3770	29	Diagnosis
ann-thyroid	3	7200	21	Diagnosis
biomed	2	209	8	Diagnosis
breast-cancer-wisconsin	2	569	30	Diagnosis
breast-cancer	2	286	9	Diagnosis
diabetes	2	768	8	Diagnosis
dna	3	3186	180	Locating exon boundaries
GMT 2w-20a-0.1n	2	1600	20	Simulated GWAS
GMT 2w-1000a-0.4n	2	1600	1000	Simulated GWAS
liver-disorder	2	345	6	Diagnosis
molecular-biology_promoters	2	106	58	Identify promoter sequences
postoperative-patient-data	2	88	8	Choose post-operative treatment

to inflated statistical significance. Nevertheless, it cannot be denied that the results are relevant to classification tasks encountered in real-world and biological contexts, since the vast majority of datasets used here are taken from those contexts.

Given these bulk results, it is tempting to recommend the top-ranked algorithm for all problems. However, this neglects the fact that the top-ranked algorithms may not outperform others for some problems. Furthermore, when simpler algorithms perform on par with a more complex one, it is often preferable to choose the simpler of the two. With this in mind, we investigate pair-wise "outperformance" by calculating the percentage of datasets for which one algorithm outperforms another, shown in Figure 2. One algorithm outperforms another on a dataset if it has at least a 1% higher 10-fold CV balanced accuracy, which represents a minimal threshold for improvement in predictive accuracy.

In terms of "outperformance," it is worth noting that no one ML algorithm performs best across all 165 datasets. For example, there are 9 datasets for which Multinomial NB performs as well as or better than Gradient Tree Boosting, despite being the overall worst- and best-ranked algorithms, respectively. Therefore, it is still important to consider different ML algorithms when applying ML to new datasets.

3.2. *Effect of Tuning and Model Selection*

Most ML algorithms contain several hyperparameters that can affect performance significantly (for example, the max tree depth of a decision tree classifier). Our experimental results allow us to measure the extent to which hyperparameter tuning via grid search improves each algorithm's performance compared to its baseline settings. We also measure the effect that model selection has on improving classifier performance.

Figure 3 compares the performance of the tuned classifier to its default settings for each algorithm across all datasets. The results demonstrate why it is unwise to use default ML al-

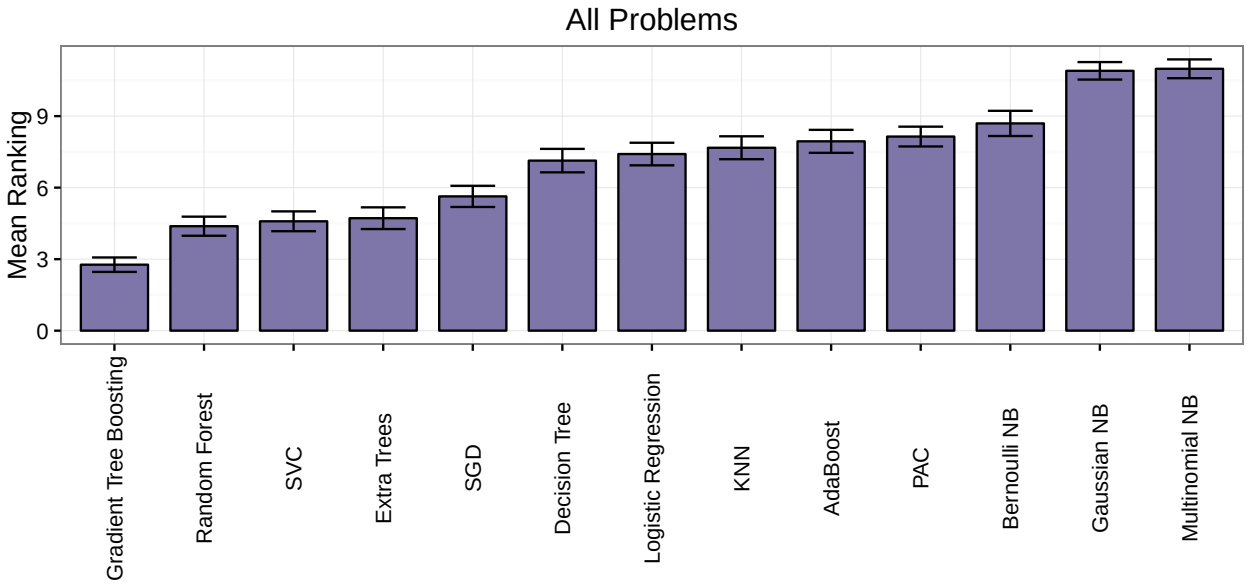


Fig. 1. Average ranking of the ML algorithms over all datasets. Error bars indicate the 95% confidence interval.

Table 3. Post-hoc Friedman test of algorithm rankings across all problems. Bold values indicate $p < 0.01$.

	GTB	RF	SVC	ERF	SGD	DT	LR	KNN	AB	PAC	BNB	GNB
RF	0.01	-	-	-	-	-	-	-	-	-	-	-
SVC	**0.001**	1	-	-	-	-	-	-	-	-	-	-
ERF	**0.0004**	1	1	-	-	-	-	-	-	-	-	-
SGD	**3e-10**	0.1	0.4	0.6	-	-	-	-	-	-	-	-
DT	**0**	**3e-09**	**1e-07**	**3e-07**	0.03	-	-	-	-	-	-	-
LR	**0**	**1e-11**	**1e-09**	**1e-07**	**0.003**	1	-	-	-	-	-	-
KNN	**0**	**1e-13**	**5e-12**	**7e-11**	**0.0002**	1	1	-	-	-	-	-
AB	**0**	**6e-15**	**4e-14**	**4e-13**	**3e-06**	0.8	1	1	-	-	-	-
PAC	**0**	**2e-16**	**3e-15**	**8e-15**	**2e-07**	0.5	0.9	1	1	-	-	-
BNB	**0**	**0**	**0**	**0**	**4e-10**	0.02	0.1	0.4	0.9	1	-	-
GNB	**0**	**0**	**0**	**0**	**0**	**0**	**2e-15**	**9e-13**	**1e-10**	**5e-09**	**2e-05**	-
MNB	**0**	**0**	**0**	**0**	**0**	**0**	**2e-15**	**7e-14**	**1e-11**	**4e-09**	**4e-06**	1

gorithm hyperparameters: tuning often improves an algorithm's accuracy by 3-5%, depending on the algorithm. In some cases, parameter tuning led to CV accuracy improvements of 50%.

Figure 4 shows the improvement in 10-fold CV accuracy attained both by model selection and hyperparameter optimization compared to the average performance on each dataset. The results demonstrate that selecting the best model and tuning it leads to approximately a 20% increase in accuracy, up to more than a 60% improvement for certain datasets. Thus, both selecting the right ML algorithm and tuning its parameters is vitally important for most problems.

3.3. *Algorithm Coverage*

Given that several of the 13 algorithms studied here have similar underlying methodologies, we would expect their performance across problems to align with the underlying assumptions

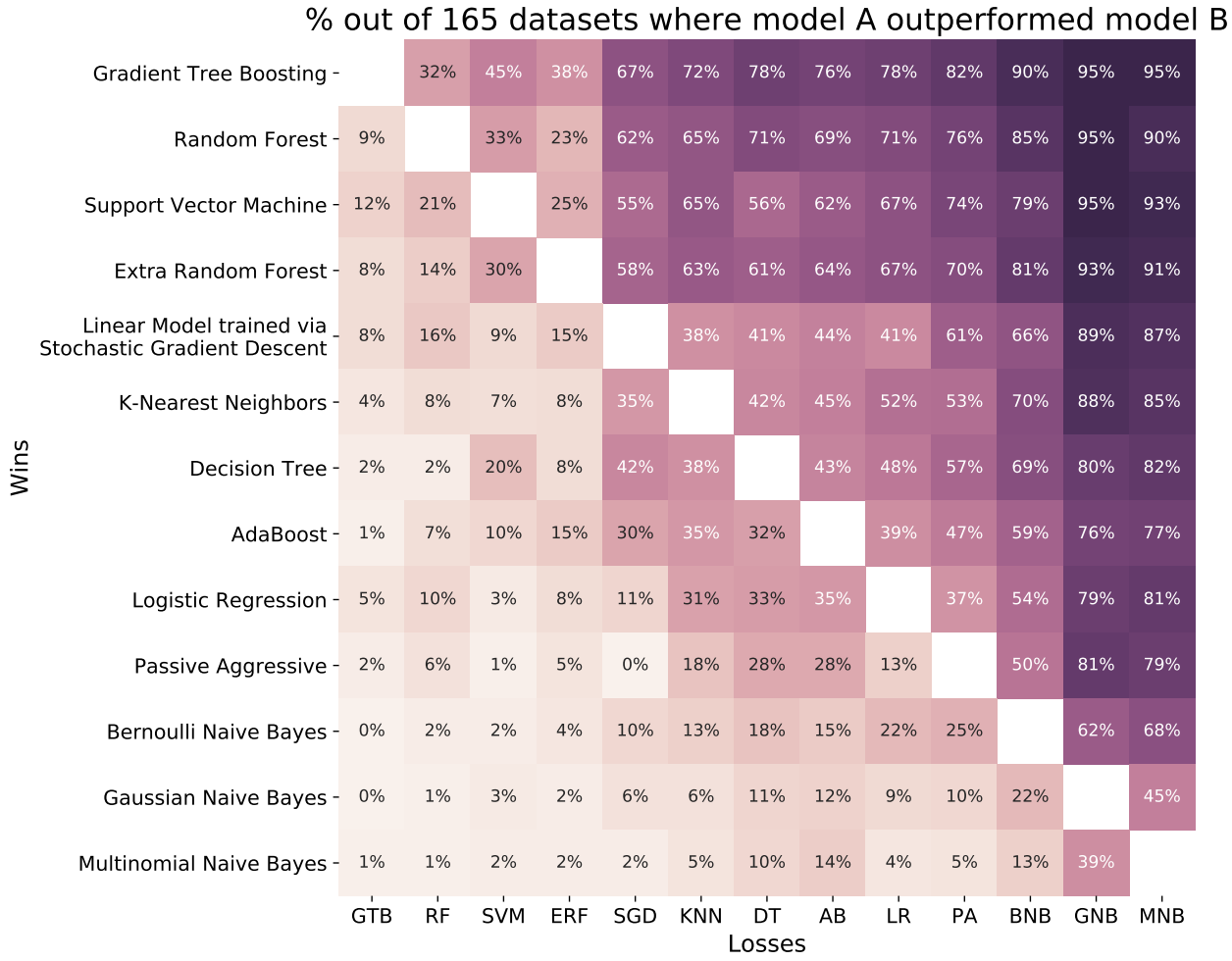


Fig. 2. Heat map showing the percentage out of 165 datasets a given algorithm outperforms another algorithm in terms of best accuracy on a problem. The algorithms are ordered from top to bottom based on their overall performance on all problems. Two algorithms are considered to have the same performance on a problem if they achieved an accuracy within 1% of each other.

that the modeling techniques have in common. One way to assess whether this holds is to cluster the performance of different algorithms across all datasets. We perform hierarchical agglomerative clustering on the 10-fold CV balanced accuracy results, which leads to the clusters shown in Figure 5. Indeed, we find that algorithms with similar underlying assumptions or methodologies cluster in terms of their performance across the datasets. For example, the Naïve Bayes algorithms (i.e., Multinomial, Gaussian, and Bernoulli) perform most similarly to each other, and the linear algorithms (i.e., passive aggressive and logistic regression) also cluster. The ensemble algorithms of Extra Trees and Random Forests, which both use ensembles of decision trees, also cluster. Support Vector Machines and Gradient Tree Boosting appear to be quite different algorithms, but given that both are able to capture nonlinear interactions between variables, it is less surprising that they cluster as well.

We present a list of five recommended algorithms and parameter settings in Table 4.

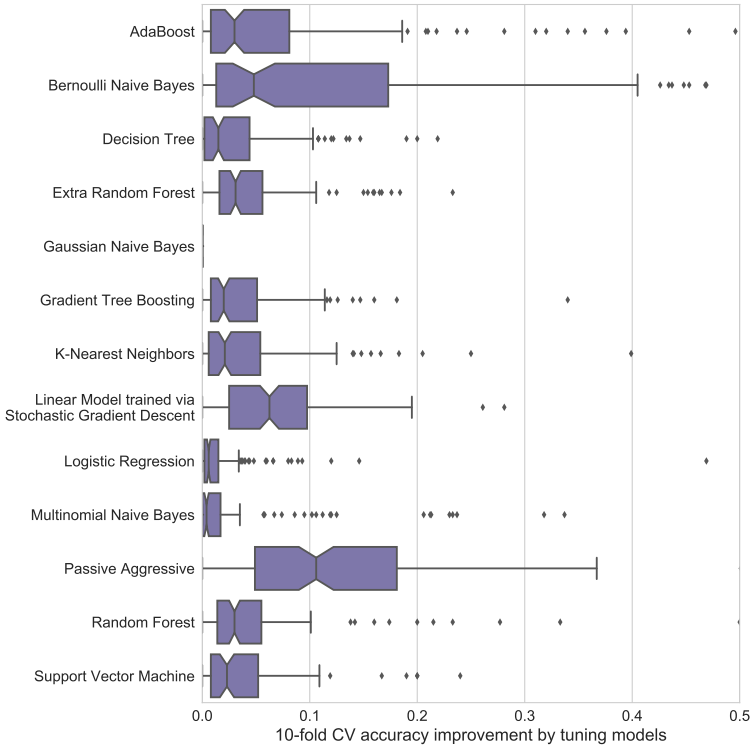


Fig. 3. Improvement in 10-fold CV accuracy by tuning each ML algorithm's parameters instead of using the default parameters from scikit-learn.

The five algorithms and parameters here are those that maximize the coverage of the 165 benchmark datasets, meaning that they perform within 1% of the best 10-fold CV balanced accuracy obtained on the maximum number of datasets in the experiment. For the datasets in PMLB, these five algorithms and associated parameters cover 106 out of 165 datasets to within 1% balanced accuracy. Notably, 163 out of 165 datasets can be covered by tuning the parameters of the five listed algorithms. Based on the available evidence, these recommended algorithms should be a good starting point for achieving reasonable predictive accuracy on a new dataset.

4. Discussion and Conclusions

We have empirically assessed 13 supervised classification algorithms on a set of 165 supervised classification datasets in order to provide a contemporary set of recommendations to bioinformaticians who wish to apply ML algorithms to their data. The analysis demonstrates the strength of state-of-the-art, tree-based ensemble algorithms, while also showing the problem-dependent nature of ML algorithm performance. In addition, the analysis shows that selecting the right ML algorithm and thoroughly tuning its parameters can lead to a significant improvement in predictive accuracy on most problems, and is there a critical step in every ML application. We have made the full set of experiments and results available online to encourage bioinformaticians to easily gather information most pertinent to their area of study.

Even with a large set of results, it is difficult to recommend specific algorithms or parameter

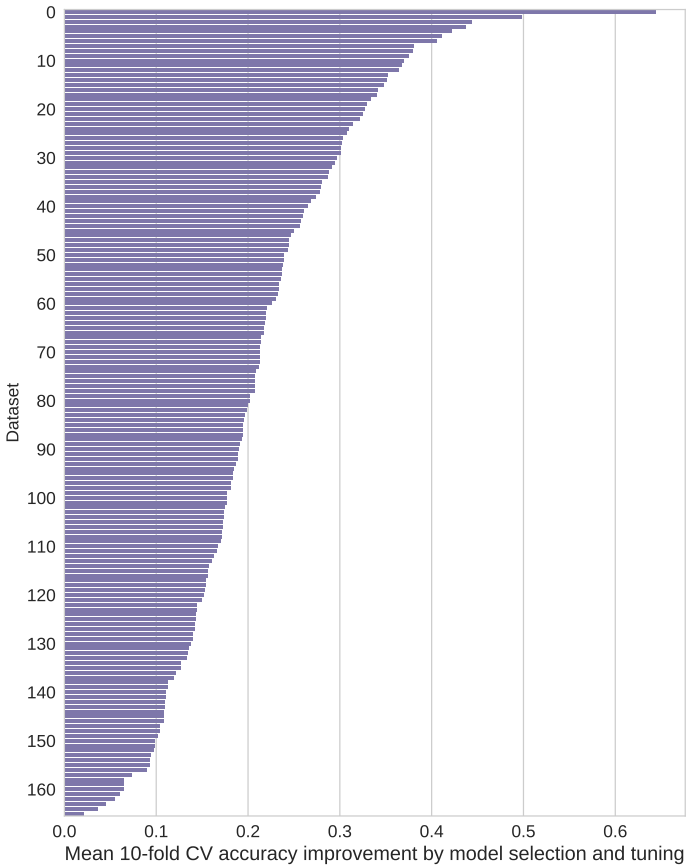


Fig. 4. Improvement in 10-fold CV accuracy by model selection and tuning, relative to the average performance on each dataset.

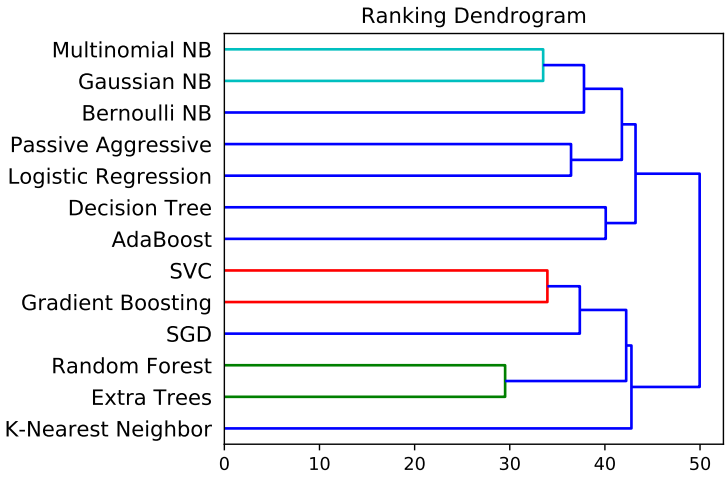


Fig. 5. Hierarchical clustering of ML algorithms by accuracy rankings across datasets.

Table 4. Five ML algorithms and parameters that maximize coverage of the 165 benchmark datasets. These algorithm and parameter names correspond to their scikit-learn implementations.

Algorithm	Parameters	Datasets Covered
GradientBoostingClassifier	loss="deviance" learning_rate=0.1 n_estimators=500 max_depth=3 max_features="log2"	51
RandomForestClassifier	n_estimators=500 max_features=0.25 criterion="entropy"	19
SVC	C=0.01 gamma=0.1 kernel="poly" degree=3 coef0=10.0	16
ExtraTreesClassifier	n_estimators=1000 max_features="log2" criterion="entropy"	12
LogisticRegression	C=1.5 penalty="l1" fit_intercept=True	8

settings with a strong amount of generality. As a starting point, we provided recommendations for 5 different ML algorithms and parameters based on their collective coverage of the 165 datasets from PMLB. However, it is important to note that these algorithms and parameters will not work best on all supervised classification problems, and they should only be used as starting points. For a more nuanced approach, the similarity of the dataset on which ML is to be applied to datasets in PMLB could be quantified, and the set of algorithms that performed best on those similar datasets could be used. In lieu of detailed problem information, one could also use automated ML tools[16,17] and AI-driven ML platforms[18] to perform model selection and parameter tuning automatically.

Of course, some bioinformaticians may value properties of ML algorithms aside from predictive accuracy. For example, ML algorithms are often used as a "microscope" to model and better understand the complex biological systems from which the data was sampled. In this use case, bioinformaticians may value the interpretability of the ML model, in which case black box predictive models that cannot be interpreted are of little use.[19] Although the logistic regression and decision tree algorithms are often outperformed by tree-based ensemble algorithms in terms of predictive accuracy (Figure 2), linear models and shallow decision trees often provide a useful trade-off between predictive accuracy and interpretability. Furthermore, methods such as LIME[19] show promise for explaining why complex, black box models make individual predictions, which can also be useful for model interpretation.

There are several opportunities to extend the analysis in this paper in future work. A natural extension should be made to regression, which is used several biomedical applications such

as quantitative trait genetics. In addition, these experiments do not take into account feature preprocessing, feature construction, and and feature selection, although it has been shown that learning better data representations can significantly improve ML performance.[20] We plan to extend this work to analyze the ability of various feature preprocessing, construction, and selection strategies to improve model performance. In addition, the experimental results contain rich information about the performance of different learning algorithms as a function of the datasets. In future work, we will take a deeper look into the properties of datasets that influence the performance of specific algorithms. By relating these dataset properties to specific areas of bioinformatics, we may be able to generate tailored recommendations for ML algorithms that work best for specific applications.

5. Acknowledgments

We thank Dr. Andreas C. Müller for his valuable input during the development of this project, as well as the Penn Medicine Academic Computing Services for the use of their computing resources. This work was supported by NIH grants P30-ES013508, AI116794, DK112217 and LM012601, as well as the Warren Center for Network and Data Science at the University of Pennsylvania.

References

1. H. Bhaskar, D. C. Hoyle and S. Singh, *Computers in Biology and Medicine* **36**, 1104 (2006), Intelligent Technologies in Medicine and Bioinformatics.
2. B. A. McKinney, D. M. Reif, M. D. Ritchie and J. H. Moore, *Applied Bioinformatics* **5**, 77 (2006).
3. Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng and M. C. Stumpe, Detecting cancer metastases on gigapixel pathology images arXiv e-print. https://arxiv.org/abs/1703.02442, (2017).
4. R. D. King, C. Feng and A. Sutherland, *Applied Artificial Intelligence an International Journal* **9**, 289 (1995).
5. A. C. Tan and D. Gilbert, An empirical comparison of supervised machine learning techniques in bioinformatics, in *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003-Volume 19*, 2003.
6. R. Caruana and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in *Proceedings of the 23rd International Conference on Machine learning*, 2006.
7. M. Fernández-Delgado, E. Cernadas, S. Barro and D. Amorim, *Journal of Machine Learning Research* **15**, 3133 (2014).
8. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
9. E. Frank, M. Hall, L. Trigg, G. Holmes and I. H. Witten, *Bioinformatics* **20**, 2479 (2004).
10. J. Vanschoren, J. N. Van Rijn, B. Bischl and L. Torgo, *ACM SIGKDD Explorations Newsletter* **15**, 49 (2014).
11. T. J. Hastie, R. J. Tibshirani and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, USA, 2009).
12. D. R. Velez *et al.*, *Genetic Epidemiology* **31**, 306 (2007).
13. R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz and J. H. Moore, PMLB:

A Large Benchmark Suite for Machine Learning Evaluation and Comparison arXiv e-print. https://arxiv.org/abs/1703.00512, (2017).
14. J. Demšar, *Journal of Machine Learning Research* **7**, 1 (2006).
15. D. H. Wolpert and W. G. Macready, *IEEE Transactions on Evolutionary Computation* **1**, 67 (1997).
16. R. S. Olson, N. Bartley, R. J. Urbanowicz and J. H. Moore, Evaluation of a tree-based pipeline optimization tool for automating data science, in *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, 2016.
17. M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter, Efficient and robust automated machine learning, in *Advances in Neural Information Processing Systems 28*, eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett (Curran Associates, Inc., 2015) pp. 2962–2970.
18. R. S. Olson, M. Sipper, W. La Cava, S. Tartarone, S. Vitale, J. H. Fu, Weixuan Holmes and J. H. Moore, A system for accessible artificial intelligence arXiv e-print. https://arxiv.org/abs/1705.00594, (2017).
19. M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (ACM, New York, NY, USA, 2016).
20. W. La Cava and J. H. Moore, Ensemble representation learning: an analysis of fitness and survival for wrapper-based genetic programming methods, in *Proceedings of the Genetic and Evolutionary Computation Conference 2017*, 2017.

Improving the explainability of Random Forest classifier – user centered approach

Dragutin Petkovic[† 1, 3], Russ Altman[2], Mike Wong[3], Arthur Vigil[4]
[1]*Computer Science Department, San Francisco State University (SFSU), 1600 Holloway Ave., San Francisco CA 94132, Petkovic@sfsu.edu*
[2]*Department of Bioengineering, Stanford University, 443 Via Ortega Drive, Stanford, CA 94305-4145*
[3]*SFSU Center for Computing for Life Sciences, 1600 Holloway Ave., San Francisco, CA 94132*
[4]*Twist Bioscience, 455 Mission Bay Boulevard South, San Francisco, CA 94158*

Machine Learning (ML) methods are now influencing major decisions about patient care, new medical methods, drug development and their use and importance are rapidly increasing in all areas. However, these ML methods are inherently complex and often difficult to understand and explain resulting in barriers to their adoption and validation. Our work (RFEX) focuses on enhancing Random Forest (RF) classifier explainability by developing easy to interpret *explainability summary reports* from trained RF classifiers as a way to improve the explainability for (often non-expert) users. RFEX is implemented and extensively tested on Stanford FEATURE data where RF is tasked with predicting functional sites in 3D molecules based on their electrochemical signatures (features). In developing RFEX method we apply *user-centered* approach driven by explainability questions and requirements collected by discussions with interested practitioners. We performed formal usability testing with 13 expert and non-expert users to verify RFEX usefulness. Analysis of RFEX explainability report and user feedback indicates its usefulness in significantly increasing explainability and user confidence in RF classification on FEATURE data. Notably, RFEX summary reports easily reveal that one needs very few (from 2-6 depending on a model) top ranked features to achieve 90% or better of the accuracy when all 480 features are used.

Keywords: Random Forest, Explainability, Interpretability, Stanford FEATURE

1. Introduction, Background and Motivation

Machine Learning (ML) methods applied on large amounts of biological, medical and life science data for use in academic, R&D and business environments are now influencing major decisions about patient care, new medical methods, drug development and their use and importance are rapidly increasing in all areas. However, algorithms and software implementing ML methods are inherently complex and often difficult to understand and explain both to non-experts as well as experts. In addition, ML training databases used to derive predictive models are often large and complex, with noisy data, and in many cases are imbalanced containing much fewer positive class samples than background samples making commonly used "average" classification accuracy measures inadequate. All this makes it very challenging to understand, evaluate and be confident about results of ML performance. The interest in explaining how ML systems work is lately also driven by general public and funding agencies given the penetration of ML in all aspects of our

lives and not only in bio-science. This is indicated by articles in popular press, many recent blogs, new DARPA program on explainable AI [1], FDA requirements for future data mining [2]), as well as by recent workshops focused on this subject (e.g. 2016 ICML Workshop on *Human Interpretability in Machine Learning*; PSB 2018 Workshop on *Machine Learning and Deep Analytics for Biocomputing: Call for Better Explainability)* . Problems arising with the lack of explainability are being increasingly documented and discussed [3]. However, review of the published scientific literature on explainability in ML shows that very few research efforts and methods focus specifically on ML explainability. In addition, there is practically no work following the tried-and-true best practice of "user centered design" in which one engages users who are the ultimate judges and beneficiaries of explainability. We believe that the importance and benefits of being able to explain why and how ML decisions models make their decisions to non-ML experts and experts alike (e.g. *explainability*) are critical and must be addressed. We can further define explainability in ML as *model explainability* - why and how the trained ML model works overall, and *sample explainability* - how ML made a decision for a specific data sample (e.g. sample under investigation or sample from the training database). ML training is outside of our scope since it is usually well explained. Note that the ML approach can be reproducible but it still may not be sufficiently explainable. The improved explainability of ML in biocomputing and other areas will result in the following benefits: a) increased confidence of application and domain experts who are key decision makers (and who are often non ML-experts) in adopting ML; b) better testing and prevention of cases where ML approach produces results based on fundamentally wrong reasons (e.g. based on features not available in real application, wrong data in training databases or imperfect algorithm); c) easier evaluation, audit and verification of ML results for granting agencies, government organizations like FDA, and editors/publishers who need to decide what is being published and with what level of detail; d) simplification and reduction of the cost of application of ML in practice (e.g. by knowing which smaller feature subsets produce adequate accuracy more cost effective systems can be built); e) improved "maintenance" where ML method has to be changed or tuned to new data or decision needs; and f) possible discovery of new knowledge and ideas (e.g. by discovering new patterns and factors that contribute to ML decisions)

1.1 Random Forest (RF) Classifiers

RF is a popular and powerful ensemble supervised classification method [4]. Due to its superior accuracy and robustness, and some ability to offer insights by ranking of its features, RF has effectively been applied to various machine learning applications, including many in bioinformatics and medical imaging. RF consists of a set of decision trees, each of which is generated by the bagging algorithm with no pruning, forming a "forest" of classifiers voting for a particular class. To train a RF, two parameters, the number of tress (*ntree*) in the forest and the number of randomly selected features/variables used to evaluate at each tree node (*mtry*), must be supplied, as well as a training database with ground-truth class labels. RF also allows adjustment of the voting threshold or *cutoff* (fraction of trees in the forest needed to vote for a given class), which is used to compute *recall, precision and f-score*. The accuracy estimate built into the RF

algorithm and all its software implementations is called Out of Bag Error (OOB), which measures the average misclassification ratio of samples not used for RF training. One of the RF algorithm's strengths, and reasons we chose it, is its ability to calculate various feature/variable importance measures which can form the basis for enhancing its explainability [4, 6]. For this work we chose *MDA (mean decrease in accuracy)* as our main feature importance (ranking) measure. MDA measures the average increase of the error rate (i.e. decrease of accuracy) against random permutation of feature values across OOB cases. With a trained RF, the values of OOB cases for a tree are first permuted along the m-*th* feature. Then error rate with and without this permutations are recorded and their difference computed. This is repeated for all decision trees and the average of these differences gives the m-*th* feature's MDA. We leverage the fact that MDA can be computed for + and – class separately (e.g. MDA+ and MDA-), thus providing better explainability. For main measure of RF classification accuracy, given that in most cases we have unbalanced training data (as in our case study), instead of commonly used OOB we use *f-score (f=2 * (precision*recall)/(precision + recall))* determined using K-fold (we use K=5) *stratified cross validation (SCV)* [5, 7] where we independently partition samples to K folds for positive and negative sample pools first, then merge positive and negative folds to form the K folds that preserves the class distribution of the original dataset. The SCV procedure is then repeated, with varying of the RF tree voting cutoff threshold to maximize f-score.

1.2 Related work on Explainability for Random Forest Classifiers

Basic RF classification results traditionally comprise: information on the training data; optimal RF parameters; and the set of estimated accuracy measures with description of evaluation methods being used. Current methods for RF explainability fall into two basic categories. *Feature ranking* uses RF-provided variable importance measures like e.g. RF-provided Gini, MDA (mean decrease in accuracy) or others, to present them in *tables or horizontal bar charts* sorted by chosen variable importance measure, as in [8, 9, 10, 22, 23]. Highly ranked features are then assumed to play important role RF predictions, which in turn may offer some insights into the observed process or can even be used to clean-up training databases [23]. However, this information is insufficient for more substantial explainability. In addition, feature ranking is seldom done for + and – class separately, thus posing problems for frequent case of imbalanced data sets. Enhanced ranked feature representation with more details for helping RF epxlainability has been reported by [11]. One innovative idea to look at *pairs* of highly ranked feature and extract positive and negative pair-wise feature interactions has been reported in [12]. The second basic approach is *rule extraction from trained RF*. This method consists of: a) performing standard RF training; b) defining rules by analyzing trained RF trees (resulting in very large set of rules, order of 100 K); and c) reducing the number and complexity of extracted rules by optimization e.g. minimizing some metrics (accuracy, coverage, rule complexity…) to reduce to 10s – 100s of rules, each with 1-10 or so conditions [13-17]. Common problem with this approach is still a large number of complex rules hard to interpret by humans and lack of tradeoffs between accuracy and number of rules used. Our prior work on explainability for RF was motivated by our original joint work with Stanford Helix team on applying Support Vector Machines (SVM) [18] and RF [5] to their

FEATURE data [19] where we show very good classification results measured by high recall and precision. In [5] we made first attempts to improve explainability by using RF-provided variable importance measures but did not analyze positive vs. negative classes separately and achieved very limited explainability improvements. The published work on explainability for RF (and other ML methods) can be summarized as follows: a) in spite of the fact that explainability is geared toward non-expert and expert *human* users no design consideration and formal evaluations related to *human usability* of proposed explanations and representations have been attempted; b) proposed explainability representations do not offer easy to use and critically important tradeoffs between accuracy and complexity of ML; c) analysis of + vs. – class separately (critical for a common case of unbalanced training data) has seldom been done; and d) feature reduction is generally not applied *before* explainability steps, thus necessitating complex approaches using large numbers of features impeding the explainability.

1.3 User-Centered Approach in Enhancing Random Forest Explainability - RFEX

RFEX method starts with standard approach to RF classification, using training database and standard RF tools/algorithms producing *base* RF accuracy estimates. In a series of steps RFEX then produces a RFEX *summary report* which is to be used by human (often non-expert) users to improve the explainability of original trained RF classifier (approach advocated in [1]). In developing RFEX we took a *user-centered-approach* (which to the best of our knowledge has not been tried by others): we guide our RFEX method by user-centered explainability questions or requirements collected by discussions and observations with interested practitioners, and then we test usefulness of RFEX as it is applied to FEATURE data by formal usability experiments. Based on our experience and investigation (especially in common case of imbalanced data) most users will lack full understanding of how and why RF works based only on the traditionally provided information (e.g. info on training data, optimized RF parameters, accuracy evaluation methods and estimates) and would pose a number of *explainability questions* to gain more insights and confidence *before* adopting it:

1. *Can the explainability analysis be done for + and – class separately (critical in frequent case of imbalanced training data)?*
2. *What are most important features contributing to ML prediction and how do they rank in importance? What is the relationship of most important features for + vs. – class, is there any overlap?*
3. *What is the loss/tradeoffs of accuracy if I use only certain subset of most important features?*
4. *What is "direction" of features? Abundance ("more of it" or "presence") or deficiency ("less of it" or "absence")? What thresholds I can use to determine this?*
5. *Which features interact together?*
6. *Can this analysis be presented in an easy to understand and simple summary for ML/domain experts and non-experts?*

We then use these explainability questions as "user-driven requirements" for the design of RFEX resulting in *one page RFEX explainability summary* report (one page was a design goal).

RFEX is implemented and extensively tested on Stanford FEATURE data. Most importantly (and to the best of our knowledge never done before) we also performed formal RFEX usability study with 13 users of various experience in RF and FEATURE to assess RFEX utility in increase of RF classification results' explainability.

2. Case Study: RFEX Applied to Stanford FEATURE data

Stanford FEATURE [19] is a system for classifying protein functional sites from electrochemical signatures/properties around those functional sites. FEATURE data is organized as feature vectors each describing a site in a three dimensional protein structure, using 80 physicochemical properties (features) in 6 concentric spherical shells, each 1.25 Ångstroms thick, yielding 480 feature values per vector. Each feature is denoted by the physicochemical property name, followed by its shell location (Si). FEATURE data i.e. the training database used for RF training, contains feature vectors at known positive (functional site) and negative (background) class labels for each protein functional model [18]. FEATURE training data is highly imbalanced e.g. there are two to three orders of magnitude more negative (background) vs. positive (functional sites) samples. For the work in this paper we used the same 7 FEATURE models selected in experiments in [5], which are subset of models analyzed in [18], see Table 1.

2.1 Creation of RFEX Summary Reports

We first estimate "*base RF classification accuracy*" by training RF on FEATURE data using all 480 features and we estimate accuracy using f-score with 5 fold stratified cross validation (SCV). We use *ntree* = 500 and vary *mtry* as {10, 20, 40} to find the optimal combination maximizing f-score. This experiment confirmed high RF predictive power for all 7 models (as reported before in [5]). Table 1 shows 7 models, and for each model their training data and several base RF accuracy measures using all 480 features. For our analysis we use Open Source packages which implement RF and provide MDA measures as well as various methods for RF training, including SCV, namely R package [20] and *caret* tool kit [21], along with Python integration and application code. We then proceed in developing "*explainable RF model/representation*" using RFEX approach which involves a series of steps and strategies (including novel explainability measures) designed to explicitly answer all 6 explainability questions above. We show details of experiments for one FEATURE model, namely ASP_PROTEASE.4.ASP.OD1 and then present RFEX one page summaries for two FEATURE models (ASP_PROTEASE.4.ASP.OD1 and EF_HAND_1.1.ASP.OD1), shown in Fig 2 and Fig 3. Detailed experimental results and RFEX summary reports for all 7 models are presented in [7]. All our analysis is performed separately for positive (functional sites) and negative (background) class (*explainability question 1*).

To rank features by importance (*explainability question 2*) we use MDA + (ranking for positive class) and MDA – (ranking for negative class) provided from above trained RF classifiers using standard RF tools, as explained in Section 1.1. By leveraging feature rankings for positive and negative class *separately* (seldom done in published literature) we achieve more explainability

given that FEATURE data is highly unbalanced. Indeed, this is justified by the fact that this method produces sets of differently ranked features for + and - class, as seen below in Table 2, showing 20 top ranked features for + and – class separately. Features appearing in both lists are bold.

Table 1. Summary of RF basic accuracy using all 480 features (described by several accuracy measures) for 7 FEATURE models used in this study

model	num.positive	num.negative	mtry	recall	precision	fscore	oob	positive.oob	negative.oob
ASP_PROTEASE.4.ASP.OD1	1585	48577	40	0.99180	0.99873	0.99525	0.00032	0.00883	0.00004
EF_HAND_1.1.ASP.OD1	1811	48145	40	0.91275	1.00000	0.95439	0.00268	0.07289	0.00004
EF_HAND_1.1.ASP.OD2	1811	50290	40	0.91496	0.99941	0.95532	0.00248	0.07013	0.00004
EF_HAND_1.9.GLN.NE2	15	47325	10	0.13333	1.00000	0.23529	0.00027	0.86667	0.00000
IG_MHC.3.CYS.SG	2017	49081	40	0.98017	0.98266	0.98141	0.00123	0.01487	0.00067
PROTEIN_KINASE_ST.5.ASP.OD1	1096	48924	40	0.94162	0.99901	0.96947	0.00112	0.05018	0.00002
TRYPSIN_HIS.5.HIS.ND1	446	50007	40	0.94177	0.99767	0.96892	0.00050	0.05381	0.00002

We then follow with critical (and seldom used by others) step of *early* complexity and dimensionality reduction where we aim to provide tradeoffs between using the subset of feature vs. loss of accuracy (*explainability question 3*). We focus on positive class and first re-train RF on top 2 ranked features from Table 2 using 5-fold SCV on original training data and record average f-score and its variation (measured by standard deviation). We then add next top ranked feature and retrain RF only on those 3 features. We repeat this adding top ranked features one by one until top 20th feature to obtain graph in Fig. 1 showing that by using very small subset of features (less than 20 from total of 480) one can achieve almost full base accuracy.

Table 2. Top 20 ranked features for ASP_PROTEASE.4.ASP.OD1) for positive class (MDA+ ranked) and negative class (MDA- ranked), with their *feature direction* (+/- columns). Features appearing in both lists are bold

Top Features by +MDA	+/-	Top Features by -MDA	+/-
NEG_CHARGE_s2	**+ (0.91)**	**RESIDUE_NAME_IS_GLY_s2**	**- (0.99)**
RESIDUE_CLASS1_IS_UNKNOWN_s2	**+ (0.84)**	**RESIDUE_CLASS1_IS_UNKNOWN_s2**	**- (0.99)**
RESIDUE_NAME_IS_GLY_s2	**+ (0.82)**	RESIDUE_CLASS2_IS_POLAR_s2	- (0.93)
SECONDARY_STRUCTURE1_IS_STRAND_s5	**+ (0.96)**	RESIDUE_NAME_IS_LEU_s5	- (0.96)
RESIDUE_NAME_IS_GLY_s3	**+ (0.88)**	**SECONDARY_STRUCTURE1_IS_STRAND_s5**	**- (0.88)**
RESIDUE_CLASS1_IS_UNKNOWN_s3	**+ (0.89)**	PEPTIDE_s2	- (0.85)
SOLVENT_ACCESSIBILITY_s5	**- (0.93)**	SOLVENT_ACCESSIBILITY_s1	+ (0.83)
SOLVENT_ACCESSIBILITY_s4	**- (0.82)**	**RESIDUE_NAME_IS_GLY_s3**	**- (0.96)**
RESIDUE_NAME_IS_THR_s4	**+ (0.86)**	**ATOM_TYPE_IS_O2_s2**	**- (0.95)**
ATOM_TYPE_IS_O2_s2	**+ (0.86)**	**NEG_CHARGE_s2**	**- (0.95)**
SECONDARY_STRUCTURE1_IS_TURN_s3	**+ (0.90)**	**RESIDUE_CLASS1_IS_UNKNOWN_s3**	**- (0.96)**
RESIDUE_CLASS2_IS_BASIC_s4	- (0.99)	MOBILITY_s5	+ (0.92)
CHARGE_WITH_HIS_s2	+ (0.95)	**SOLVENT_ACCESSIBILITY_s4**	**+ (0.92)**
CHARGE_s2	+ (0.93)	**SECONDARY_STRUCTURE1_IS_TURN_s3**	**- (0.89)**
NEG_CHARGE_s3	**+ (0.88)**	**RESIDUE_NAME_IS_THR_s4**	**- (0.94)**
RESIDUE_NAME_IS_THR_s3	+ (0.77)	**RESIDUE_CLASS2_IS_POLAR_s3**	**- (0.95)**
SECONDARY_STRUCTURE1_IS_TURN_s2	+ (0.84)	**SOLVENT_ACCESSIBILITY_s5**	**+ (0.92)**
RESIDUE_CLASS2_IS_POLAR_s3	**+ (0.83)**	RESIDUE_CLASS1_IS_HYDROPHOBIC_s5	- (0.82)
SECONDARY_STRUCTURE1_IS_STRAND_s4	+ (0.94)	**NEG_CHARGE_s3**	**- (0.86)**
RESIDUE_NAME_IS_ASP_s3	+ (0.93)	ELEMENT_IS_ANY_s4	+ (0.50)

To understand the *feature direction (+/- columns in Table 2)* we introduce novel measure *DIR(I)* as + *(n)* or – *(n)* denoting fraction of times (n) when feature I was above (+) (*abundance*) or below (-) (*deficiency*) the threshold when making correct prediction, for all trees in the forest making a correct prediction, and for all test samples. We measure feature direction for top ranked 20 features, separately for positive and negative class (*explainability question 4*), shown in Table 2

as +/- columns. We also recorded histograms of threshold values used for top 5 ranked features but this information proved to be hard to use due to its variability. The table 2 also reveals some important confidence building explainability information: a) set of features best predicting positive vs. negative class is different and/or ranked differently; b) some of these features overlap (e.g. appear in both lists), and in those cases their direction is *opposite*; and c) all features are clearly either abundant or deficient (e.g. have high value of n). To measure which features "interact" or co-occur in making correct classifications (*explainability question 5*), we compute novel measure of *Mutual Feature Interaction MFI(I,J) for features I and J* as a count of times features I and J appear on the same tree path making a correct prediction, for all trees in RF ensemble, and for all test samples. We show top 3 co-occurring features for each of the top 10 ranked features (see Fig. 2). Note that MFI only measures statistical pair-wise feature co-occurrences and not necessarily causality.

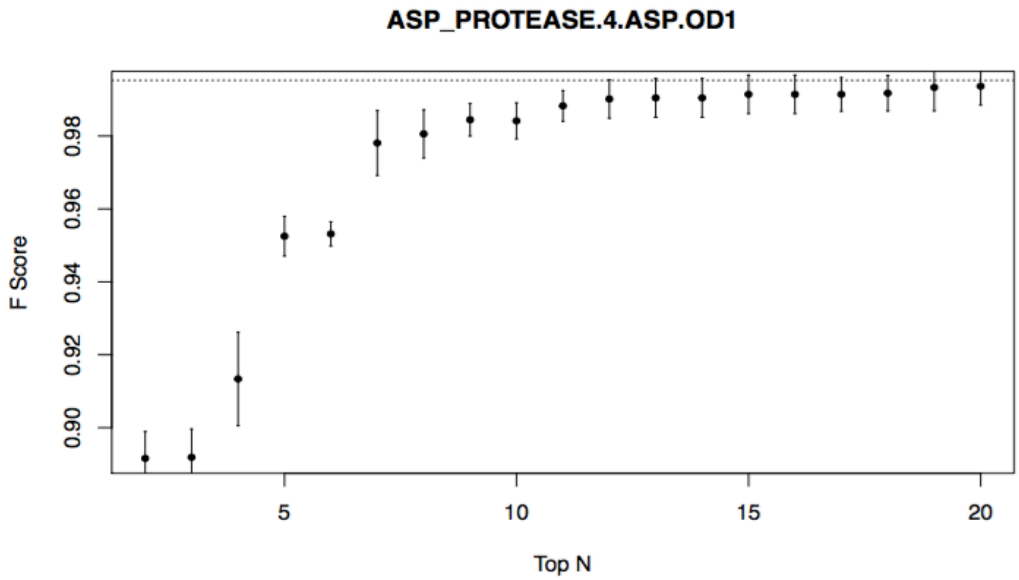
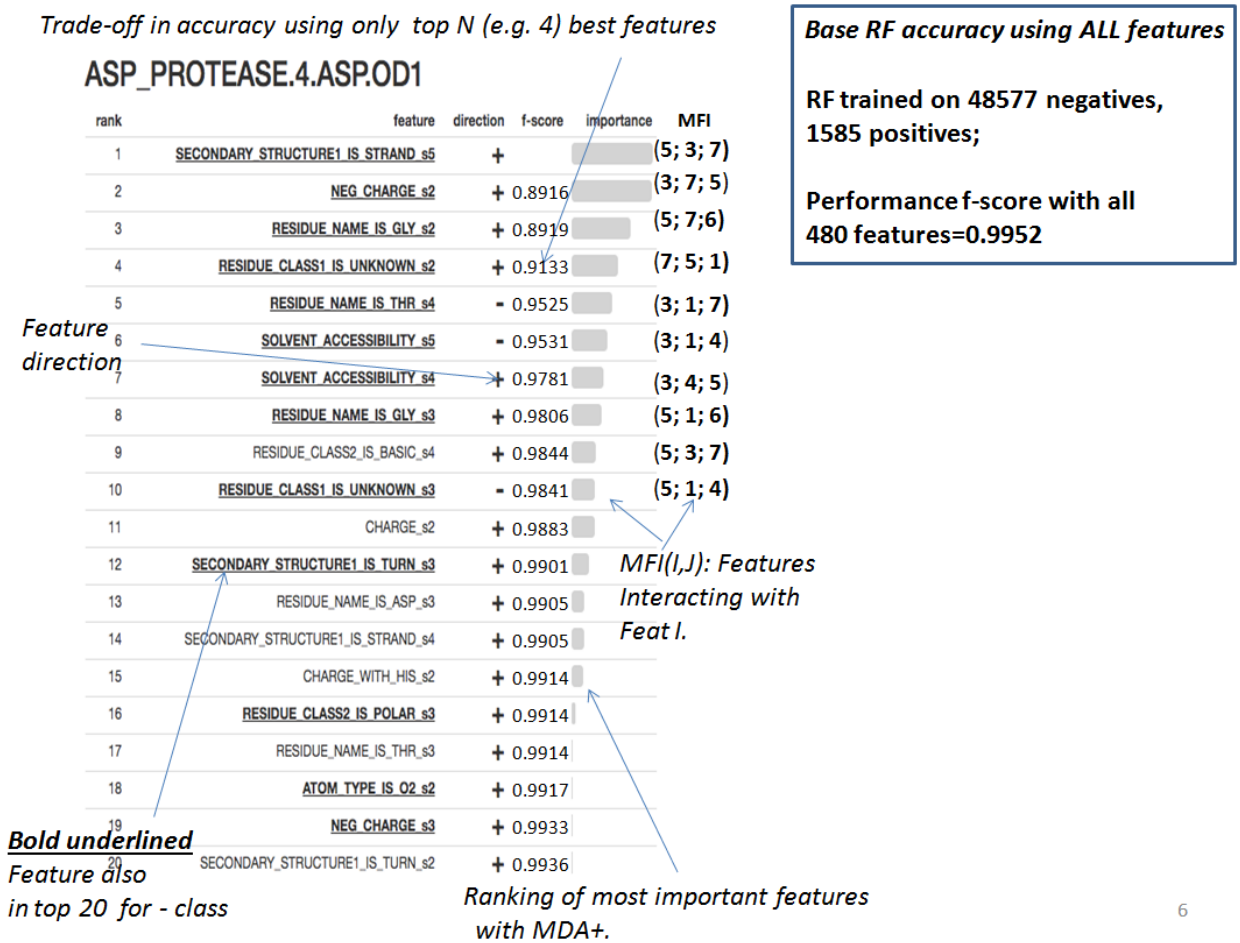


Fig.1 Trade-off of accuracy (average f-score and its variance from 5-fold CV) using Top N ranked subset of features (top 2, top 3 and so on), for positive class. Base accuracy using all 480 features is at dotted line

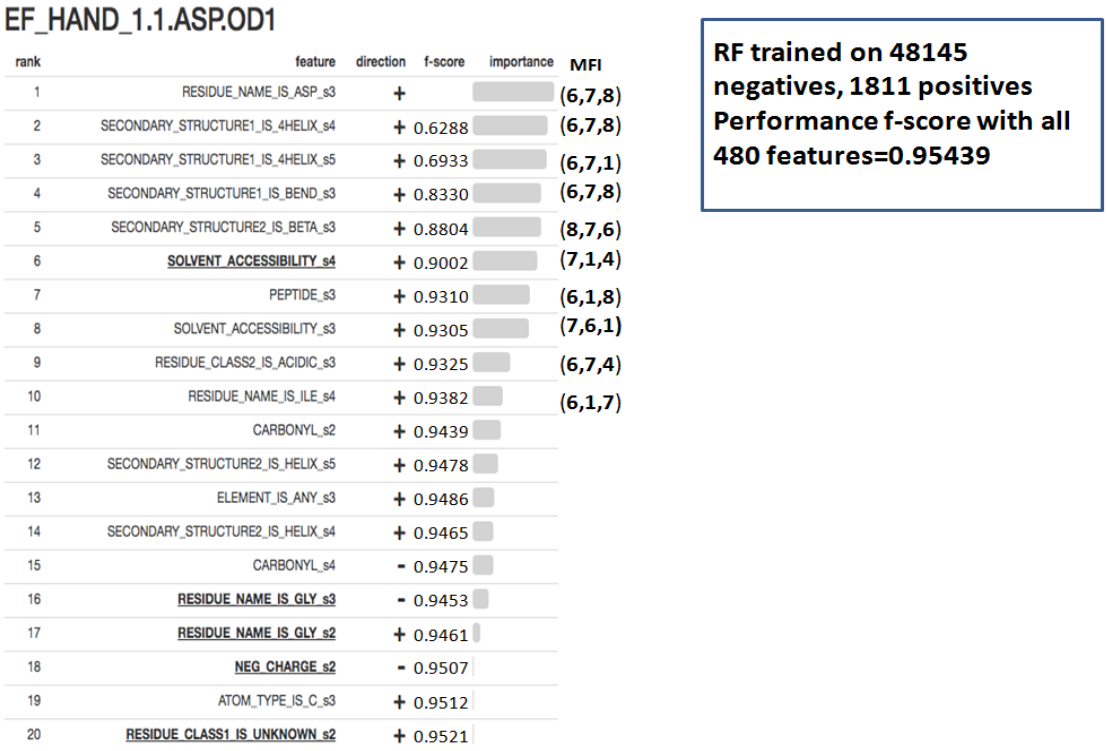
Finally, we carefully designed a *one page summary* RFEX report (*explainability question 6*) intended to be easy to read and interpret for expert and non-experts alike, and to answer all 6 explainability questions above. It is provided for positive and negative class separately. We show two RFEX summary reports, first for ASP_PROTEASE.4.ASP.OD1 (positive class) in Fig. 2, annotated with explanations of what elements relate to 6 explainability questions (in italics), and the second one for EF_HAND_1.1.ASP.OD1 (positive class) in Fig. 3. One way is to use RFEX summary report is to verify whether it matches known intuition or biochemical patterns already known (e.g. by looking for "presence" or abundance (marked +) or "absence" or deficiency (marked -) of highly ranked features. This in turn would increase users' confidence in RF predictions. Indeed, for the given two examples (ASP_PROTEASE, EF_HAND_1), there is evidence that the ranked features match our intuitive understanding of the active site structure, supported by the PROSITE [24] pattern matching. For ASP_PROTEASE, there is a required

glycine residue that is one amino acid away from the active site; the ranked list indicates that atoms belonging to glycine residue(s) 2.5 to 3.75 Ångstroms (shells S2 and S3; *RESIDUE_NAME_IS_GLY_s2, RESIDUE_NAME_IS_GLY_s3*) are a positive predictor, and negative for background. For EF_HAND_1, alpha helix residues near the coordinating ASP residue is part of the motif definition. Reassuringly this rule contributes as two of the top 3 features for positive prediction (*SECONDARY_STRUCTURE_IS_4HELIX_s4*, *SECONDARY_STRUCTURE_IS_4HELIX_s5*). Another way of interpreting RFEX in general is to look at highly ranked features and use their presence or absence to indicate main predictive factors, which potentially can bring new insights (the approach we used in [22]). Finally, one can *easily and efficiently* use RFEX summary reports to explore tradeoffs between number of features used (with their names and direction) and classification accuracy by looking at f-score column for RFEX summary reports. This shows that for all 7 investigated FEATURE models (except for EF_HAND_1.9.GLN.NE2 which had only 15 training samples), it suffices to use only from 2-6 (depending on a model) top ranked features to achieve 90% or better of the accuracy (f-score) when all 480 features are used.



rank	feature	direction	f-score	importance	MFI
1	SECONDARY_STRUCTURE1_IS_STRAND_s5	+			(5; 3; 7)
2	NEG_CHARGE_s2	+	0.8916		(3; 7; 5)
3	RESIDUE_NAME_IS_GLY_s2	+	0.8919		(5; 7;6)
4	RESIDUE_CLASS1_IS_UNKNOWN_s2	+	0.9133		(7; 5; 1)
5	RESIDUE_NAME_IS_THR_s4	-	0.9525		(3; 1; 7)
6	SOLVENT_ACCESSIBILITY_s5	-	0.9531		(3; 1; 4)
7	SOLVENT_ACCESSIBILITY_s4	+	0.9781		(3; 4; 5)
8	RESIDUE_NAME_IS_GLY_s3	+	0.9806		(5; 1; 6)
9	RESIDUE_CLASS2_IS_BASIC_s4	+	0.9844		(5; 3; 7)
10	RESIDUE_CLASS1_IS_UNKNOWN_s3	-	0.9841		(5; 1; 4)
11	CHARGE_s2	+	0.9883		
12	SECONDARY_STRUCTURE1_IS_TURN_s3	+	0.9901		
13	RESIDUE_NAME_IS_ASP_s3	+	0.9905		
14	SECONDARY_STRUCTURE1_IS_STRAND_s4	+	0.9905		
15	CHARGE_WITH_HIS_s2	+	0.9914		
16	RESIDUE_CLASS2_IS_POLAR_s3	+	0.9914		
17	RESIDUE_NAME_IS_THR_s3	+	0.9914		
18	ATOM_TYPE_IS_O2_s2	+	0.9917		
19	NEG_CHARGE_s3	+	0.9933		
20	SECONDARY_STRUCTURE1_IS_TURN_s2	+	0.9936		

Fig 2. RFEX one page summary report for ASP_PROTEASE.4.ASP.OD1, with explanation of graphical elements as they relate to explainability questions



EF_HAND_1.1.ASP.OD1

rank	feature	direction	f-score	importance	MFI
1	RESIDUE_NAME_IS_ASP_s3	+			(6,7,8)
2	SECONDARY_STRUCTURE1_IS_4HELIX_s4	+	0.6288		(6,7,8)
3	SECONDARY_STRUCTURE1_IS_4HELIX_s5	+	0.6933		(6,7,1)
4	SECONDARY_STRUCTURE1_IS_BEND_s3	+	0.8330		(6,7,8)
5	SECONDARY_STRUCTURE2_IS_BETA_s3	+	0.8804		(8,7,6)
6	SOLVENT_ACCESSIBILITY_s4	+	0.9002		(7,1,4)
7	PEPTIDE_s3	+	0.9310		(6,1,8)
8	SOLVENT_ACCESSIBILITY_s3	+	0.9305		(7,6,1)
9	RESIDUE_CLASS2_IS_ACIDIC_s3	+	0.9325		(6,7,4)
10	RESIDUE_NAME_IS_ILE_s4	+	0.9382		(6,1,7)
11	CARBONYL_s2	+	0.9439		
12	SECONDARY_STRUCTURE2_IS_HELIX_s5	+	0.9478		
13	ELEMENT_IS_ANY_s3	+	0.9486		
14	SECONDARY_STRUCTURE2_IS_HELIX_s4	+	0.9465		
15	CARBONYL_s4	-	0.9475		
16	RESIDUE_NAME_IS_GLY_s3	-	0.9453		
17	RESIDUE_NAME_IS_GLY_s2	+	0.9461		
18	NEG_CHARGE_s2	-	0.9507		
19	ATOM_TYPE_IS_C_s3	+	0.9512		
20	RESIDUE_CLASS1_IS_UNKNOWN_s2	+	0.9521		

Fig 3. RFEX one page summary report for EF_HAND_1.1.ASP.OD1

3. RFEX Usability Evaluation

The goals of RFEX usability evaluation were to assess: a) the *increase* of explainability e.g. users' confidence and understanding of how and why RF works by using our RFEX approach compared to traditional methods of presenting RF classification results; b) obtain users' feedback on the utility of each of the RFEX explainability summary report features. The usability evaluation was anonymous and was performed by users on their own time and place, based on the package of information and usability questionnaire sent to them with 11 questions. There were 13 users of varied experience in FEATURE and RF. User skill level in RF and FEATURE was assessed by users rating their expertise with 4 and 5 (e.g. "expert" level), or 1,2,3 (e.g. "non-expert"). Users in this study were grouped in 4 groups: a) FEATURE and RF NON-experts (4 users) ; b) FEATURE experts, RF non-experts (3 users); c) FEATURE non-experts, RF experts (2 users); and d) FEATURE experts, RF Experts (4 users). Group a) in some way corresponds to high level management or general public; group b) to bio scientists who are not versed in computational ML like RF; and group d) to computational bio scientists versed in biology (e.g. FEATURE) domain as well as ML (e.g. RF). Users were first asked to review *Exhibit A* (traditional ways of presenting RF results on FEATURE as in our prior work in [5]) for two models - ASP_PROTEASE.4.ASP.OD1 and EF_HAND_1.1.ASP.OD1) and assess their understanding of how and why RF works. Users were then given RFEX one page explainability summary reports for the above models (*Exhibit B*) and then asked to rate any *gain* in confidence and understanding

of why and how RF works. Users were also asked to grade usefulness of particular RFEX summary report graphical and presentation features, as well as assess RFEX applicability to other RF and ML applications. In Table 3 below we show averages of answers to most important usability questions (4 out of 11), grouped by user expertise level as explained above (higher number indicates better rating).

Table 3: Average of user responses to 4 most important usability questions for various user groups

Question	ALL users (13)	FEATURE and RF NON-experts (4)	FEATURE experts and RF NON-experts (3)	FEATURE NON-experts and RF experts (2)	FEATURE and RF experts (4)
Estimate your *increase in confidence of* RF classification *of FEATURE* data after using RFEX summaries	2.7 (SD 2.2)	2.5	2	0.5	4.5
Estimate your *increase in understanding why and how RF works on FEATURE* data (e.g. RF Explainability) using RFEX one page summaries	3.3 (SD 1.7)	3.25	3.7	1	4.25
I believe RFEX approach will be useful for other applications of RF	4.4 (SD 0.5)	4.5	4.3	4	4.5
I believe RFEX approach (or its modifications) will be useful for other machine learning methods	4.0 (SD 0.8)	4.5	3.3	3.5	4.25

For *all 13 users*, increase in understanding was significant (average of 3.3 for second question on a scale of 1…5) and increase in confidence in RF was good (average of 2.7 for first question). *All users* rated usefulness of RFEX method for other applications of RF and other ML approaches with 4 or 5, indicating strong RFEX promise at least at the conceptual level (averages for third and fourth questions were 4.4 and 4.0 respectively with small standard deviation). Analysis of usefulness of each feature of RFEX reporting where *all users* graded (not rated) them on a scale 1 to 5, indicated that most useful features were indeed those used to guide our design: one-page RFEX explainability summary design; feature ranking; presenting tradeoffs between number of features used and accuracy. In fact, all RFEX presentation features except thresholds used to test for abundance or deficiency of features were graded as above 3.7 in their usefulness. Positive user feedback on specific visualization format of RFEX summary points to importance of careful user-centered design for general ML explainability. Users also preferred to see up to top 3 Mutual Feature Interactions (MFI). The biggest increase both in confidence and understanding of how RF works on FEATURE data was achieved by user group *d) (FEATURE and RF experts e.g. "Computational bio-scientists"),* with averages of 4.5 and 4.25 on third and fourth questions. Users in group a) (FEATURE and RF NON-experts *e.g. "Managers or general public")* showed significant increase in understanding (average of 3.25 on second question) and good increase in confidence (average of 2.5 on first question). Users in group b) *(FEATURE experts and RF NON-experts e.g. "Bio scientists not versed in RF)"* also showed strong increase in understanding of why and how RF works (average of 3.7 on second question) but moderate increase in overall

confidence (average 2 on first question) which in part could be explained by their lack of RF expertise.

4. Conclusions and Future Work

Our RFEX method focuses on enhancing Random Forest (RF) classifier explainability by augmenting traditional information on RF classification results with one page RFEX summary report which is easy to interpret by users of various levels of expertise. RFEX method was designed and evaluated by never used before *user-centered-approach* driven by explainability questions and requirements collected from discussions with interested practitioners. It was implemented and extensively tested on Stanford FEATURE data. To assess usefulness of RFEX method for users, we performed formal usability testing with 13 expert and non-expert users which indicated its usefulness in increasing epxlainability and user confidence in RF classification on FEATURE data. Notably, RFEX summary reports easily reveal that one needs very few top ranked features (from 2-6 depending on a model) to achieve 90% or better of the accuracy achieved when all 480 features are used. Based on user feedback and our analysis we believe RFEX approach is directly applicable for other RF applications and to other ML methods where some form of feature ranking is available. Our future work includes applying RFEX on other RF applications and creation of a toolkit to automate RFEX creation.

Acknowledgments

We thank Prof. L. Kobzik, Dr. L. Buturovic and Prof. K. Okada for valuable feedback, J. Yang for helping with paper graphics and T. Murray and J. Schwartz for help in organizing our usability study. We also thank 13 anonymous usability reviewers. The work has partially been supported by NIH grant R01 LM005652 and by SFSU Center for Computing for Life Sciences.

References

1. DARPA program on Explainable AI (http://www.darpa.mil/program/explainable-artificial-intelligence), downloaded 07/04/17
2. Duggirala HJ, et al: "Use of data mining at the Food and Drug Administration", J Am Med Inform Assoc 2016; **23**:428-434
3. S. Kaufman, S. Rosset, C. Perlich: "Leakage in Data Mining: Formulation, Detection, and Avoidance", ACM Transactions on Knowledge Discovery from Data 6(4):**1-21**, December 2012
4. L. Breiman, "Random forests," Machine Learning, vol. **45**, no. 1, pp.5–32, 2001
5. K. Okada, L. Flores, M. Wong, D. Petkovic: "Microenvironment-Based Protein Function Analysis by Random Forest", Proc. ICPR (International Conference on Pattern Recognition), Stockholm, 2014
6. Liaw and M. Wiener, "Classification and regression by randomforest," R News, vol. **2**, no. 3, pp. 18–22, 2002. [Online],: http://CRAN.R-project.org/doc/Rnews/
7. A.Vigil: "Building Explainable Random Forest Models with Applications in Protein Functional Analysis", MS Thesis, San Francisco State University, Computer Science Department, December 2016

8. H. Malik, I. Chowdhury, H. Tsou, Z. Jiang, A. Hassan: "Understanding the rationale for updating a function's comment", IEEE Int. Conf. on Software Maintenance, Oct 2008
9. D. Delen: "A comparative analysis of machine learning techniques for student retention management", Decision Support Systems, Volume **49**, Issue 4, November 2010, Pages 498–506
10. J. Dale, L. Popescu, P. Karp:" Machine learning methods for metabolic pathway prediction", BMC Bioinformatics 2010, **11**:15
11. S. Cheng: "Unboxing the Random Forest Classifier: The Threshold Distributions", Airbnb Engineering and Data Science, https://medium.com/airbnb-engineering/unboxing-the-random-forest-classifier-the-threshold-distributions-22ea2bb58ea6, downloaded 07/04/17
12. C. Kelly, K. Okada: "Variable Interaction measures with Random Forest Classifiers", IEEE Int. Symposium on Biomedical Imaging**,** ISBI 2012
13. M. Mashayekhi, , R. Gras:"Rule Extraction from Random Forest: the RF+HC Methods", Advances in Artificial Intelligence, Volume **9091** of the series Lecture Notes in Computer Science pp 223-237, 29 April 2015
14. S. Liu, S. Dissanayake, S. Patel, X Dang, , T. Milsna, Y. Chen, D. Wilkins, D.:" Learning accurate and interpretable models based on regularized random forests regression", *BMC Systems Biology*, ***8***(Suppl 3), S5, 2014
15. S. Naphaporn, S, Sinthupinyo:" Integration of Rules from a Random Forest ", 2011 International Conference on Information and Electronics Engineering ,IPCSIT vol.**6**, 2011, Singapore
16. S. Hara, K. Hayashi: "Making Tree Ensembles Interpretable", ICML Workshop on Human Interpretability in Machine Learning , WHI 2016, NY, USA
17. L. Phung, V. Chau, N. Phung:" Extracting Rule RF in Educational Data Classification: From a Random Forest to Interpretable Refined Rules", Int. Conf. on Advanced Computing and Applications (ACOMP), 2015
18. L. Buturovic, M. Wong, G. Tang, R. Altman, D. Petkovic: "High precision prediction of functional sites in protein structures", PLoS ONE **9**(3): e91240. doi:10.1371/journal.pone.0091240
19. L. Wei and R. B. Altman, "Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function," J. Bioinform Comput Biol., vol. **1**, no. 1, pp. 119–38, 2003
20. R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. 2013. http://www.R-project.org
21. M. Kuhn: "The caret package", https://topepo.github.io/caret/, downloaded 07-04-17
22. D. Petkovic, M. Sosnick-Pérez, K. Okada, R. Todtenhoefer, S. Huang, N. Miglani, A. Vigil: "Using the Random Forest Classifier to Assess and Predict Student Learning of Software Engineering Teamwork" Frontiers in Education FIE **2016**, Erie, PA, 2016
23. B. Aeverman, J. McCorison et al.:"Production of Preliminary Quality Control Pipeline for Single Niclei RNA-SQ and its Application in the Analysis of Cell Type Diversity of Post-Mortem Human Brain Neocortex", PSB 2017, January 2017, Hawaii

C. Sigrist, E. de Castro, L. Cerutti, B. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios: *"New and continuing developments at PROSITE ",*Nucleic Acids Res. **2012**; doi: 10.1093/nar/gks1067, PubMed: 23161676

Tree-based Methods for Characterizing Tumor Density Heterogeneity

Katherine Shoemaker
Statistics Department, Rice University
Houston, Texas, 77005, USA

Brian P. Hobbs
Biostatistics, MD Anderson Cancer Center
Houston, Texas, 77030, USA
E-mail: bphobbs@mdanderson.org

Karthik Bharath
School of Mathematical Sciences, University of Nottingham
Nottingham, NG7 2RD, UK

Chaan S. Ng
Radiology, MD Anderson Cancer Center
Houston, Texas, 77030, USA

Veerabhadran Baladandayuthapani
Biostatistics, MD Anderson Cancer Center
Houston, Texas, 77030, USA

Solid lesions emerge within diverse tissue environments making their characterization and diagnosis a challenge. With the advent of cancer radiomics, a variety of techniques have been developed to transform images into quantifiable feature sets producing summary statistics that describe the morphology and texture of solid masses. Relying on empirical distribution summaries as well as grey-level co-occurrence statistics, several approaches have been devised to characterize tissue density heterogeneity. This article proposes a novel decision-tree based approach which quantifies the tissue density heterogeneity of a given lesion through its resultant distribution of tree-structured dissimilarity metrics computed with least common ancestor trees under repeated pixel re-sampling. The methodology, based on statistics derived from Galton-Watson trees, produces metrics that are minimally correlated with existing features, adding new information to the feature space and improving quantitative characterization of the extent to which a CT image conveys heterogeneous density distribution. We demonstrate its practical application through a diagnostic study of adrenal lesions. Integrating the proposed with existing features identifies classifiers of three important lesion types; malignant from benign (AUC = 0.78), functioning from non-functioning (AUC = 0.93) and calcified from non-calcified (AUC of 1).

Keywords: Radiomics, Imaging Features, Heterogeneity, Galton-Watson Trees

1. Introduction

One of the critical aspects to the study of solid lesions is intra-tumor heterogeneity (ITH). Solid lesions are often heterogeneous phenotypically, physiologically, and genetically, due to variations in processes such as cell proliferation, cell death, and local environmental factors.[1–3] Cellular diagnostic techniques such as biopsies are not only invasive, but they also do not allow for a thorough or complete investigation of the entire tumor environment. In order to get a more comprehensive picture of the entire lesion environment without having to take multiple biopsies or depend on qualitative visual assessments, quantitative imaging features can be mined with analytical techniques often called radiomics.[4] These radiomic features are objectively assessed and quantitatively descriptive of the lesion phenotypes and can be used to develop models that can be used in prediction, classification or diagnosis. The "radiomics hypothesis" that is central to this strategy is that advanced levels of analytics on imaging data can capture information that would not otherwise be available.[5] It has been hypothesized that this information on phenotypic patterns are reflective of complementary tumor characteristics at molecular, cellular and genetic levels.[6]

There are many possible ways to extract this radiomic data from routine images: features that describe size and shape, features that describe the intensity of and relationship between pixel values, textures features, and fractal features.[6] The ability to access a large number of quantitative features from images is now possible due to the progress made in imaging techniques, but issues like high levels of correlation between these features have led to the need to determine which of these features to use in downstream analyses and interpretation. As with any big data problem, when working with such a large number of variables, it is imperative to balance interpretability vs. computational tractability.

Currently the most commonly used radiomic features are texture-seeking. They can be divided into two categories, intensity and texture. Intensity features capture the shape of the histogram of the pixel values, while texture features describe the spatial distribution and pattern of the pixel values.[7] Texture-seeking methods can be divided into four categories: Non-spatial methods (NSM), Spatial Grey Level methods (SGLM), Fractal Analysis and Filters and Transforms. As NSM and SGLM are the methods used in the majority of analysis, we will focus in on those and give a brief description before highlighting potential new features derived from the tumor heterogeneity trees (THT).

NSM are intensity based features, comprised of first order statistics computed on the image grey level data. These are basic metrics, which include metrics such as the first order features, taken from the grey level image, include minimum and maximum, as well as computations such as range, mean, standard deviation, variance, median, skewness, kurtosis, entropy, root mean square (RMS) and total energy.[7]

The most frequently used of the central moments are variance, skewness and kurtosis.[8] Variation gives an idea of the size of the spread of the distribution around the mean, skewness is a measure of asymmetry around the mean, and kurtosis is a measure of the sharpness of the histogram. Intensity features give insight into how the pixel densities are distributed, but cannot give insight into their relative spatial positions, which limits their potential for describing the texture features of the image. The size of the images is a confounding factor for

several of these metrics, but the simplicity of these metrics is an advantage, and they contain a nontrivial amount of information about the image.

SGLM are texture features that are used to interrogate the spatial relationships between the grey levels of the image. A Grey Level Co-occurrence Matrix (GLCM) is an object that describes the spatial relationship of the grey levels of pixels in an image by counting the number of times two grey level valued pixels appear a specific distance and angle from each other. Before calculating the matrix, the number of grey levels must be chosen. This is done while considering the level of detail desired, the number of unique pixel values available, and the distribution of the density of these pixel values. These grey level "buckets" can be of consistent size, but the choice can alternatively be made to split the pixel value distribution along percentiles. This approach is a potential way to account for outlier pixel values within the pixel distribution. For example, the grey level matrix of angle 0°and distance 1, the cell (i, j) will contain the number of times a pixel of grey level j appears immediately to the right of a pixel of grey level i. This matrix can be made symmetric or not, and can be normalized by dividing each count by the number of pixels in the image. The angle can vary, allowing for comparison vertically or along the diagonal as well as the horizontal example given previously. The distance can be changed as well, with no restriction beyond the size of the image along the chosen angle.

GLCM features were first proposed by Haralick in 1973[9] and these features are easily computable from the GLCM matrix and include features that measure attributes such as image coarseness, symmetry, energy, and heterogeneity.[7] These features are often computed on multiple combinations of angle, distance and number of grey levels, leading to a large set of features that can be used in model building and data analysis.

While the above metrics work well in capturing the morphological characteristics of the tumor, they limited in their characterization of IHT. In consideration of lesion heterogeneity, tree-structured objects offer hierarchical dissimilarity processes that may better reflect the "relationship" between the pixels as a representation of the cellular evolution of cancer and of the lesion as it grows and develops. It is well-established that cancer as a disease starts at a "single point," a cell, which divides and proliferates outward to an extent that is allowed by the local immune and tissue environments. Each cell division is a biological bifurcation and this process is repeated up to the moment of the diagnostic scan capturing the cross-sectional state of the tumor. Conceptually, malignant cell proliferation is well characterized by a binary decision tree, which describes a hierarchical splitting process that divides iteratively from a common root until arriving at the final state of nodes or leaves. Considering tissue density as a surrogate[10] for the cellular division process, the growth process of a tumor may be well characterized by dissimilarity measures of pixel intensities obtained from tree-structured objects.

With this in mind, it is the goal of this paper to briefly discuss some of the various methods used to interrogate tumor texture, and to present a potential additional method based on tree-based analysis of the lesions, which will be used in conjunction with the currently used methods on a set of solid adrenal lesions to capture various aspects of cancer progression and development. Specifically, we use the feature set to classify benign from malignant, functioning

from non-functioning, and calcified from non-calcified, lesions with encouraging results.

2. Methods

Trees are a data type that is a specific subset of graphs. They are a directed, acyclic set of linked nodes that are connected by edges. The parental node is called the root, while the terminal nodes are called leaves. Depending on construction, they can either start at the leaves and repeatedly combine pairs (in a binary tree) until all leaves are grouped together, or start at the root and divide until each leaf is separate from all others. This branching process shows the relationship between the leaves and the history of how they separated from each other, when and in what order. If this process is applied to pixels from an image, and we consider that a radiological image is a representation the cells present inside the body, the tree can give a "history" of the representation of these cells and how they have divided from an original source, to a reasonable degree. In cancer, the pattern of growth is critical to the lesion development and this growth can be affected by the cellular environment, and is reflective of the ITH. A goal of introducing this feature is capturing this ITH via building the tree-based relationship between the pixels as intermediates for cells. Figure 2 gives a high level summary of the steps for creating tree-based features from radiological images which we describe in ensuing sections.

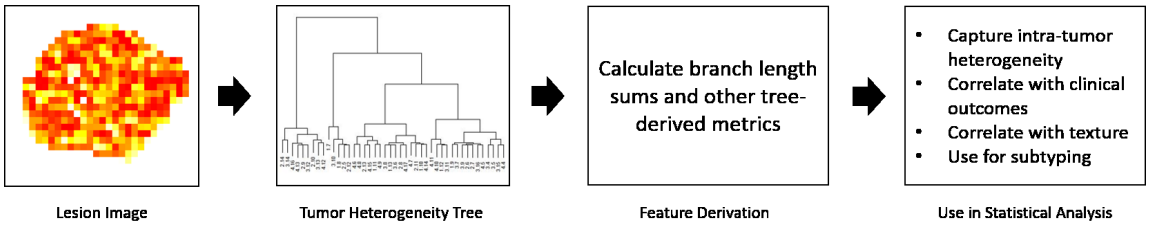


Fig. 1. A pictorial description of the methodology for extracting the tree-based feature metrics from a radiological image: First, a hierarchical tree is built from the image. Then, the branch length sums and other derived metrics are calculated from the tree. These metrics capture ITH and then can be used in further statistical analytics.

2.1. *Constructing Trees from pixel-level ROI data*

The mathematical objects underlying the first three aspects of Figure 2 are further diagrammed in Equation 1. THT are constructed from radiological images, from which Least-Common Ancestor (LCA) trees are drawn. This sample of LCA trees can then be summarized into metrics to be used as features such as the ones outlined previously, to join these other features in modeling and analysis.

$$\underbrace{\mathcal{I}_i}_{\text{Radiologic Image}} \rightarrow \underbrace{\mathcal{T}_i}_{\text{Tumor Heterogeneity Tree}} \rightarrow \underbrace{c_i(k)}_{\text{LCA Tree}} \rightarrow \underbrace{s_i}_{\substack{\text{Branch length sum,} \\ \text{a tree-based metric}}} \tag{1}$$

To construct the THT, we follow the work of Bharath et al,[11] and denote a rooted finite tree with n vertices as τ_n, where τ_n is a point in the space $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$. $\mathcal{T}_n$ is the set of all

finite trees on n vertices. A convenient notation for the tree is $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$, where $\mathcal{V}(\tau_n) = (root, v_1, \ldots, v_{2n-1})$, the set of vertices and $\mathcal{E} = (e_1, \ldots, e_{2n-1})$, the set of edges. Note that τ_n denotes a tree with n vertices, including the root and $\tau(n)$ denotes a tree with n terminal vertices.[11,12] The tree τ_n is not itself a probabilistic structure, so a stochastic process is placed on the growth of the tree in order to build a probabilistic model on the tree-structured data, and further steps are taken to provide a consistent family of densities. A Galton-Watson (GW) process $\{X_n\}_{n\geq 0}$ is a stochastic process that takes on positive integer values in discrete time, often used to model populations. It has an offspring distribution $(\pi_k, k = 0, 1, 2, \ldots)$. When this process is conditioned to have n vertices, the resulting tree τ_n is known as a conditioned GW tree.

These conditioned GW trees come from offspring distributions π_k, where k is equal to the number of leaves in the tree. To obtain information about variations in branch structure and to incorporate information about branch lengths, we must move to the Continuum Random Tree (CRT) through weak convergence. The CRT is the asymptotic limit of the GW tree, and in this limit, σ^2, the variance parameter of the GW tree's offspring distribution appears. Least common ancestor (LCA) trees are randomly chosen binary subtrees of conditioned GW trees, and can be understood as marginals of the CRT and provide dimension reduction. To create an LCA tree from τ_n, choose $k < n$ then uniformly choose k vertices from the n vertices of $\mathcal{V}(\tau_n)$. The density of the family of consistent CRT binary trees $C(k)$ from which these LCA trees with $k < n$ leaves are drawn is shown in equation 2.

$$f_{k,\sigma^2}(c(k)) = \left[\prod_{i=1}^{k-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{k-1}} (\sigma^2)^k s \exp\left(\frac{-s^2\sigma^2}{2}\right) \tag{2}$$

In summation, properties of the CRT allow us to use this density to approximate the density from which the LCA tree from any conditioned GW tree is drawn. The σ^2 term, gained by taking the CRT of the tree, captures variability in the branching process between different GW trees, while the LCA tree provides the ability to reduce the dimension of the data to s. Thus, for each image's tree $\mathcal{T}_i$, we create a LCA-tree $c_i(k_i)$ by choosing k_i of the n_i vertices, then calculate the value s_i by taking the sum of the lengths of the branches of LCA-tree. The density above has the kernel of a Gamma distribution with respect to s. Further, we know that s is non-negative, as the branch length components are non-negative, and these CRT branch lengths also asymptotically follow a Gamma distribution in this CRT construction of trees. As the sum of Gamma random variables is also Gamma, this allows for exploration of this feature in a generalized linear model setting. A full reasoning for the choice of the Gamma distribution on the trees can be found in K. Bharath et. al.

The trees produced from the images, as well as informative variables derived from these trees, will be the focus of the analysis in this paper. In practice, for each image, $\mathcal{I}_i$, hierarchical clustering is done on the vector of pixel densities, v_{ij}, where $j = 1, \ldots, n_i$, and n_i is equal to the number of pixels in image $\mathcal{I}_i$. The agglomerative clustering was done using the UPGMA (average) linkage method[13] and Euclidean distance to produce a tree, $\mathcal{T}_i$, from each image. Sensitivity analysis to the selection of distance metrics and clustering methods was performed. From this tree $\mathcal{T}_i$, a LCA tree $c_i(k)$ is randomly sampled and the branch length sum s from

$c_i(k)$ is calculated.

2.2. *Deriving metrics of ITH from tree representations*

In order to account for the randomness of the selection of leaves in the LCA trees, we randomly sampled 100-fold from the same image. The median value of the sum of the branch lengths and a measure of the spread of these values were collected as the variables of interest. This process is summarized and depicted in Figure 2, where the multi-modality of the empirical distribution highlights the need to take the median as the measure of center.

It is hypothesized that the edge sum value for each lesion can be a feature that is reflective of the ITH. A group of pixels that are more diverse will produce a tree that is taller; a tree that, for example, clusters somewhat quickly into various groups but then those groups do not merge into one cluster until much later. If an image has a large amount of density values that are similar, those will cluster quickly, leading to short branch lengths. A reflection of this hypothesis can be seen in the left hand column of Figure 3, a graph using images from the case study described below. Tumors with a large amount of similarly valued pixels have low branch length sums, while those that have sharp differences have higher median edge sum values. In fact, the lesion with the highest valued median edge sum has a large group of extremely dense pixels, surrounded by more moderately valued pixels. Trees produced from this lesion have very long branches from the split of the group and non-group pixels, which is reflected in its' very large branch sum value. While some of the difference in visual levels of heterogeneity can be explained by the pixel size of the images, there are differences in the small and large valued groups of the median.

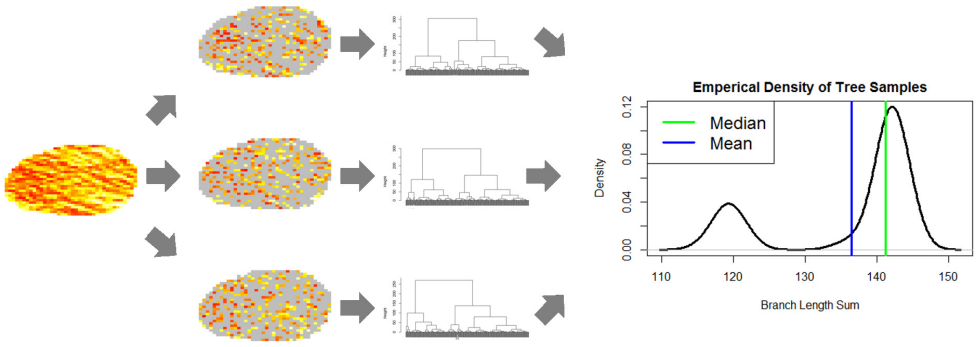


Fig. 2. A depiction of the LCA sampling process. To the left, the original image. Next, a 30% sample of pixels is taken, illustrated by the three images with only 30% of their pixels still in color. From each of these, a hierarchical tree is built, where subtle differences can be seen in the third set of images. The final image to the right is the empirical density plot of the 100 LCA samples taken from this image, and it provides an instance in which taking the median instead of the mean is important.

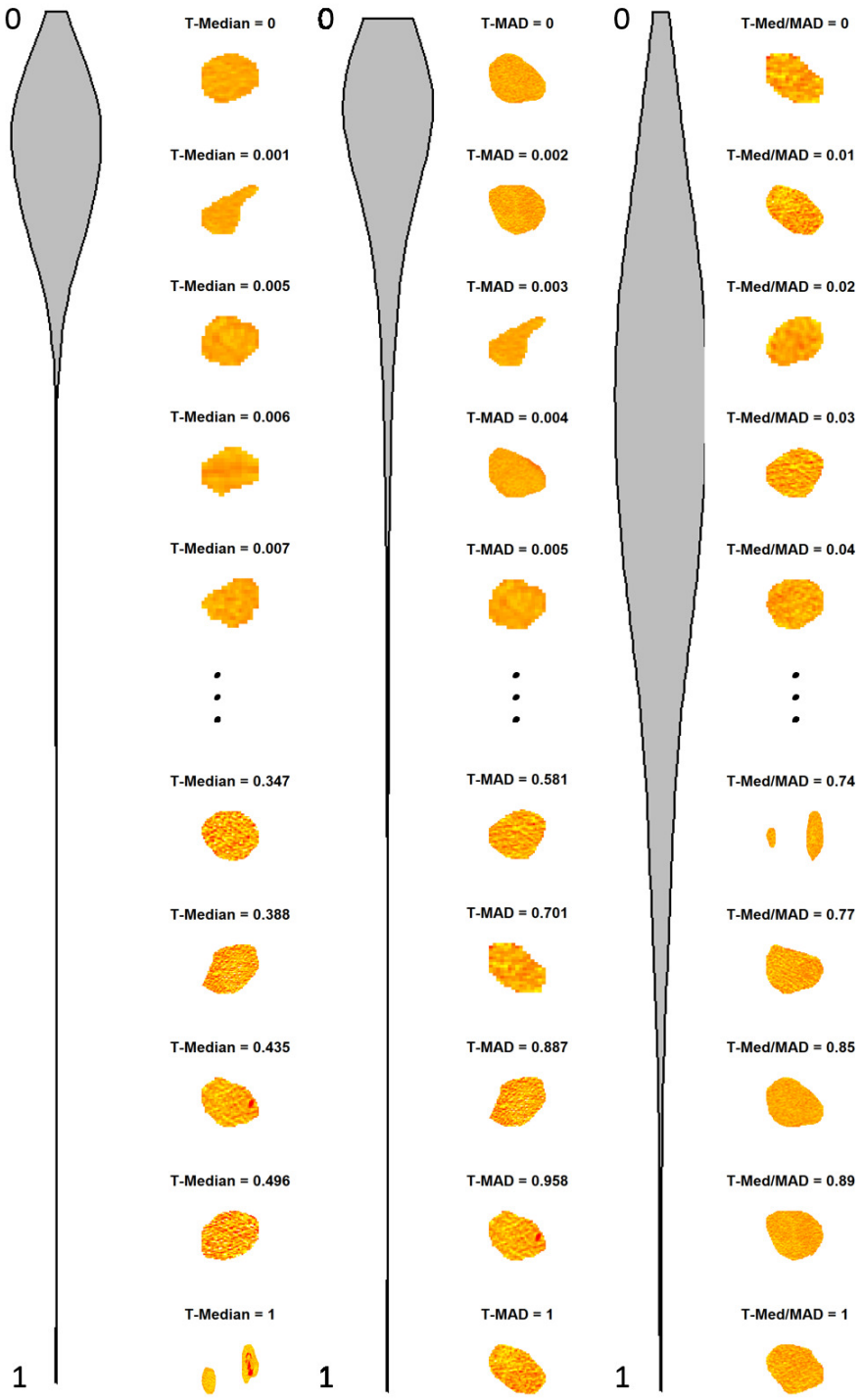


Fig. 3. From left to right, the violin plots show the empirical densities of the Median, Median Absolute Deviation (MAD), and Median/Median Absolute Deviation features for the adrenal lesions used in the case study below, normalized to be between 0 and 1. To the immediate right of the density for each feature are the lesion images for the five highest and lowest values of each, along with their normalized value. The image colors have been scaled so that the individual mean pixel intensities correspond to the same color across all images.

3. Application to Solid Adrenal Lesions

3.1. *Adrenal Lesions*

Adrenal masses are common and can be either functioning or non-functioning. Within each side of this divide, they can be either benign or cancerous. Cancerous lesions can be either first degree, primary, tumors or second degree, metastatic tumors. In patients without known cancer, these masses are often benign adenomas and of little clinical significance, but the proportion that are malignant increases slightly with previous knowledge of cancer[14] as well as with age.[15] A primary non-functioning tumor is a rare malignancy known as an adrenocortical carcinoma (ACC)[16] and a non-functioning benign lesion is an adenoma, which make up approximately 50% of the non-functional lesions.[17] There are many different types of functioning lesions, such as paragangliomas and pheochromocytomas, both of which can present as either benign or malignant.

3.2. *Data*

Our retrospective data consists of 379 CT scans from 356 patients. Of the lesions, 195 are malignant and are 182 benign, 334 are non-functioning lesions and 43 are functioning. For tumors that are metastatic, the information about the primary type of cancer and the timing is available as well. Their pathologies have been verified by the radiologist and are available in table form in the supplementary material, located at `http://kas23.web.rice.edu/PSB_supplementary_material.pdf`. There are 13 calcified lesions, 13 fatty lesions, 202 heterogeneous lesions and 134 homogeneous ones. The density of lesion size (calculated by pixel count) was heavily skewed to the right, with large lesions presenting as outliers. An unsupervised clustering was performed on the pixel size using k-nearest neighbors in order to produce a distinction between the main group of lesions and the large outlier lesions. To attempt to remove these outlier affects, this cluster of large tumors was not included in calculations based solely on the THT features.

Using the methodology described above, a GW tree was built for each adrenal lesion image using Matlab, averaging 0.98 seconds per tree. Then 100 branch length samples were taken for each of the lesions. This was done with a C++ program accessed through R. It took, on average, 4 seconds to compute one LCA sample of one tree. This was done on a computer with a 3.3 GHz processor and 16 GBs of RAM. The average number of pixels in an image was 2064 and the median number was 812. Several of the lesions' branch sum distributions were multi-modal, see Figure 2 as an example, so it was decided to take the median in place of the mean as the measure of central tendency to account for this. The median absolute deviation (MAD) of the samples was calculated to capture the spread of these samples.

The median branch length sum empirical density from this truncated group of lesions is plotted vertically in the far left violin bar in Figure 3, along with the densities of the MAD of the sample draws and the normalized feature. For all three, the curve is unimodal and varying degrees of right skewed, but the normalized feature on the right does present a smaller tail. As mentioned previously, a visual difference between the upper and lower groups can be seen, particularly in the median and the MAD features (the left and middle columns, respectively).

Note that when time is referenced, it is not in the usual temporal sense, but rather time within the tree similarity space. The group of lesions with small median values are very smooth with similarly valued pixels, leading to trees that go a long distance without branching out, with pixels tending to stay in the cluster instead of breaking apart. The group with large median values have large differences in color, some even with visible sections of pixels that are isolated and a much different value from the rest. Trees for an image such as this are going to have clusters that break apart quickly due to the large variation of intensity values present, leading to tall trees that have large branch length sums. The extreme outlier lesion at the bottom of the column is a perfect example of this, with a large cluster of high-density pixels that cause there to be a fast division into two primary clusters that stay clustered with themselves for a very long time. Lesions with low MAD values appear to be more homogeneous and uniform, likely from the similar pixels causing the sampled trees to be relatively similar as well, i.e., regardless of random sample taken as in Figure 2, the resulting tree is similar. Higher MAD values correspond to lesions with large visible pixel differences; the sampling of pixels from these lesions can make a large difference in the height of the resulting tree. While the normalized feature has less of a visible differential between the high and low image groupings, it contains the information from both other features and has the advantage that it is less correlated with the commonly used radiomics features.

3.3. *Connection with Other Radiomics Features*

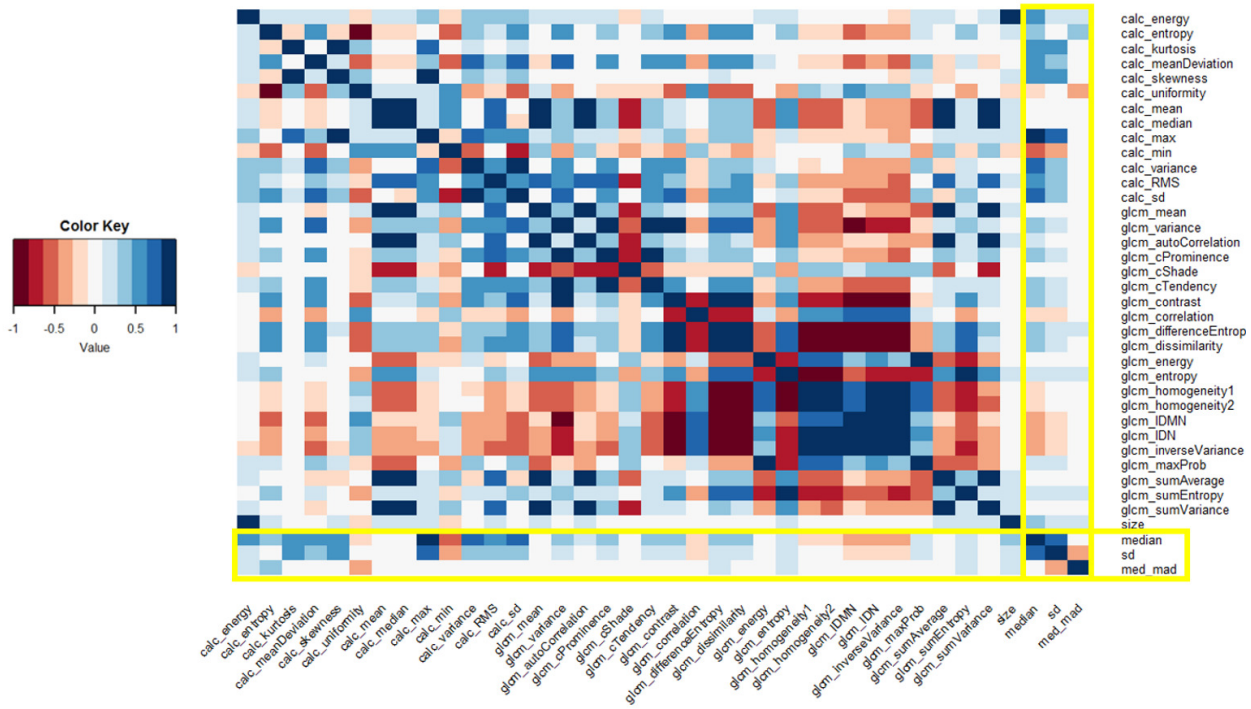


Fig. 4. A heatmap showing the correlation between a set of the NSM and SGLM features with the THT based features (highlighted in yellow).

When the group of small lesions' branch length sums are compared against each of the set of preexisting radiomics features, as in Figure 4, various relationships and lack of relationships can be observed. Less informative features such as minimum and maximum appear dramatically related with the branch lengths, but more expository features such as contrast, which measures local intensity variation, and entropy, which can be used to distinguish tissue with structure, also appear to be correlated with branch length. As can be seen in Figure 4, the median is slightly correlated with the other radiomics features. This correlation led to the decision to divide the median by the calculated MAD, and a normalized feature that was much less correlated with the other radiomics features emerged. This feature is used as the THT feature in further statistical analysis. Besides the THT feature computed for each lesion, the first order NSM features as well as the second order SGLM features were computed. For simplicity and ease of computation, only one GLCM was used for these features, the GLCM with distance 1 and angle of 0°. Thus, for each lesion, there were a total of 37 features collected using a combination of NSM, SGLM, and THT methods.

3.4. *Characterization and Classification*

In order to search for separation caused by the groups of features, a Principal Component Analysis (PCA) was performed on these 37 features in order to construct an orthogonal set of features. The components produced were further used in a 5-fold cross validated logistic regression on three qualitative features of the data set to determine the discriminatory abilities of these features when modeled conjointly. The feature loadings as well as the coefficients of the logistic regression can be found in Section 2 of the supplementary material.

We decided to use the first 6 PCs, as this was sufficient to explain 90% of the variance. The scree plot of the variances can be found as Figure S1 in the supplementary material. When each of these 6 PCs are plotted against each of the other components, the points cluster in a line with the outliers scattered to the side. As can be seen in Figure S2, especially in the 4th and 5th PCs, when compared against the various qualitative information available about the lesions, it was found that the majority of these outlier lesions were denoted by the radiologist as calcified. Calcification is typically distinguishable on CT scans, but doesn't signify one subtype of lesion over any other.[18] This delineation of the calcified lesions is likewise apparent in the results of the logistic regression, presented in Figure 5. The set of PCs has a very high level of accuracy for characterizing calcified lesions from non-calcified lesions (AUC = 1) and functioning from non-functioning (AUC = 0.93) and does moderately well at distinguishing malignant from benign (AUC = 0.78).

At a qualitative level, as can be seen above in Figure 3, a difference between the textural and visual presentation of the high and low value images along the top of the density plot can be seen for the median. This leads to the conclusion that this feature is capturing some aspect of the ITH. When looking at the lesions in the right column of Figure 3, there appears to be a difference in ITH. In general, while the lower valued images have larger particles of density clusters, the ones with higher values have a much finer grain of texture. At a quantitative level, the feature set was able to make a perfect characterization of calcified tumors (AUC of 1) and was highly accurate for determining functioning tumors (AUC of 0.93).

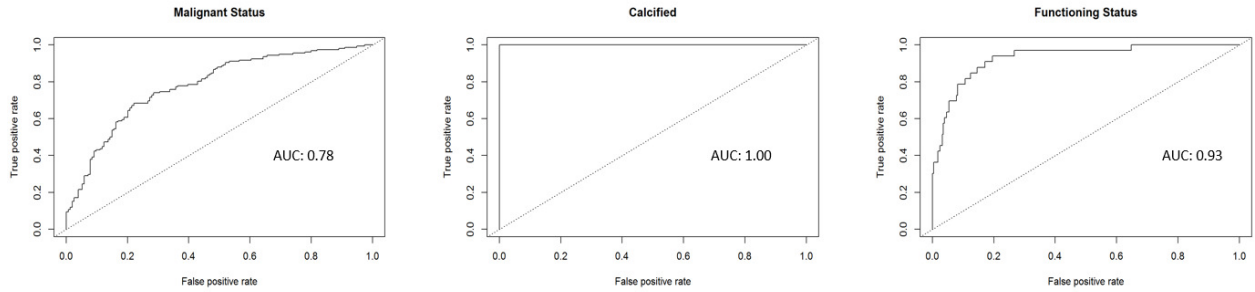


Fig. 5. Receiver Operating Curves for three endpoints chosen for discrimination. Perfect stratification was achieved in the calcified lesions, and near perfection for the functioning status.

4. Discussion

One of the foundational ideas behind radiomics is that analytics can detect nuances in tumors that a human eye might be unable to distinguish. By capturing unique aspects of ITH, as demonstrated through lack of correlation with existing quantitative image features, as well as yielding accurate tissue characterization in integrative analysis, THT-derived features represent a valuable contribution to the parameter domain of radiomics. By their formulation, tree-structured objects offer the potential to better reflect the biological and evolutionary processes that give rise to solid lesions. Summing the branch lengths of GW trees, the feature demonstrated in our case study, and interrogating their distributions under repeat sampling is straightforward and highly interpretable. THTs potentially access much broader information, however, whether that be features from the empirical distribution of the tree samples, the point process inherent to the branch breaking pattern, or using the rationale of the statistical grounding of the known distribution for the branch length sums. More exploration can be done, on this data set or others, to determine the full extent of the THT for producing features for characterization, classification or more.

Translation and dissemination of code is ongoing, available upon request. For further detail about the selection of the random LCA trees, see `www.github.com/pkambadu/DyckPaths`, where it is available under a BSD-style license

Acknowledgements

This work was supported by NIH R01-CA194391, NIH R01160736 and NSF 1463233 (to VB). KS was partially supported by NIH grant T32 - CA09652.

References

1. L. Alic, W. J. Niessen and J. F. Veenland, *PLOS ONE* **9**, 1 (10 2014).
2. M.-C. Asselin, J. P. OConnor, R. Boellaard, N. A. Thacker and A. Jackson, *European Journal of Cancer* **48**, 447 (2012).

3. A. Sottoriva, I. Spiteri, S. G. M. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts and S. Tavar, *Proceedings of the National Academy of Sciences* **110**, 4009 (2013).
4. H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, *Nature communications* **5** (2014).
5. P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker and H. J. Aerts, *European Journal of Cancer* **48**, 441 (2012).
6. S. S. F. Yip and H. J. W. L. Aerts, *Physics in medicine and biology* **61**, R150 (2016).
7. C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. U. Shankar, R. Kikinis, B. Haibe-Kains, P. Lambin and H. J. W. L. Aerts, *PLOS ONE* **9**, 1 (07 2014).
8. N. Aggarwal and R. K. Agrawal, *Journal of Signal and Information Processing* **3**, 146 (2012).
9. R. M. Haralick, K. Shanmugam *et al.*, *IEEE Transactions on systems, man, and cybernetics* , 610 (1973).
10. M. Fassnacht, M. Kroiss and B. Allolio, *The Journal of Clinical Endocrinology & Metabolism* **98**, p. 4551 (2013).
11. K. Bharath, P. Kambadur, D. K. Dey, A. Rao and V. Baladandayuthapani, *Journal of the American Statistical Association* (2016).
12. D. Aldous, *Ann. Probab.* **19**, 1 (01 1991).
13. R. R. Sokal and C. D. Michener, *University of Kansas Science Bulletin* **38**, 1409 (1958).
14. J. C. Miller, M. A. Blake and G. W. L. Boland, *BMJ* **338** (2009).
15. G. W. L. Boland, M. A. Blake, P. F. Hahn and W. W. Mayo-Smith, *Radiology* **249**, 756 (2008), PMID: 19011181.
16. E. Duregon, M. Volante, E. Bollito, M. Goia, C. Buttigliero, B. Zaggia, A. Berruti, G. V. Scagliotti and M. Papotti, *Human Pathology* **46**, 1799 (2015).
17. A. B. Grossman, *Nonfunctional Adrenal Masses*. Merck Manual.
18. P. J. Kenney and R. J. Stanley, *Urologic radiology* **9**, 9 (Dec 1988).

How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures?

Yogasudha Veturi[1] and Marylyn D. Ritchie[1]
[1]*Biomedical and Translational Informatics Institute, Geisinger*
Danville, PA
Email: yveturi@geisinger.edu, mdritchie@geisinger.edu

Transcriptome-wide association studies (TWAS) have recently been employed as an approach that can draw upon the advantages of genome-wide association studies (GWAS) and gene expression studies to identify genes associated with complex traits. Unlike standard GWAS, summary level data suffices for TWAS and offers improved statistical power. Two popular TWAS methods include either (a) imputing the *cis* genetic component of gene expression from smaller sized studies (using multi-SNP prediction or MP) into much larger effective sample sizes afforded by GWAS –- TWAS-MP or (b) using summary-based Mendelian randomization –- TWAS-SMR. Although these methods have been effective at detecting functional variants, it remains unclear how extensive variability in the genetic architecture of complex traits and diseases impacts TWAS results. Our goal was to investigate the different scenarios under which these methods yielded enough power to detect significant expression-trait associations. In this study, we conducted extensive simulations based on 6000 randomly chosen, unrelated Caucasian males from Geisinger's MyCode population to compare the power to detect *cis* expression-trait associations (within 500 kb of a gene) using the above-described approaches. To test TWAS across varying genetic backgrounds we simulated gene expression and phenotype using different quantitative trait loci per gene and *cis*-expression /trait heritability under genetic models that differentiate the effect of causality from that of pleiotropy. For each gene, on a training set ranging from 100 to 1000 individuals, we either (a) estimated regression coefficients with gene expression as the response using five different methods: LASSO, elastic net, Bayesian LASSO, Bayesian spike-slab, and Bayesian ridge regression or (b) performed eQTL analysis. We then sampled with replacement 50,000, 150,000, and 300,000 individuals respectively from the testing set of the remaining 5000 individuals and conducted GWAS on each set. Subsequently, we integrated the GWAS summary statistics derived from the testing set with the weights (or eQTLs) derived from the training set to identify expression-trait associations using (a) TWAS-MP (b) TWAS-SMR (c) eQTL-based GWAS, or (d) standalone GWAS. Finally, we examined the power to detect functionally relevant genes using the different approaches under the considered simulation scenarios. In general, we observed great similarities among TWAS-MP methods although the Bayesian methods resulted in improved power in comparison to LASSO and elastic net as the trait architecture grew more complex while training sample sizes and expression heritability remained small. Finally, we observed high power under causality but very low to moderate power under pleiotropy.

Keywords: TWAS, summary-based, SMR, expression-trait associations, power

1. Introduction

Genome-wide association studies (GWAS) have discovered a large number of variants associated with a host of complex traits and diseases[1]. However, these GWAS-significant variants explain a very limited proportion of the overall trait heritability, a phenomenon that is widely referred to as "missing heritability"[2]. Moreover, traditional GWAS have also largely ignored the relationship that exists between genetic variants, DNA functional elements (e.g. gene expression/protein levels) and complex traits and diseases. eQTL studies can help identify the extent of influence that a variant can have on gene expression. However, the extent to which this variant can *modulate* gene expression to also influence complex traits and diseases remains a topic of great interest in the genetics and public health community.

One way to address this question is to conduct studies in which both gene expression and trait measurements are available on the same set of individuals. However, such studies are extremely limited in number and are hampered by small sample sizes owing to the costs involved in data collection. Alternatively, one could combine the features of eQTL studies and GWAS (performed on different populations) to illuminate gene-trait relationships using a transcriptome-wide association study (TWAS). Such a study exploits the relationship between a genetic variant and gene expression as well as the large sample sizes afforded by GWAS to help identify novel gene-trait associations in a powerful manner.

Many "flavors" of TWAS have been published already[3–8]. These approaches include determining whether GWAS-significant variants are also enriched for eQTLs[3,4], detecting co-localization of expression signals at known GWAS loci[7], performing Mendelian Randomization using summary-statistics for gene expression-genotype and genotype-phenotype associations[9], and performing multi-SNP prediction (MP) analysis that can more explicitly model linkage disequilibrium (LD) when causal variants are not genotyped[5,8]. Additionally, TWAS-MP methods also use different regression models to "impute" *cis*-gene expression into much larger GWAS datasets; for instance, Gusev et al.[5] use the best linear unbiased predictor (BLUP) while PrediXcan[8] applies elastic net regression to achieve the same goal.

The type of data required by each of these approaches is also different; for instance, some methods require individual-level genotype and phenotype as well as gene expression data [e.g. TWAS-MP (elastic net) in PrediXcan[8]], while others only need summary-level data at one or both levels (e.g. TWAS-MP (elastic net) in MetaXcan[10], TWAS-MP (BLUP)[8], summary-based Mendelian Randomization or TWAS-SMR[9], and COLOC[7]). At the expense of introducing some bias, summary-based approaches can vastly improve computation efficiency. Some approaches also attempt to incorporate distinctions between different kinds of genetic models in their model assumptions. For instance, while TWAS-MP assumes either direct/indirect **causality** (when expression mediates between genotyped/non-genotyped SNP and trait) or **pleiotropy** (when the genetic variant has direct and independent effects on gene expression as well as the phenotype), TWAS-SMR distinguishes pleiotropy from linkage (in effect, when two causal variants that are in LD with each other independently influence either gene expression or phenotype) using a post-hoc method called heterogeneity in dependent instruments (HEIDI)[9].

Thus far, no study has compared the power (to detect gene-trait associations) of these methods under a range of complex genetic architectures. In this study, we compare the statistical power

afforded by TWAS-MP and TWAS-SMR in hitherto unexplored scenarios. This work can help us recognize genetic patterns underlying complex trait variation. We consider two different genetic models: causality and pleiotropy (as described above). We also investigate the influence on power of trait heritability, expression heritability, number of quantitative trait loci (QTL), sample size for training the imputation algorithm (relevant to TWAS-MP methods) and finally, the GWAS sample sizes. We compare different variable selection and shrinkage-based methods that can perform TWAS-MP (e.g. BLUP/Bayesian Ridge Regression, Bayesian LASSO, Bayesian spike-slab, elastic net and LASSO) to TWAS-SMR, GWAS, and eQTL-based GWAS (eGWAS). We have integrated Bayesian LASSO with TWAS for the first time in this study. Under the assumption of causality, TWAS-MP methods yielded the highest (and consistently identical) power under different simulation scenarios while TWAS-SMR, eGWAS and GWAS yielded consistently lower power. For TWAS-MP, Bayesian methods were at least as powerful as elastic net and LASSO, and surpassed their power as trait complexity increased, expression heritability remained low, and training sample size was small. Interestingly, we observed that traditional GWAS resulted in higher power than TWAS under the assumption of pleiotropy, although there was a massive overall loss in power from before.

2. Methods

In this section, we describe the data structure and quality control procedures, the simulation pipeline (modified from Gusev et al.[5]) as well as the statistical methods employed for calculating the power of detecting gene-trait associations.

2.1. Genotype Data

Individuals included in this simulation study came from a patient cohort in the MyCode® Community Health Initiative of Geisinger Health System[11]. We used participants that were genotyped using the Illumina Human Omni Express plus exome beadchip in the DiscovEHR study (a collaboration between Geisinger Health System and Regeneron Genetics Center). The genetic data was imputed using the Haplotype Reference Consortium panel and the dataset contained 60,000 individuals and approximately 600K variants after some initial quality control measures. For this analysis, we removed any related samples (up to 1st cousins) as well as those that did not pass a sample call rate filter of 90%. We filtered variants that did not pass a genotype call rate filter of 99% and a minor allele frequency filter of 1% (so as to restrict ourselves to common variants only). We finally selected at random 6000 males of European American ancestry to ensure as much homogeneity in the population as possible.

2.2. Simulation pipeline

2.2.1. Simulating gene expression

We started with 6000 randomly chosen unrelated European American males from the MyCode® population. We then sampled 100 genes at random from across the genome, each of length between 100 and 200 SNPs, as annotated using Biofilter[12]. We selected the region 100 kb upstream and

downstream of each chosen gene. We chose 5 different seeds per gene, giving us a total of 500 replications in the power simulation.

In each replication, we divided the total sample size into two sets: training (100, 250, 500, 1000 individuals each) and testing (5000 individuals). In each training set, we first simulated gene expression under an additive genetic model at each of four levels of causal variants per gene (*number of QTL* = 5%, 10%, 25% and 50%) as well as three levels of *cis*-expression heritability (*cis*-h_e^2 = 5%, 17% and 30%). The *cis*-h_e^2 levels were chosen based on their published distributions for significant (i.e. *cis*-h_e^2 >>0 by likelihood ratio test) *cis*-eQTLs in three different SNP-expression cohorts[5].

Let the sample size be represented by n, the number of SNPs by p and the number of QTL by m. The model to simulate gene expression can be expressed as follows:

$$E = X\beta + \varepsilon \quad (1)$$

where E is the $n\times1$ vector of standardized gene expression values for the n individuals in the training set, β is the $m\times1$ vector of marker effects for the m QTL in the gene and is drawn from a normal distribution with mean zero and variance *cis*-h_e^2, X is the $n\times m$ matrix of genotypes and ε is the vector of the normally distributed errors with mean zero and variance 1-*cis*-h_e^2.

2.2.2. Simulating phenotype

We simulated the phenotype in the testing set (5000 individuals) under eight different levels of trait heritability per gene (h^2 = 0%, 0.005%, 0.001%, 0.025%, 0.05%, 0.1%, 0.5%, 1%), wherein $h^2 = 0$ corresponded to the null model. In the testing set, two genetic models were used to simulate the phenotype: causality (when expression mediates the relationship between SNP and phenotype) and pleiotropy (when gene expression and phenotype independently share the same causal variant). Phenotypes using either genetic model were simulated under an additive genetic model as follows:

(1) Causality

$$Y = Eb_1 + \varepsilon_1 \quad (2)$$

where Y is the 5000×1 vector of the standardized response for the 5000 individuals in the testing set, E is the 5000×1 vector of gene expression values for testing set, b_1 is transcript effect drawn with zero mean and variance h^2, and ε_1 is vector of the normally distributed errors with mean zero and variance 1-h^2.

(1) Pleiotropy

$$Y = Xb_2 + \varepsilon_2 \quad (3)$$

where Y is the 5000×1 vector of the standardized response for the 5000 individuals in the testing set, X is the $5000\times m$ matrix of genotypes (same as those used to simulate gene expression), b_2 is the $m\times1$ vector of marker effects drawn from a normal distribution with mean zero and variance h^2, and ε_2 is vector of the normally distributed errors with mean zero and variance 1-h^2.

To reach precision corresponding to a large-sized GWAS, we repeated the phenotype generation with different environmental noise terms: 10 iterations resulted in a GWAS sample size of 50,000, 30 iterations resulted a GWAS sample size of 150,000 and 60 iterations resulted in a GWAS sample size of 300,000.

2.2.3. Power analysis

The following were the null and alternative hypotheses in this study:
H_0: There is no association between gene and phenotype; i.e. cis-$h_e^2 = 0$ or $h^2 = 0$
H_1: There is a non-zero association between gene and phenotype; i.e. cis-$h_e^2 > 0$ and $h^2 > 0$

In this study, we only considered the $h^2 = 0$ scenario as our null model. We first conducted eQTL analysis on the training set to identify the $p \times 1$ vector of z-scores (Z_{eQTL}) by regressing gene expression on the p SNPs in the chosen gene. Subsequently, we obtained p-values corresponding to expression-trait associations from 8 different models:

1. GWAS: For each GWAS set (50K, 150K, or 300K individuals), we conducted meta-analysis across the smaller sets to obtain a $p \times 1$vector of z-scores (Z_{GWAS}) and corresponding p-values for all SNP-trait associations. The gene was considered to be detected if at least one SNP in the gene had a p-value < 5E-8.
2. eGWAS: In this eQTL-based GWAS, we used the GWAS p-value of the single most significant SNP from eQTL analysis. The gene was considered to be detected if this p-value < 0.05/15,000 (where 15,000 corresponds to the number of genes across the genome).
3. TWAS-MP: This approach involves imputation of expression-trait association statistics directly into GWAS summary statistics and involves three different steps:

Obtaining weights: The first step here was to obtain estimated coefficients (weights obtained on regressing gene expression on SNPs) on the training set using five different penalized regression/Bayesian regularization approaches; elastic net, LASSO, Bayesian ridge regression (BRR), Bayesian LASSO (BL) and Bayesian spike slab or BayesC (BC). LASSO and elastic net are penalized regression methods that differ in the choice of the penalty function; LASSO[13] uses the L1 norm as the penalty function whereas elastic net[14] uses the weighted average of the L1 and L2 norms. While both methods perform a combination of variable selection and shrinkage on marker effects, elastic net also accounts for correlated predictors better than LASSO. BRR, BL and BC are Bayesian shrinkage estimators that use a Gaussian prior, thick-tailed (double-exponential) prior and spike-slab (point-of-mass at zero and Gaussian slab) prior, respectively, for marker effects. BRR and BL perform homogeneous and differential shrinkage respectively, whereas BC performs a combination of variable selection and homogeneous shrinkage on marker effects[15]. The weights W for LASSO and elastic net were obtained using the glmnet[16] package in R while those for BRR, BL and BC were obtained using the BGLR[17] package in R.
Accounting for LD: Irrespective of the training sample size used to obtain weights, the covariance matrix among all the chosen SNPs in the gene Σ was obtained using the full training set of 1000 individuals. This is reasonable because, [i] in practice, publicly available human genotype data (e.g. 1000 genomes data[18]) can be used for this purpose and [ii] we wanted to keep the influence of LD consistent between training sets.
Imputing the weights into GWAS: TWAS was conducted by imputing the weights W obtained using each of the ***five*** above-described penalized/Bayesian regularized regression approaches into the GWAS summary statistics. The single imputed z-score (normally distributed with zero mean and unit variance) of *cis*-genetic effect on the phenotype can be obtained as follows:

$$z_{TWAS-MP} = W' Z_{GWAS} / (W' \Sigma W)^{\frac{1}{2}} \quad (4)$$

Similar to eGWAS, the gene was considered to be detected if its p-value < 0.05/15,000.

4. TWAS-SMR: For the given gene, we obtained the TWAS-SMR-based z-score by combining the z-score of the single most significant SNP from eQTL analysis ($z_{eQTL} = \min(Z_{eQTL})$) with the z-score of the corresponding SNP from GWAS (z_{eGWAS}), which can be expressed as follows:

$$z_{TWAS-SMR} \approx (z_{eQTL} * z_{eGWAS}) / \sqrt{(z_{eQTL}^2 + z_{eGWAS}^2)} \quad (5)$$

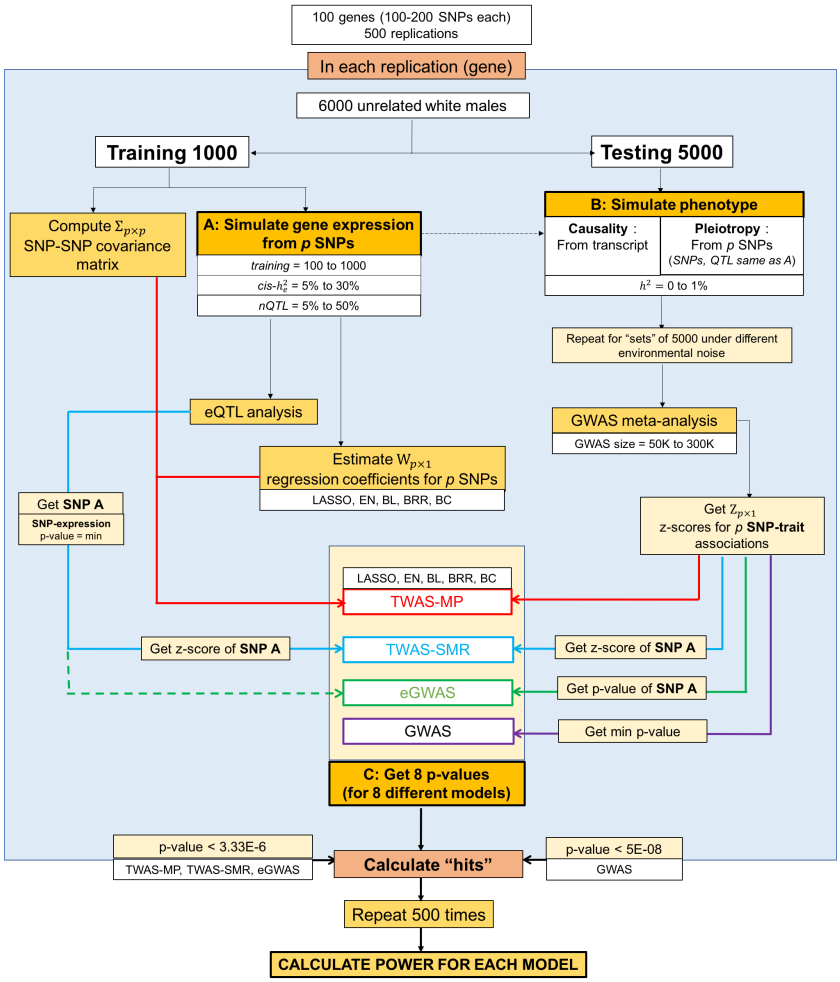


Figure 1. Simulation pipeline per gene for power analysis

Similar to eGWAS and TWAS-MP, the gene was considered to be detected if its p-value < 0.05/15,000.

The entire procedure was repeated 500 times and power was calculated as the fraction of instances where the given gene was detected. A summarized version of the power analysis pipeline is given in **Figure 1**. All models were fit using the 3.2.1 version of software R.

3. Results

We observed that the power of detecting an expression-trait association varied not only with the genetic architecture of the trait but also with the sample size. Let's first consider the genetic model corresponding to **causality** (**Figure 2**). Broadly, power was observed to increase with: (1) the sample size used for training the TWAS imputation algorithm and for eQTL analyses, (2) the sample size used to conduct GWAS meta-analysis, (3) the trait heritability as well as (4) the expression heritability. We observed that GWAS sample size had a bigger effect on power than the training sample size (we only considered realistic GWAS and training sample sizes).

Across all cases, we observed that a trait heritability of less than 0.001% resulted in low to zero power, irrespective of the considered sample sizes. For a GWAS sample size as large as 150,000 individuals, trait heritability less than 0.025% yielded low to zero power across all methods (even when the expression heritability was as high as 30%). In addition, eGWAS and

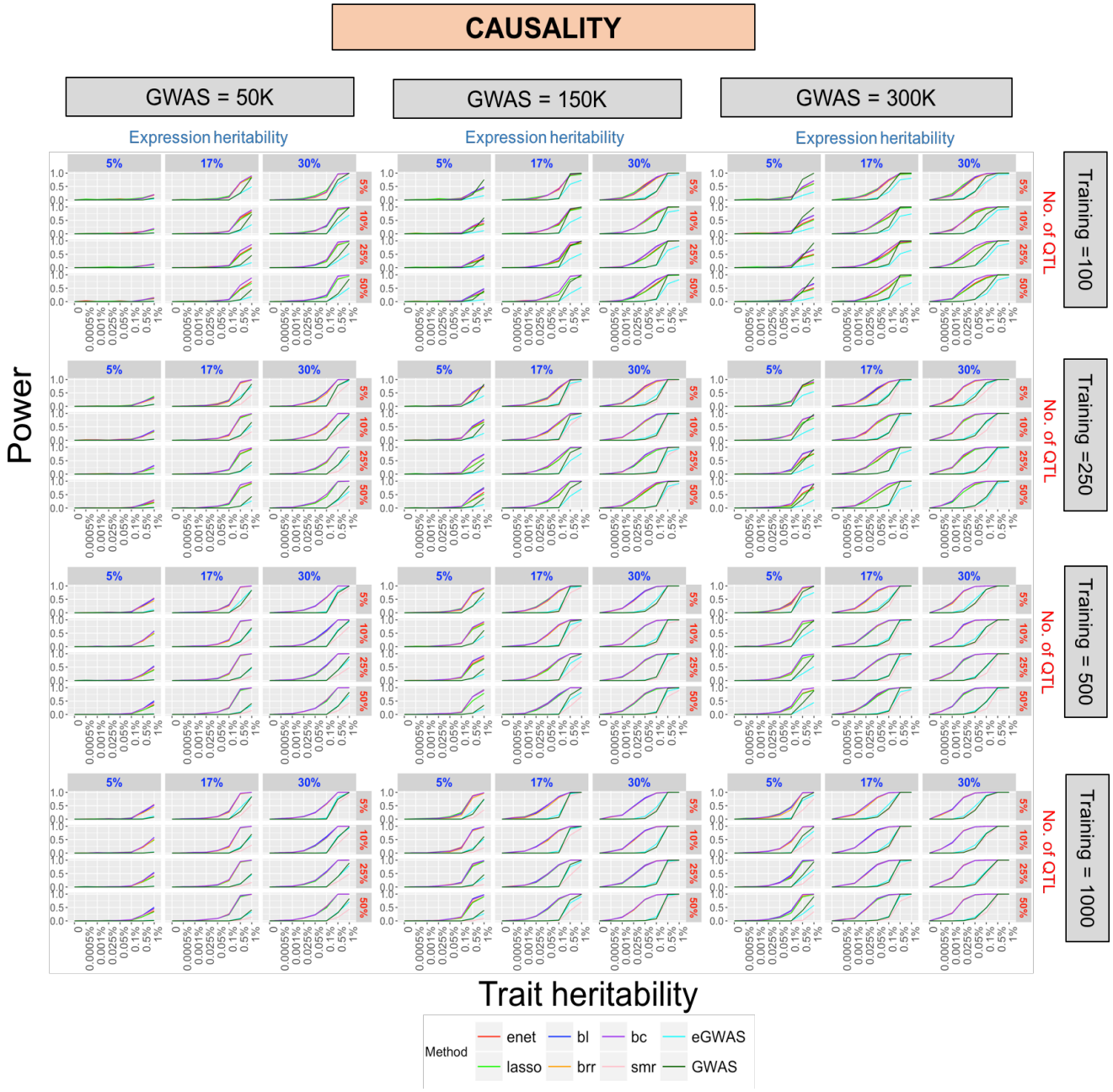


Figure 2. **Power distributions under causality (full)**: Power (y-axis) across trait heritability (x-axis) under varying levels of expression heritability (in blue—subplot columns), no. of QTL (in red—subplot rows), training sample size (main plot rows) and GWAS sample size (main plot columns). Each plot represents power from 8 different models; TWAS-MP (LASSO, elastic net, BL, BC, BRR), TWAS-SMR, eGWAS, and GWAS. **Enlarged views of this plot can help identify differences between methods. Relevant differences between TWAS-MP methods can be seen in Figure 2.**

TWAS-SMR did significantly worse than all other considered methods, except when trait heritability, expression heritability and GWAS sample size were very high (~ 1%, ~30% and >=150,000, respectively).

TWAS-SMR achieved peak performance and offered power comparable to eGWAS and GWAS (across all levels of trait heritability) when the expression heritability, training sample size and GWAS sample sizes were all at their highest levels (30%, 1000 and 300,000 respectively).

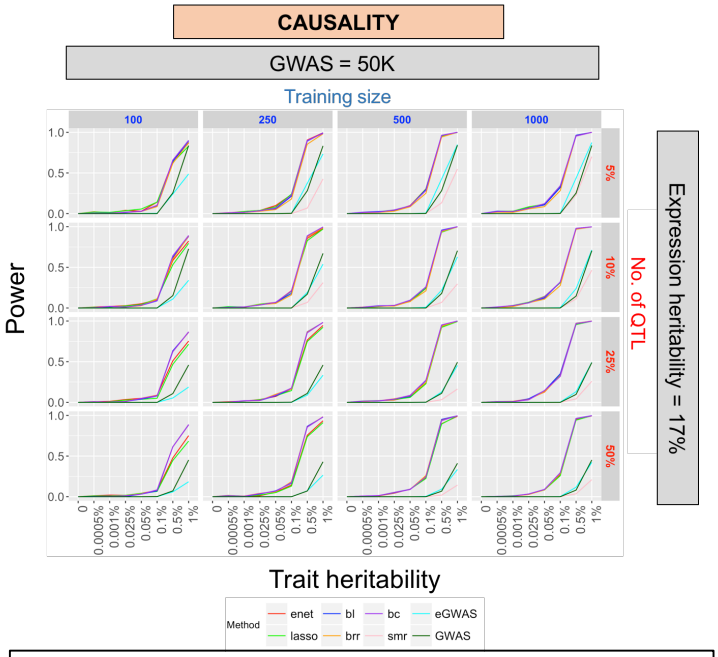


Figure 3. **Power distributions under causality (reduced)**: Power (y-axis) across trait heritability (x-axis) under varying levels of training sample sizes (in blue-columns) and no. of QTL (in red-rows), at **GWAS sample size of 50K** and **expression heritability of 17%.** Each plot represents power from 8 different models; TWAS-MP (LASSO, ENet, BL, BC, BRR), TWAS-SMR, eGWAS, and GWAS.

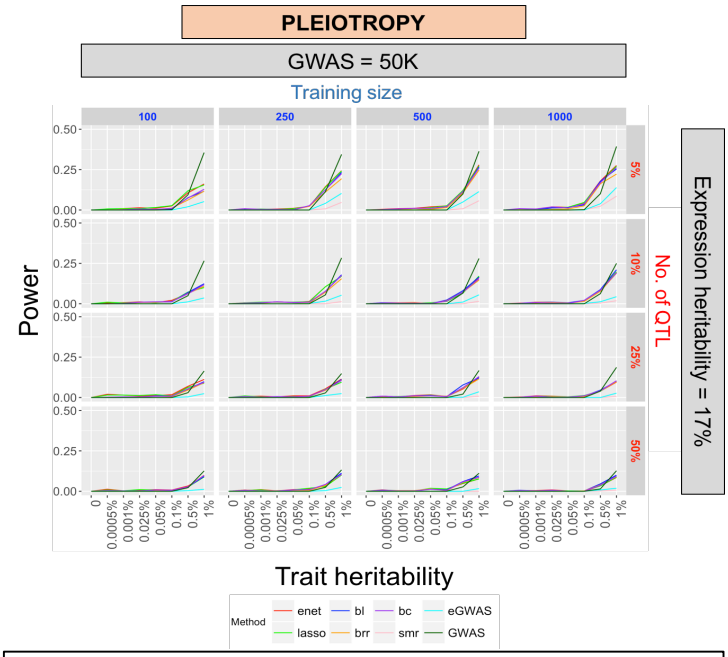


Figure 4. **Power distributions under pleiotropy (reduced)**: Power (y-axis) across trait heritability (x-axis) under varying levels of training sample sizes (in blue-columns) and no. of QTL (in red-rows), at **GWAS sample size of 50K** and **expression heritability of 17%**. Each plot represents power from 8 different models; TWAS-MP (LASSO, ENet, BL, BC, BRR), TWAS-SMR, eGWAS, and GWAS.

However, its performance was still worse than that of eGWAS and GWAS, especially when expression heritability was low. Although eGWAS's performance was also poor under low expression heritability, it made up for this loss as GWAS sample size increased. As expected, power afforded by GWAS was unaffected by expression heritability and training sample size; it increased only with trait heritability and GWAS sample size. Interestingly, GWAS resulted in marked improvement in power over all other methods when expression heritability, number of QTL, and training sample sizes were at their lowest levels and GWAS sample size was high (see first subplot column under top right main plot panel in **Figure 2**; GWAS is in dark green).

TWAS-MP always resulted in the highest power, except when expression heritability and training sample sizes were at their lowest (**Figure 2**). For instance, given an expression heritability of 17% and a trait heritability of 0.1%, moderate sample sizes for training and GWAS (250 and 150,000 respectively) were sufficient to achieve >=75% power using any of the TWAS imputation methods. Also, genes with average to high expression heritability were found to have very high power of detecting a significant gene-trait association even when GWAS and training sample sizes were low; the power ranged from approximately 0% at $h_e^2 = 5\%$ to approximately 100% at $h_e^2 = 30\%$ for a gene that had a trait heritability of greater than 0.5% (see top- and left-most panel in **Figure 2**). In general, the TWAS-MP methods yielded almost identical power. However, Bayesian methods performed better than LASSO and elastic net when the expression heritability was low to moderate (5%-17%), number of QTL was high

(>=25%) and training sample size was low to moderate (<=500) (**Figure 2)**. In particular, when the expression heritability is low (5%), BL achieved a maximum improvement in power (~ 17%-18%) as compared to the elastic net and LASSO under a trait heritability of at least 0.5% and using a GWAS sample size of 150K (**Table 1**).

Table 1. **Comparison of power of TWAS-MP methods under causality (reduced)**: Power using 5 TWAS-MP approaches (elastic net, LASSO, BL, BRR, BC) at training sample sizes of 100, 250, 500, 1000, **GWAS sample size of 150K**, **expression heritability of 5%** and trait heritability of 0.5% and 1%, respectively.

TWAS-MP	Training sample size															
	100				250				500				1000			
	nQTL				nQTL				nQTL				nQTL			
	25%		50%		25%		50%		25%		50%		25%		50%	
	h^2		h^2		h^2		h^2		h^2		h^2		h^2		h^2	
	0.50%	1%	0.50%	1%	0.50%	1%	0.50%	1%	0.50%	1%	0.50%	1%	0.50%	1%	0.50%	1%
ENET	0.207	0.352	0.196	0.585	0.313	0.848	0.312	0.370	0.635	0.592	0.514	0.787	0.945	0.994	0.908	0.982
LASSO	0.179	0.317	0.185	0.578	0.304	0.808	0.289	0.339	0.591	0.525	0.509	0.791	0.934	0.992	0.893	0.988
BL	0.250	0.484	0.258	0.748	0.498	0.938	0.484	0.488	0.734	0.768	0.672	0.932	0.988	1.000	0.984	0.998
BRR	0.258	0.444	0.242	0.722	0.472	0.910	0.452	0.468	0.706	0.736	0.664	0.918	0.968	1.000	0.960	0.998
BC	0.256	0.458	0.234	0.732	0.486	0.930	0.468	0.462	0.716	0.716	0.660	0.898	0.974	0.998	0.986	0.998

Under **pleiotropy**, GWAS always resulted in the best power among all the other methods considered (**Figure 4**) and the trend was consistent across training and GWAS sample sizes as well as levels of expression heritability (data not shown). Accordingly, the power peaked when number of causal variants was small. Interestingly, even with a trait heritability as high as 1%, we could only achieve a maximum power of approximately 40% with GWAS.

4. Discussion

TWAS have been introduced as a way to combine SNP-expression information and GWAS to identify genes whose expression levels are associated with a trait. A recent study has applied TWAS to over 30 different complex human traits to identify functional signatures in pleiotropic traits[19]. However, the scenarios under which different flavors of TWAS can achieve improved power as compared to eQTL-based GWAS and GWAS have not yet been explored. In this study, we examine the influence of complex genetic architectures and sample size on power afforded by different TWAS-based approaches (five TWAS-MP methods and TWAS-SMR), eGWAS and GWAS. We vary several simulation parameters including the number of QTL, the training sample size, the GWAS sample size, the trait heritability and the expression heritability under two genetic models (causality and pleiotropy) and examine the influence of each on power.

4.1 Training sample size

Training sample size is important since eQTL studies are typically limited in sample size. The NIH Common Fund project called Genotype Tissue Expression Project (GTEx[20]) is assembling a database of SNP-expression associations spanning 43 different tissues. However, for any given tissue, the sample size is fairly low, ranging from approximately 77 (small intestine terminal ileum) to 161 (muscle skeletal). Other currently available SNP-expression studies are also limited in size, e.g. the Netherlands Twin Register (1,247 peripheral blood samples), the Metabolic Syndrome in Men study (563 adipose samples[21–23]), the Genetic European Variation in Health and Disease (460 lymphoblastoid cell lines[8,24]), Depression Genes and Network (922 whole blood samples[25]), and

Braineac (130 individuals with brain region samples[26]). Accordingly, we explored training sample sizes ranging from 100 to 1,000 in this study. Under the assumption of causality, we see that even a sample size as small as 100 is sufficient to achieve 100% power for a gene with moderate expression heritability (17%) as long as the GWAS sample size is at least 150,000. Training sample size was not observed to have a marked influence under pleiotropy (**Figure 4**).

4.2 GWAS sample size

GWAS sample sizes have been increasing over the years using meta-analyses across multiple cohorts and a multitude of common variants have been detected for a host of complex traits and diseases. We observe that GWAS sample size plays a crucial role in also detecting gene-trait associations, especially under the assumption of causality. A high GWAS sample size can help detect genes with low expression heritability (and moderate to high trait heritability) even when the training sample size is small, especially under the assumption of causality (**Figure 2**).

4.3. Number of QTL

We chose genes with of sizes between 100 and 200 SNPs and included the region 500 kb upstream and downstream of the gene into our analyses to investigate the impact of the number of causal variants as well as the extent of LD between markers and causal variants on statistical power. It is known that these factors affect the prediction accuracy of a trait in whole-genome regression based studies[27,28]. Under the assumption of causality (**Figures 2** and **3**), the number of QTL had a noticeable impact on power obtained using eGWAS, GWAS, and TWAS-SMR while that obtained from TWAS-MP was not significantly affected. This is understandable given that eQTL-guided GWAS, GWAS and TWAS-SMR only choose the top-most significant SNP/eQTL in the gene and lose a considerable portion of the genetic signal when the number of QTL forms a large proportion of the gene. This behavior, albeit muted, was also observed under pleiotropy (**Figure 4**).

4.4. Expression heritability

Few studies have thus far shed light on the average heritability of gene expression across different cohorts and tissues. This parameter refers to the proportion of variation in gene expression that can be explained by genotype. Under the genetic model of causality (**Figures 2** and **3**) we observe that expression heritability has a profound influence on power, especially when training sample size and GWAS sample sizes are moderate to low (e.g. top left-most panel in **Figure 2**). Under the genetic model of pleiotropy, expression heritability only has a slight influence on TWAS-MP but no effect on the other methods (eGWAS, GWAS and TWAS-SMR), irrespective of training and GWAS sample sizes (data not shown). This intuitive result confirms that even a gene with very high expression heritability is not likely to have high power to detect a gene-trait association when gene expression does not mediate between the SNP and the phenotype.

4.5. Trait heritability

Complex traits have widely varying heritability measures ranging from ~80-90% for height[29] to between 30%-70% for lipid traits[30]. We chose an upper limit of 1%, which would correspond to a large-effect gene that explains almost 1% of the overall trait heritability. Under causality, we

observed that TWAS-MP methods were powerful in detecting genes even with moderate trait heritability (17%) as long as the sample sizes were high (**Figures 2** and **3**). Under pleiotropy, we observed that a gene needed to have very high trait heritability (>1%) to be detected with moderate power (<40%) at best (**Figure 4**).

4.6. Genetic model

We only considered two genetic models in this study. The power obtained under pleiotropy was significantly lower than that obtained under causality, which demonstrates the weaknesses of TWAS methods when genes operate under non-causal genetic models (**Figures 3-4**).

4.7. Statistical model

In general, all TWAS-MP methods (LASSO, elastic net, BC, BRR, and BL) performed uniformly well and achieved high power under the assumption of causality. However, in particular, Bayesian methods performed better than LASSO and elastic net as the trait architecture grew more complex, expression heritability remained low and training sample sizes were small (**Figure 2** and **Table 1**). This shows that LASSO and elastic net are more conducive for variable selection than BL, BRR, and BC and their performance worsens as a greater number of predictors in the model carry genetic signal. On the other hand, TWAS-SMR did much worse than TWAS-MP under all considered simulation scenarios. As expected, eGWAS, GWAS and TWAS-SMR had better power when the number of QTL was small although their performance still lagged behind that of TWAS-MP methods. As expression mediated weakly between SNP and trait, performance of TWAS worsened and assuming no mediation at all (pleiotropy), GWAS performed better than TWAS-MP, TWAS-SMR and eGWAS (which had uniformly poor power).

A limitation of this work is that our GWAS "meta-analysis" only comprised Caucasian males, which is likely to have resulted in a sample with far more homogeneous LD patterns than what can be expected in reality. Also, our meta-analysis (sampling with replacement) is likely to have resulted in inflated power due to sample relatedness. We will exploit more heterogeneous GWAS samples in the future and will also conduct type I error experiments to ensure type I error is well controlled. Also, we assumed that all causal variants were included in our model whereas in reality we might only have SNPs tagged to the causal variants. Finally, it is a worthwhile future exercise to compare power of TWAS-MP to TWAS-SMR when *both* eQTL and GWAS data have summary-level data.

In conclusion, we have presented a comprehensive power analysis for detecting gene-trait associations under a range of complex genetic architectures using approaches based on individual-level and/or summary-level data. In future, these methods could also be applied to integrate GWAS with other kinds of "omic" information aside from gene expression (e.g. metabolomics, methylation). This is a starting step to better understand methods that can illuminate genetic patterns and functional mechanisms underlying complex trait variation in a powerful yet computationally efficient manner.

5. References

1. Welter, D. *et al*. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. **42,** D1001-6 (2014).

2. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
3. Nicolae, D. L. *et al*. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet*. **6,** e1000888 (2010).
4. Schadt, E. E. Novel integrative genomics strategies to identify genes for complex traits. *Anim. Genet*. **37,** 18–23 (2006).
5. Gusev, A. *et al*. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet*. **48,** 245–252 (2016).
6. Zhu, Z. *et al*. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet*. **48,** 481–487 (2016).
7. Giambartolomei, C. *et al*. Bayesian Test for Colocalisation Between Pairs of Genetic Association Studies Using Summary Statistics. (2013).
8. Gamazon, E. R. *et al*. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet*. **47,** 1091–1098 (2015).
9. Zhu, Z. *et al*. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet*. **48,** 481–487 (2016).
10. Barbeira, A. N. *et al*. Integrating tissue specific mechanisms into GWAS summary results. (2016). doi:10.1101/045260
11. Carey, D. J. *et al*. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet. Med*. **18,** 906–13 (2016).
12. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput*. 368–79 (2009).
13. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58,** 267–288 (1996).
14. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B (Statistical Methodol*. **67,** 301–320 (2005).
15. Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. Additive genetic variability and the Bayesian alphabet. *Genetics* **183,** 347–63 (2009).
16. Hastie, T. & Qian, J. Glmnet Vignette. (2016).
17. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198,** 483–95 (2014).
18. Auton, A. *et al*. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
19. Mancuso, N. *et al*. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet*. **100,** 473–487 (2017).
20. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet*. **45,** 580–5 (2013).
21. Wright, F. A. *et al*. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet*. **46,** 430–437 (2014).
22. Nuotio, J. *et al*. Cardiovascular risk factors in 2011 and secular trends since 2007: the Cardiovascular Risk in Young Finns Study. *Scand. J. Public Health* **42,** 563–71 (2014).
23. Gusev, A. *et al*. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet*. **48,** 245–252 (2016).
24. Lappalainen, T. *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
25. Battle, A. *et al*. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. **24,** 14–24 (2014).
26. Ramasamy, A. *et al*. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci*. **17,** 1418–1428 (2014).
27. Wimmer, V. *et al*. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* **195,** 573–87 (2013).
28. VanRaden, P. M. *et al*. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci*. **92,** 16–24 (2009).
29. Silventoinen, K. *et al*. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res*. **6,** 399–408 (2003).
30. Kettunen, J. *et al*. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet*. **44,** 269–76 (2012).

Democratizing Health Data for Translational Research

Philip R.O Payne
Institute for Informatics, Washington University School of Medicine
St. Louis, MO, USA
Email: prpayne@wustl.edu

Nigam H. Shah
Center for Biomedical Informatics Research, Stanford University School of Medicine
Stanford, CA, USA
Email: nigam@stanford.edu

Jessica D. Tenenbaum
Department of Biostatistics and Bioinformatics, Duke University School of Medicine
Durham, NC, USA
Email: jessie.tenenbaum@duke.edu

Lara Mangravite
Sage Bionetworks
Seattle, WA, USA
Email: lara.mangravite@sagebase.org

There is an expanding and intensive focus on the accessibility, reproducibility, and rigor of basic, clinical, and translational research. This focus complements the need to identify sustainable ways to generate actionable research results that improve human health. The principles and practices of *open science* offer a promising path to address both issues by facilitating: 1) increased transparency of data and methods which promotes research reproducibility and rigor; and 2) cumulative efficiencies wherein research tools and the output of research are combined to accelerate the delivery of new knowledge. While great strides have been in made in terms of enabling the open science paradigm in the biological sciences, progress in sharing of patient-derived health data has been more moderate. This lack of widespread access to common and well characterized health data is a substantial impediment to the timely, efficient, and multi-disciplinary conduct of translational research, particularly in those instances where hypotheses spanning multiple scales (from molecules to patients to populations) are being developed and tested. To address such challenges, we review current best practices and lessons learned, and explore the need for policy changes and technical innovation that can enhance the sharing of health data for translational research.

Keywords: Open Science, Open Data, Translational Research, Data Science

1. Introduction

There is an emergent national and international dialogue concerned with the accessibility, reproducibility, and rigor of all types of biomedical research. Simultaneously, it has been recognized that the scientific community needs new approaches to make its work sustainable during times of both decreased funding and increased demand for timely and actionable outcomes of research programs. One potential solution to both of these challenges is the adoption of *open science* models that allow: 1) increased transparency of data and methods, which promotes research reproducibility and rigor [1-4]; and 2) cumulative efficiencies wherein research tools and the output of research are combined to accelerate the delivery of new knowledge [5-7]. For the purposes of this manuscript, we provide the following working definition for open science:

"Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods."[8]

Unfortunately, contradictory and sometimes conflicting positions on open science - and the way the open science paradigm might best be operationalized - demonstrate the need for greater community engagement to test the theory that open science in the health sciences can indeed improve the rigor and efficiency of research. This challenge is exemplified by the recent controversy regarding research "parasites" [9], and the vigorous debate that ensued as a result. Furthermore, while there have been some notable and early successes in terms of enabling open science frameworks, such as the creation of data sharing "commons" or the enforcement of data sharing policies concomitant with formal publication of research results, most if not all of these efforts have been focused on biologic data sets such as those related to either: 1) –omics focused measurements of bio-molecular phenomena with some small volume of associated clinical phenotype annotations; or 2) limited scale and highly synthesized data derived from clinical trials of new therapies or diagnostic methods [10-14]. Often, these data are used as a "reference" for secondary interrogation, for example, in studies involving the derivation of phenotypic "signatures" that correlate disease risk, severity, or potential response to therapy [15-19], or for the identification of candidates for drug repositioning/repurposing [20-22], to name a few of many potential examples.

In contrast, the wide-spread and comprehensive sharing of "reference" data sets containing patient-derived health data, such as that found in Electronic Health Records (EHRs), Clinical Research Management Systems (CRMS), or Electronic Data Capture (EDC) systems remains far less prevalent or developed. A number of rationales have been given for this lack of sharing and re-use of patient-derived data, including concerns surrounding patient consent and privacy [10, 23], the mechanisms for attribution of such data and its sources [12-14, 24], and uncertainty surrounding the quality and completeness of data sets collected for primarily clinical or administrative purposes [11, 13, 25]. Unfortunately, in the absence of "reference" data sets that include adequate amounts of patient-derived health data and that are well annotated and understood from a content, quality, and

provenance standpoint, we risk substantial inefficiencies as well as the potential for non-reproducible research, where such studies involve the development of novel ways to reason across multiple scales of phenotype, a situation that is intrinsic to what we often refer to as translational research.

Emergent efforts such as All of Us (formerly the Precision Medicine Initiative) in the United States [26] and the Observational Health Data Science and Informatics (OHDSI) that seek to create mechanisms and best practices for the sharing of patient-derived health data [27], as well as an increasing emphasis on the creation and maintenance of registries for "Real World Evidence" (RWE) generation by both diseases focused groups and biotechnology and pharmaceutical firms [2, 3, 7], provide the basis for a path forward. However, all of the preceding efforts remain both formative and early in their development. As such***, there remain substantial and unanswered questions concerning how to achieve a vision of "democratized" health data for translational research*** wherein all of the preceding challenges and opportunities have been adequately addressed. Ideally, such a vision of "democratized" health data would involve the adoption and widespread use of open science approaches that include a full spectrum of data types and assets.

2. Background

Never before has the health and life sciences communities been able to access such a wide variety of open data resources. Such data sets include measurements at the bio-molecular, clinical, and population levels, and are often derived from observational, clinical, and broader public health studies, not to mention an ever-increasing number of patient-reported indicators of health and wellness as well as sensors and other ubiquitous computing sources [4, 7, 11-13, 28, 29]. When viewed as a whole, these open data resources represent an opportunity for a paradigm shifting approach to discovery science, one in which we move away from the collection and curation of high-cost and project-specific data sets in which a small number of hypotheses are tested, and towards a model in which large-scale and heterogeneous data sets are collected, integrated, shared, and interrogated in a high throughput manner. This **open science** paradigm has the potential to substantially increase the speed and impact of research, while also reducing costs and barriers to answering critical questions by making the pursuit of such question cumulative in nature [7, 14, 30].

Given the promise of open science, one must ask why such an approach is not more common and widespread. Unfortunately, there exist a number of notable impediments in the contemporary scientific environment that preclude or inhibit the pursuit of open science, including:

- Confusing, overlapping, or contradictory **regulatory frameworks** governing the sharing of and access to research data, particularly data derived from humans;
- Misaligned **incentives and "community standards"** corresponding to research **career development and peer recognition**; and
- The dearth of **suitable platforms, technologies, and best practices** that can serve to support or enable the pursuit of open science by geographically, temporally, or otherwise distributed research teams that span traditional organizational boundaries and settings.

The papers and presentations associated with our session at the 2018 Pacific Symposium and Biocomputing (PSB 2018), entitled "**Democratizing Health Data for Translational Research**", explore each of these areas in further detail, particularly as they relate to implementing open science paradigms when seeking to understand the critical relationships between bio-molecular and clinical phenotypes in both health and wellness. As part of this collection of papers and presentations, there is both an assessment of the current state-of-the-art as well as evolving approaches and solutions to such impediments. Ultimately, it is our belief that the benefits of, and momentum behind, open science paradigms will overcome these barriers as a result in solutions that address fundamental flaws associated with "traditional" and highly compartmentalized approaches to translational research. This momentum will be amplified by the way in which open science democratizes access to and participation in research endeavors, thus supporting true communities-of-practice and the economy of ideas and thinking such collaborative constructs provide for. However, the trajectory of open science we envision will not be easy, as it represents a fundamental culture change in the health and life sciences communities, and it is well understood and documented that culture change is challenging and fraught with peril for early adopters or advocates of such change. As such, we believe that this session and its content serve as a critical "marker" concerning the future directions and research agendas needed to realize such a vision for high impact discovery science and translational research.

3. Democratizing Health Data for Translational Research

As was noted above, a selection of papers and presentations concerning the current state-of-the-art in terms of democratizing health data for translational research, as well as evolving approaches to fundamental impediments in implementing open science, was curated as part of the proceeding of PSB 2018. Below is a thematic summary of the three major areas of endeavor highlighted by those papers and presentations:

- **Leveraging data models and standard to improve the discoverability and utility of public data repositories:** Exemplars of this theme included: 1) efforts by ***Madhavan*** and colleagues to employ syntactic and semantic standards in order to improve the accessibility and usefulness of comprehensive biomolecular and clinical data sets created by the ClinGen initiative; 2) methods being developed by ***Tenenbaum*** and colleagues to ensure the reproducibility of analyses using publically available data assets relevant to Alzheimer's research; and 3) similar methodological development efforts by ***Sharma*** and colleagues to extract meaningful health outcomes and natural product therapeutic information from adverse event report systems.
- **Synthesizing and simulating data sets that are comparable to human-derived source data:** Exemplary of this theme is the work by ***Moore*** and colleagues to employ simulation techniques to synthesize health data that can be used to inform the design and evaluation of a variety of machine learning algorithms that can identify potentially informative patterns spanning multidimensional data.
- **Educating informatics investigators to systematically and responsibly access and utilize emerging sources of health data for translational research:** Finally, this theme is represented by the work of ***Van Horn*** and colleagues to define pedagogical approaches for

preparing biomedical informatics and data science investigators to responsibly and reproducibly utilize health data as found in a variety of domains in order to support hypothesis generating and testing science.

4. Conclusions

PSB 2018 is a unique venue for thought leadership and technical direction setting that we believe can enable widespread action to "democratize" health data and to support timely as well as high impact translational research. This capability is exemplified by the work presented in our session. Ultimately, the themes and findings presented by our authors chart a path forward for this important area, involving:

- The **creation, verification and validation of tools** and methods that can assist in the sharing, discovery, and analysis of open health data in a primary or secondary manner, including the development of databases, algorithms, and modeling techniques therein;
- The **conduct of discovery science in data-intensive experimental contexts** that leverage such open health data resources across scales from molecules to patient to populations; and
- The **preparation and interaction of multidisciplinary computational, biology, clinical, and population health science teams** to conduct research that serves to identify policy, technical, and socio-cultural needs associated with the implementation of open science paradigms that include patient-derived health data.

This research agenda can and should advance our collective understanding of the role of "democratized" health data in advancing the state-of-the-art in translational research writ large with demonstrable benefit in terms of human health and wellness.

Acknowledgements

The authors wish to acknowledge the contributions of all of the authors who submitted content to this session at PSB 2018, as well as the scientific and editorial oversight provided by the conference's scientific program committee.

References

1. Goodman, S.N., D. Fanelli, and J.P. Ioannidis, *What does research reproducibility mean?* Science translational medicine, 2016. **8**(341): p. 341ps12-341ps12.
2. Iqbal, S.A., et al., *Reproducible research practices and transparency across the biomedical literature.* PLoS Biol, 2016. **14**(1): p. e1002333.
3. Nosek, B.A., et al., *Promoting an open research culture.* Science, 2015. **348**(6242): p. 1422-1425.
4. Warren, E., *Strengthening research through data sharing.* New England Journal of Medicine, 2016. **375**(5): p. 401-403.
5. Holve, E., *Open Science and eGEMs: Our Role in Supporting a Culture of Collaboration in Learning Health Systems.* eGEMs, 2016. **4**(1).
6. McKiernan, E.C., et al., *How open science helps researchers succeed.* Elife, 2016. **5**: p. e16800.
7. Moher, D., et al., *Increasing value and reducing waste in biomedical research: who's listening?* The Lancet, 2016. **387**(10027): p. 1573-1586.
8. FOSTER. *Open Science Taxonomy*. 2016 [cited 2016 September 14]; Available from: https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition.
9. Longo, D.L. and J.M. Drazen, *Data sharing.* New England Journal of Medicine, 2016. **374**(3): p. 276-277.
10. Aitken, M., et al., *Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies.* BMC medical ethics, 2016. **17**(1): p. 73.
11. Ross, J.S. and H.M. Krumholz, *Open Access Platforms for Sharing Clinical Trial Data.* Jama, 2016. **316**(6): p. 666-666.
12. Taichman, D.B., et al., *Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors.* JAMA, 2016. **315**(5): p. 467-468.
13. Vallance, P., A. Freeman, and M. Stewart, *Data Sharing as Part of the Normal Scientific Process: A View from the Pharmaceutical Industry.* PLoS Med, 2016. **13**(1): p. e1001936.
14. Wilbanks, J. and S.H. Friend, *First, design for data sharing.* Nature biotechnology, 2016.
15. Denny, J.C., et al., *Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data.* Nature biotechnology, 2013. **31**(12): p. 1102-1111.
16. Payne, P.R. and P.J. Embi, *An Introduction to Translational Informatics and the Future of Knowledge-Driven Healthcare*, in *Translational Informatics*. 2015, Springer. p. 3-19.

17. Plenge, R.M., E.M. Scolnick, and D. Altshuler, *Validating therapeutic targets through human genetics.* Nature reviews Drug discovery, 2013. **12**(8): p. 581-594.
18. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype-phenotype interactions.* Nature Reviews Genetics, 2015. **16**(2): p. 85-97.
19. Shah, N.H., *Mining the ultimate phenome repository.* Nature biotechnology, 2013. **31**(12): p. 1095-1097.
20. Chen, B. and A.J. Butte, *Leveraging big data to transform target selection and drug discovery.* Clinical Pharmacology & Therapeutics, 2016. **99**(3): p. 285-297.
21. Li, J., et al., *A survey of current trends in computational drug repositioning.* Briefings in bioinformatics, 2016. **17**(1): p. 2-12.
22. Liu, Z., et al., *In silico drug repositioning–what we need to know.* Drug discovery today, 2013. **18**(3): p. 110-115.
23. Joly, Y., et al., *Are Data Sharing and Privacy Protection Mutually Exclusive?* Cell, 2016. **167**(5): p. 1150-1154.
24. Krumholz, H.M., S.F. Terry, and J. Waldstreicher, *Data acquisition, curation, and use for a continuously learning health system.* Jama, 2016. **316**(16): p. 1669-1670.
25. Krumholz, H.M. and J. Waldstreicher, *The Yale Open Data Access (YODA) project—a mechanism for data sharing.* New England Journal of Medicine, 2016. **375**(5): p. 403-405.
26. Ashley, E.A., *Towards precision medicine.* Nature Reviews Genetics, 2016. **17**(9): p. 507-522.
27. *OHDSI*. 2017; Available from: http://www.ohdsi.org/.
28. Frasier, M., *Perspective: Data sharing for discovery.* Nature, 2016. **538**(7626): p. S4-S4.
29. Payne, P., et al., *Enabling Open Science for Health Research: Collaborative Informatics Environment for Learning on Health Outcomes (CIELO).* Journal of Medical Internet Research, 2017. **19**(7).
30. Watson, M., *When will 'open science' become simply 'science'?* Genome biology, 2015. **16**(1): p. 101.

ClinGen Cancer Somatic Working Group – standardizing and democratizing access to cancer molecular diagnostic data to drive translational research

Subha Madhavan[1], Deborah Ritter[2], Christine Micheel[3], Shruti Rao[1], Angshumoy Roy[2], Dmitriy Sonkin[4], Matthew Mccoy[1], Malachi Griffith[5], Obi L Griffith[5], Peter Mcgarvey[1], Shashikant Kulkarni[2] on behalf of the ClinGen Somatic Working Group

1 Innovation Center for Biomedical Informatics, Georgetown University, Washington D.C.; 2. Baylor College of Medicine and Texas Children's Hospital, Houston, TX.; 3. Vanderbilt University School of Medicine, Nashville, TN.; 4. National Cancer Institute, Rockville, MD.; 5. The McDonnell Genome Institute, Washington University, St. Louis, MO.

Abstract

A growing number of academic and community clinics are conducting genomic testing to inform treatment decisions for cancer patients (1). In the last 3-5 years, there has been a rapid increase in clinical use of next generation sequencing (NGS) based cancer molecular diagnostic (MolDx) testing (2). The increasing availability and decreasing cost of tumor genomic profiling means that physicians can now make treatment decisions armed with patient-specific genetic information. Accumulating research in the cancer biology field indicates that there is significant potential to improve cancer patient outcomes by effectively leveraging this rich source of genomic data in treatment planning (3). To achieve truly personalized medicine in oncology, it is critical to catalog cancer sequence variants from MolDx testing for their clinical relevance along with treatment information and patient outcomes, and to do so in a way that supports large-scale data aggregation and new hypothesis generation. One critical challenge to encoding variant data is adopting a standard of annotation of those variants that are clinically actionable. Through the NIH-funded Clinical Genome Resource (ClinGen) (4), in collaboration with NLM's ClinVar database and >50 academic and industry based cancer research organizations, we developed the Minimal Variant Level Data (MVLD) framework to standardize reporting and interpretation of drug associated alterations (5). We are currently involved in collaborative efforts to align the MVLD framework with parallel, complementary sequence variants interpretation clinical guidelines from the Association of Molecular Pathologists (AMP) for clinical labs (6). In order to truly democratize access to MolDx data for care and research needs, these standards must be harmonized to support sharing of clinical cancer variants. Here we describe the processes and methods developed within the ClinGen's Somatic WG in collaboration with over 60 cancer care and research organizations as well as CLIA-certified, CAP-accredited clinical testing labs to develop standards for cancer variant interpretation and sharing.

Keywords: ClinGen, Somatic variants, predictive biomarkers, MVLD, data sharing

ClinGen

To address these needs of capturing, standardizing and sharing clinically relevant variants, the Clinical Genome Resource, ClinGen (4) collaboration was established in 2012 and has been developing interconnected community resources to improve our understanding of genomic variation and enhance its use in clinical care. ClinGen represents a strong partnership among public, academic, and private institutions that relies on collaboration between the NIH, academic and commercial laboratories operating in both the research and clinical realms. ClinGen is also engaging numerous entities, including professional societies, to ensure that the resources that are produced meet community expectations. The Somatic working group (Somatic WG) is a clinical domain working group within the ClinGen consortium and was established in 2015 to address standardization and sharing of cancer MolDx test results described here.

The Standard

In order to standardize the collection of clinically relevant somatic data, the Somatic WG of ClinGen created a framework of consensus data elements titled "Minimum Variant Level Data" (MVLD) (5). MVLD was developed with input from multiple stakeholders ranging from database engineers to researchers and somatic clinical laboratory directors, as well as input from multiple current databases that collect cancer variant data. Briefly, MVLD consists of three sections: allele descriptive, allele interpretive and somatic interpretive. The allele descriptive section contains data elements that describe the genome position, gene, chromosome, genomic location, reference transcript and protein. The allele interpretive section contains data elements describing the somatic classification (confirmed somatic, confirmed germline or unknown), the DNA and protein substitution, the variant type and consequence and PubMed identifiers associated with interpretation. The somatic interpretive section contains the most clinically relevant data, and is the section that required the most discussion and consensus-building among the working group members. The somatic interpretive section contains a description of the cancer type (NCI Thesaurus, Oncotree, Disease Ontology), the Biomarker Class (Diagnostic, Prognostic, Predictive), the Therapeutic Context (associated drugs), Effect (Resistant, Responsive, Not-Responsive, Sensitive, Reduced-Sensitivity), Level of Evidence (a tiered system similar to the recent AMP/CAP/ASCO guidelines) (6) and Sub-Level of Evidence (reporting of trials, metadata analysis, preclinical data or inferential data). Readers are referred to the publication for a more detailed description of these data elements.

Since the publication of MVLD, recent guidelines on somatic variant interpretation have been published through a joint effort of the Association for Molecular Pathology (AMP), College of American Pathologists (CAP) and the American Society of Clinical Oncology (ASCO)(6). We intend to fully harmonize MVLD elements with these guidelines; mapping any specific criteria to the current

version of MVLD and revising MVLD to accommodate new elements. There are distinct areas of agreement between MVLD and AMP/CAP/ASCO guidelines, such as using HUGO-approved nomenclature and HGVS formatting for variants. However, there are also sizable and nuanced differences that need resolution to sync the guidelines with the MVLD data structure.

One area of immediate critical harmonization needed is in the Somatic Interpretive Level of Evidence and Sub-Level of Evidence in MVLD, which was drawn from the Cancer Driver Log (CanDL) (7). The AMP/CAP/ASCO guidelines contain classification for uncertain (Tier III) and benign (Tier IV) variants, while MVLD was not initially designed to incorporate these types of variants. However, the necessity and relevance of uncertain or benign variants is apparent in that they too can aid clinical diagnosis. The AMP/CAP/ASCO guidelines Tier I Level A and MVLD Tier 1 are the same, but AMP/CAP/ASCO further provides Level B to sustain interpretations that derive from well-established studies that are not yet FDA or NCCN approved. Similarly, there are numerous nuanced differences between AMP/CAP/ASCO Tier II Level C and D and MVLD Tier 2, 3 and 4. The Sub-Level of Evidence in MVLD is incorporated in AMP/CAP/ASCO at various Tiers as well. Instead of partially modifying the MVLD Level of Evidence and Sub-Level of Evidence, we propose to absorb the Sub-Level of Evidence element into the Level of Evidence and to fully adopt the classification system proposed by AMP/CAP/ASCO into the Somatic Interpretive Level of Evidence shown in **Figure 1**. Cancer-type might be agnostic, for example with PD-L1 testing for Keytruda.

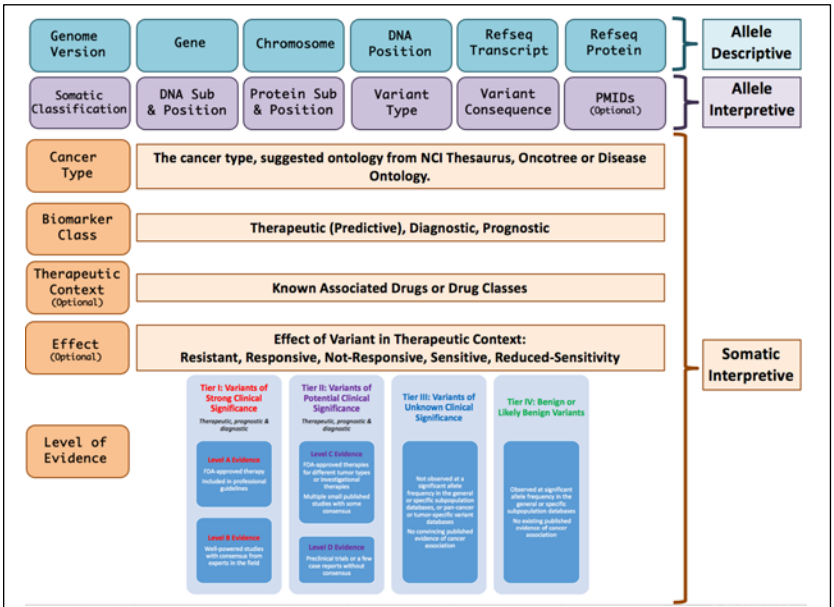


Fig.1. Mapping AMP guidelines to the ClinGen MVLD framework

In addition, there are other moderate differences. AMP/CAP/ASCO suggested the variant allele fraction of the somatic variant was important to report, and MVLD did not include a field for this. In general, MVLD did not focus on negative results or include fields for lack of supporting variant evidence, but will consider doing so in future revisions by incorporating them into the Tier III and Tier IV categories. For Biomarker Classifications, AMP/CAP/ASCO uses the term "therapeutic" which is similar to the MVLD "predictive"; we are proposing to adopt "therapeutic" as well, and will start by using a combination of both "therapeutic/predictive" to ensure clarity. AMP/CAP/ASCO did

not touch upon standardized ontology for cancer type, while MVLD proposed use of NCI Thesaurus, Oncotree or SNOMED. We will be revising MVLD to adopt Disease Ontology as well due to its popularity and prevalence.

The Somatic WG of ClinGen plans to continue reviewing the above detailed distinctions and again use a consensus-driven, round table approach to resolve and revise the MVLD to accommodate and support somatic variant interpretation guidelines. While the AMP/CAP/ASCO somatic interpretation guidelines have been very well received, consistent feedback has promoted discussions of increased granularity for the guidelines, similar to the germline variant interpretation guidelines of the American College of Medical Genetics and Genomics (ACMG). The Somatic WG of ClinGen is preparing a joint effort with ACMG and AMP/CAP/ASCO to incorporate this level of detail into somatic interpretation guidelines.

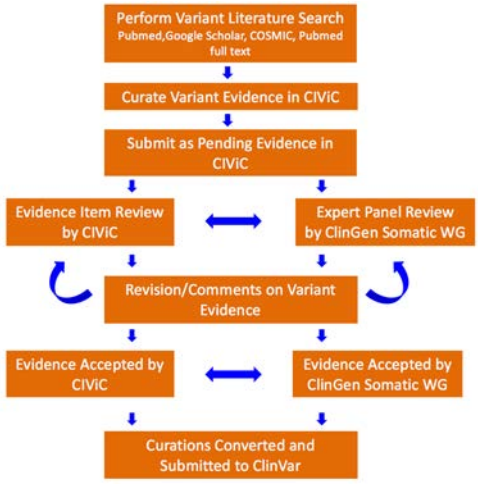


Fig. 2. ClinGen Somatic Variant Curation SOP

Variant Curation SOP and expert review

Our variant curation and interpretation process leverages the strengths of ClinGen Somatic WG, the consortium of multi-disciplinary experts in somatic variants in cancers, CIViC (8), a cancer variant knowledgebase and crowdsourced curation system and ClinVar (9), an NCBI submission-driven database for variants. ClinGen brings/develops organized clinical and biomedical expertise, best practices and SOPs, CIViC provides a curation interface and interpretation portal, and ClinVar allows widespread dissemination of the expert-curated content and provides patient-level observations of variants in clinical settings back into ClinGen/CIViC.

The ClinGen Somatic variant curation and expert review process is shown in Figure 2. New submissions or revisions are made through data entry pages in CIViC that support dynamic form adjustments, live type-ahead suggestions, ontology look-ups, and warnings for merge conflicts. Discussion pages track the complete history of comments and revisions. Curators and editors have the option to “follow” any entry (gene, variant, evidence) to receive notifications of comments, proposed changes or additions. Curators can also communicate with others in the CIViC community directly through site mentions, updates and messages. All curated entries can be “flagged” for problems or revisions can be proposed. Flagging allows for easy marking of content needing immediate review or users can flag entries which require more caution with use as diagnostic markers, while revisions are tracked and displayed with detailed GitHub-style diffs and comments.

Curators can create detailed profiles so that their efforts are recognized by awarding badges for curation activity milestones to encourage and recognize participation. Curators can also join formal curation organizations within the CIViC community, for example the ClinGen Somatic Working Group exists as a CIViC organization, currently with 12 active members.

The high quality of CIViC content is encouraged through several mechanisms. First, content creation is completely transparent with detailed provenance for all additions and revisions. Problems can be identified quickly by publicly accessible comment or flag features. Anyone can become a curator with the ability to flag, comment, or submit new/revised content. To ensure all content is reviewed by at least 2 CIViC reviewers, new entries must be reviewed by a site editor and users cannot accept their own contributions. There is an additional layer of expert review by the ClinGen Somatic WG disease or gene centric taskforces, which are described in Community Engagement.

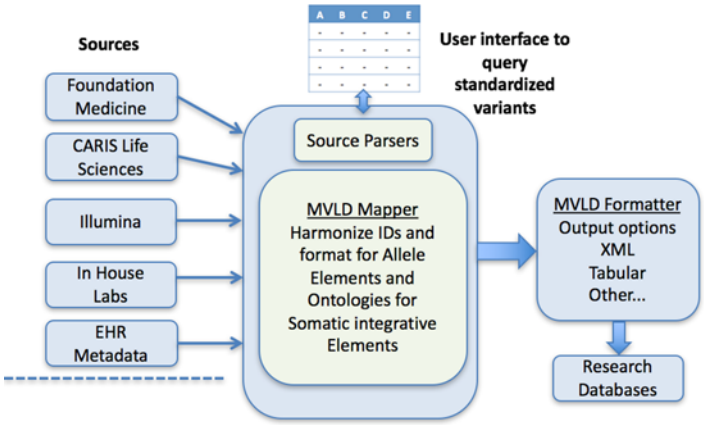


Fig. 3. Cancer MolDx – MVLD software schematic

MolDx2MVLD mapping tools

To complement the crowdsourced expert variant curation process, members of the ClinGen WG are designing and implementing tools to support mapping of clinical MolDx data to standards and automated importation of this data into research databases, for example, G-DOC (10) (Georgetown Database of Cancer) and SEER (11) to drive new hypothesis generation for translational research. The primary goal of this tool is to enable broad sharing of de-identified MolDx data from clinical laboratories for novel hypothesis generation and evidence collection for clinical actionability.

The general components of the tool to map MolDx data to the MVLD system are outlined in Figure 3. The four main components of the tool are: 1) ETL (Extract, Transform, Load) tools to parse individual sources, extract and format information required for MVLD descriptive and interpretive elements; 2) MVLD Mapper to map the extracted information to the MVLD standard, harmonize the elements to standard identifiers and ontologies used in public data repositories, identify missing data elements, and attempt to fill in missing values if possible; 3) a simple QA/QC interface for checking results and correcting or adding missing values; and 4) MVLD formatter to output the information in

various formats, e.g., xml, tabular, or others to interface with EHRs and translational research databases depending on user needs.

ETL tools

MolDx laboratory sources use different internal formats and standards to identify a variant and describe the results of their tests. For example, some labs use only gene names and protein changes as a description; others provide transcript identifiers and the specific DNA change in the transcript and/or the exact chromosomal location. To date we have not seen a report using a complete HGVS (Human Genome Variation Society) formatted sequenceID+variation name, though some labs provide the information to create the description. There are also differences in how data is distributed, with some labs providing results in tabular formats while others provide XML. We will create parsers and logic for each source to extract and perform initial transformations. **Figure 4** shows an example of an XML excerpt from the Foundation Medicine report on a patient with breast carcinoma. This patient's molecular testing identified a variant E545K in the PIK3CA gene with potential benefit from mTOR inhibitors such as Everolimus or Temsirolimus. The figure shows mapping of XML output from the lab to elements in the MVLD standard. Similar mapping is being conducted for all 18 data elements in MVLD to various commercial labs. Lab formats to integrate are prioritized by our stakeholder community based on most widely used labs.

MVLD mapper

This is the core component of the system. We expect variations in how the descriptive and interpretive properties in MVLD will be expressed, and the mapper attempts to harmonize these representations to the most

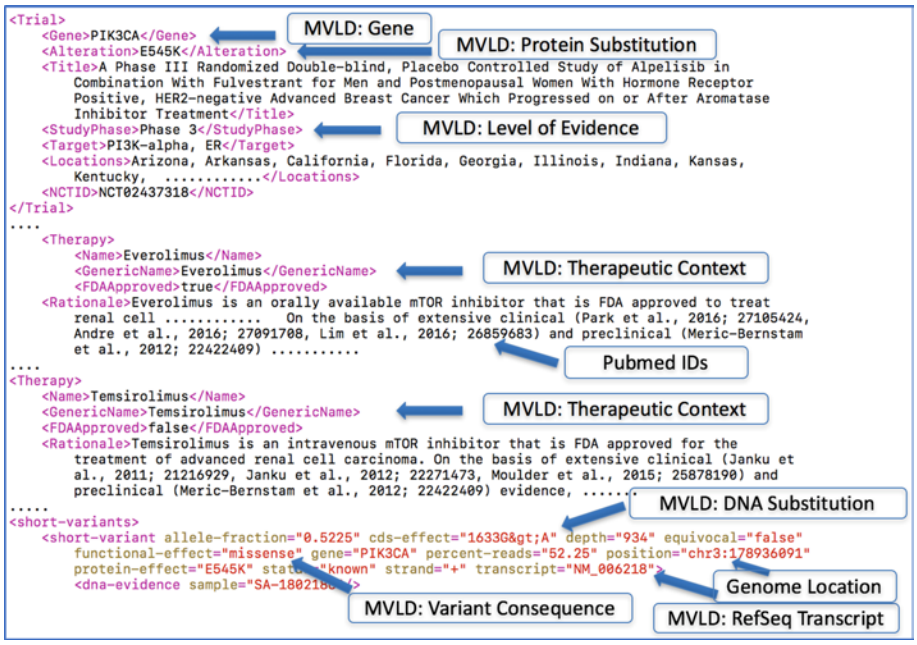


Fig. 4. A sample mapping of FM XML file to MVLD standard is shown. Not all lab reports are as well structured

informative common representation using variant-specific APIs connected to international variation databases and ontology servers. For allele properties, we use NCBI's E-utilities including their new variation API services that allow users to compare and return all equivalent alleles using multiple NCBI identifiers including a canonical identifier (12) . The tool first checks the ClinVar database to see if a representation of the variation exists or the test is registered with the Genetic Testing Registry, in which case ClinVar may contain all the variants tested (13). ClinGen has released an Allele Registry with APIs that also attempts to link equivalent variant alleles to a canonical representation but is not NCBI centric in that ENSEMBL IDs and even ExAC alleles are supported and novel alleles can be submitted (14).

MVLD Interpretive elements like cancer type can be standardized using APIs for terminology servers. The WG identifies terminology standards used in key data fields in lab reports, such as disease and drug names, ICD codes, or the 10-digit national drug code and map them to MVLD-recommended standards. If any term is unknown, the MolDx processor attempts to automatically map to terminologies in the NCI Thesaurus (NCIt) using LexEVS Terminology Server APIs or BioPortal APIs (15, 16). In all cases, the original value is stored along with the selected mappings to defined terminologies. This allows the data to be in a uniform format that follows standards and ontologies, while allowing for more integrated search functions within systems like ClinVar.

MVLD Formatter

The formatter module provides multiple output options for target research databases. The Initial output will be delimited tables or XML mainly for consumption by institutional databases (EHRs) that want to store the MVLD standardized data. We will work with the community to define and build in additional XML or other formats (e.g. JSON) from labs.

Community Engagement

Many standards and frameworks fail due to lack of community engagement and adoption. To avoid this, the ClinGen clinical domain working group engaged both strategic leaders and tactical implementers from over 60 cancer centers, industry partners and federal agencies. These include active participation from organizations such as Georgetown University Lombardi Cancer Center, Baylor College of Medicine, Vanderbilt University Medical Center, Washington University School of Medicine, Moffitt Cancer Center, Illumina, Molecular Match, NCI, NHGRI and FDA. An initial survey of participating organizations identified major challenges in somatic variant assessment, clinical interpretation pipelines and open tools for variant curation and expert review. The survey results also indicated the use of a tiered system of variants for clinical actionability (FDA-approved/NCC guidelines, clinical trials data, pre-clinical data, mechanistic/pathway level evidence).

These results motivated the efforts to develop the MVLD to help standardize how clinical labs report MolDx data to patients, clinicians and regulatory agencies. We also engaged members from AMP (Association of Molecular Pathologists) and CAP (College of American Pathologists) somatic practice guideline committees to help drive adoption of MVLD within their professional societies and members. ClinGen Somatic WG is also actively working with Global Alliance for Genomic Health (GA4GH)'s Variant Interpretation for Cancer Consortium (VICC). The VICC seeks to integrate global efforts for the clinical interpretation of cancer variants. The ClinGen Somatic WG engages various experts in the cancer research and care communities through taskforces. These taskforces are self-organized expert groups in a particular cancer type, gene or a pathway. Three such taskforces that are currently operational are described below with other taskforces being routinely formed.

Pediatric somatic taskforce

Cancers in children differ from those in adult patients in at least three key aspects. First, a number of childhood cancers occur exclusively in children and adolescents. The genomic landscape of such tumors, including medulloblastoma, neuroblastoma, and Wilms tumors, have been the focus of several recent large-scale research projects, such as the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (17), the Pediatric Cancer Genome Project (PCGP) (18), the Peds-MiOncoSeq (19) and the BASIC3 (20) projects, and have been found to harbor not only shared but several novel recurrent alterations when compared to the most common genomic alterations seen in adult epithelial malignancies. Such alterations include, in addition to SNVs and indels, a higher preponderance of gene fusions and copy number alterations. Second, genomic profiling of certain childhood cancers with particularly poor outcomes and with well-recognized adult counterparts such as glioblastoma have revealed remarkable non-overlap in the key genetic drivers and signaling pathways between the same disease histology in the two age groups (21). These studies highlight the diversity of genomic alterations that are exclusive or enriched in childhood cancers. Assessment of the extent of curation of key childhood cancer genes and variants in CIViC reveal many of the most common childhood cancer specific genes (e.g., *H3F3A*, *HIST1H3B*, *ACVR1*, *ATRX*, *DDX3X*, *DROSHA*, *DICER*) and variants to be absent or sparsely curated in CIViC and other cancer knowledgebases highlighting the urgent need for expert curation of novel childhood cancer genes. Third, even when childhood tumors are found to harbor therapeutically targetable alterations that have been validated in clinical trials of adult patients, the feasibility and efficacy of such treatment options in children are often unclear; expert curation of such 'actionable' alterations in childhood cancers will require the insight of pediatric oncologists and other domain-specific experts in pediatric oncology. The ClinGen Somatic Pediatric taskforce is addressing these issues through the expert variant curation and adjudication process described above.

Pancreatic somatic taskforce

Pancreatic ductal adenocarcinoma (PDAC) is one of the most difficult cancers to treat, as physiological hurdles to preventative monitoring and an absence of early detection biomarkers result in many late stage diagnoses. The more advanced disease is typically metastatic and chemoresistant, with a 5-year survival is around 8% (22). The lack of noninvasive, PDAC specific biomarkers highlights the need to identify novel molecular signals associated with onset and progression, and to develop a better understanding of the role of somatic variation to improve individualized therapy decisions. The last stage diagnosis of PDAC typically results in the convergence of mutations in several known oncogenes and tumor suppressor; mutations activating KRAS are found in 95% of cases, inactivation of TP53 occurs in 75% of cases, and CDKN2A is inactivated in 95% of cases (23). These genes are known for their regulatory influence on cell proliferation and are particularly relevant to regulating a shift in active metabolic pathways promoting growth in a hypoxic microenvironment. This metabolic reprogramming found in PDAC not only increases glucose uptake and enhances glycolysis (24), but also increase the expression of glutamine metabolic pathways and promote production of NADPH/NADP+ (25). These cellular mechanisms are adapted to meet the energy requirement for cell division at a reduced oxygen consumption rate and represent an emerging set of PDAC targets that require somatic curation (26).

The occurrence of somatic variants in PDAC which overlap with known oncogenes and tumor suppressors provide the opportunity for targeted therapies which have proven effective in other cancers. In particular, a mutator phenotype of PDAC associated with a high mutation burden is driven by mutation to mismatch repair genes (27), and may benefit from any number of therapies developed to target cancers with DNA repair deficiencies. By extension, somatic variants found in other PDAC subtypes could benefit from successful advances in other cancerous tissues, and underscores the need for their high quality curation. Somatic variants in PDAC are being cataloged as part of the PANCAN Know Your Tumor program(1), and to date have identified 5473 unique variants from 432 genes. Overall, the variants remain uncurated with about 38% of genes represented in CIViC, covering only 1-2% of variants observed in PDAC. Clearly there is a significant need for the expert curation of novel PDAC somatic variants, and the potential for improvement over current therapy options.

TP53 taskforce

Tumor suppressor genes play important roles in cancer biology and could be inactivated by number of different mechanisms, such as deletions, insertions, inversions, loss of function point mutations, and loss of expression. Tumor suppressor gene TP53 is one of the most frequently altered genes across multiple tumor types. To fulfil its proper biological function four identical TP53 polypeptides must form a tetramer which functions as a transcription factor. Mostly due to this

underlying molecular biology majority of TP53 inactivating alterations are loss of function point mutations. Majority of the loss of function point mutations are concentrated in the DNA binding domain and have different degree of dominant negative phenotype. MDM2-TP53 interaction inhibitors efficacy is currently under investigation in multiple clinical trials. Inactivating TP53 mutations prevent on target activity and efficacy of MDM2-TP53 interaction inhibitors, therefore only patients with intact TP53 could benefit from such inhibitors. Using preclinical (28) and clinical publications (29) on MDM2-TP53 interaction inhibitors efficacy in relation to TP53 status corresponding evidence items have been submitted to CIViC. At this point in time CIViC does not support entry of information on pathogenicity of somatic alterations; however this functionality will be added in the future. Due to current lack of ability to enter pathogenicity information MDM2-TP53 interaction inhibitors related evidence items for individual alteration have been limited to the ones which are directly mentioned in Saiki et al. 2015 publication (30).

Future directions

A future goal of the ClinGen Somatic Working Group is to review and harmonize existing guidelines and guideline efforts related to curation, interpretation, and reporting of somatic alterations in cancer to provide a unified guideline to clinical labs to represent and interpret cancer variants. The working group is forming a task force that will bring together representatives from ClinGen, ACMG, AMP, ASCO and CAP, among other relevant organizations. This harmonization task force will seek to build upon recently published work such as the AMP/ASCO/CAP guidelines and the ClinGen Somatic Working Group's MVLD for curation of somatic alterations in cancer. Further, this task force will work to make its standard compatible with other related guidelines, such as the ACMG/AMP guideline for interpretation of germline variants (31), and an ongoing effort within ACMG on the interpretation of copy number variants in neoplastic diseases. The task force will review and include work such as the Variant Interpretation in Cancer Consortium and AACR GENIE's efforts to describe and standardize curation practices in the somatic cancer space (32). Finally, the task force will review and include efforts of somatic cancer knowledgebases to map terminologies and levels of evidence schemes across knowledgebases (e.g., CIViC, OncoKB, PMKB, JAX-CKB, CGI, PCT, CanDL, G-DOC). The ClinGen gene and disease focused taskforces will provide usecases and variant examples to this harmonization effort, enabling a continually updated, unified guideline for somatic variant interpretation agreed upon by experts and serving the cancer MolDx community in their mission to democratize access to these important clinical datasets.

Acknowledgements – Authors acknowledge support from the ClinGen grant NIH U01 HG007437. CIViC is supported by NIH U01 CA209936. MG is supported by NIH R00 HG007940. OLG is supported by NIH K22 CA188163.

References

1. PANCAN. Know Your Tumor 2017 [Available from: https://www.pancan.org/facing-pancreatic-cancer/patient-services/know-your-tumor.
2. Samuel N, Villani A, Fernandez CV, Malkin D. Management of familial cancer: sequencing, surveillance and society. Nat Rev Clin Oncol. 2014;11(12):723-31.
3. Meric-Bernstam F, Brusco L, Shaw K, Horombe C, Kopetz S, Davies MA, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. J Clin Oncol. 2015;33(25):2753-62.
4. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235-42.
5. Ritter DI RS, Roy A, Rao S, Landrum MJ, Sonkin D, Shekar M, Davis CF, Hart R, Micheel C, Weaver M, Allen EV, Parsons DW, McLeod HL, Watson MS, Plon SE, Kulkarni S, Madhavan S. Somatic Cancer Variant Curation and Harmonization through Consensus Minimum Variant Level Data. Genome Medicine. 2016.
6. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19(1):4-23.
7. Damodaran S, Miya J, Kautto E, Zhu E, Samorodnitsky E, Datta J, et al. Cancer Driver Log (CanDL): Catalog of Potentially Actionable Cancer Mutations. J Mol Diagn. 2015;17(5):554-9.
8. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170-4.
9. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862-8.
10. Bhuvaneshwar K, Belouali A, Singh V, Johnson RM, Song L, Alaoui A, et al. G-DOC Plus - an integrative bioinformatics platform for precision medicine. BMC Bioinformatics. 2016;17(1):193.
11. Altekruse SF, Rosenfeld GE, Carrick DM, Pressman EJ, Schully SD, Mechanic LE, et al. SEER cancer registry biospecimen research: yesterday and tomorrow. Cancer Epidemiol Biomarkers Prev. 2014;23(12):2681-7.
12. NCBI. Services for variation data processing 2017 [Available from: https://api.ncbi.nlm.nih.gov/variation/v0/.
13. NCBI. New Web Services for Comparing and Grouping Sequence Variants 2017 [Available from: https://ncbiinsights.ncbi.nlm.nih.gov/2017/02/09/new-web-services-for-comparing-and-grouping-sequence-variants/.
14. Patel RY, Shah N, Jackson AR, Ghosh R, Pawliczek P, Paithankar S, et al. ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. Genome Med. 2017;9(1):3.
15. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37(Web Server issue):W170-3.
16. Wiki, N. LexEVS Servers and APIs Summary 2014 [Available from: https://wiki.nci.nih.gov/display/EVS/LexEVS+Servers+and+APIs+Summary.

17. Walz AL, Ooms A, Gadd S, Gerhard DS, Smith MA, Guidry Auvil JM, et al. Recurrent DGCR8, DROSHA, and SIX homeodomain mutations in favorable histology Wilms tumors. Cancer Cell. 2015;27(2):286-97.
18. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, et al. Novel mutations target distinct subgroups of medulloblastoma. Nature. 2012;488(7409):43-8.
19. Mody RJ, Wu YM, Lonigro RJ, Cao X, Roychowdhury S, Vats P, et al. Integrative Clinical Sequencing in the Management of Refractory or Relapsed Cancer in Youth. JAMA. 2015;314(9):913-25.
20. Parsons DW, Roy A, Yang Y, Wang T, Scollon S, Bergstrom K, et al. Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. JAMA Oncol. 2016.
21. Sturm D, Pfister SM, Jones DTW. Pediatric Gliomas: Current Concepts on Diagnosis, Biology, and Clinical Management. J Clin Oncol. 2017;35(21):2370-7.
22. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin. 2017;67(1):7-30.
23. Lu S, Ahmed T, Du P, Wang Y. Genomic Variations in Pancreatic Cancer and Potential Opportunities for Development of New Approaches for Diagnosis and Treatment. Int J Mol Sci. 2017;18(6).
24. Ying H, Kimmelman AC, Lyssiotis CA, Hua S, Chu GC, Fletcher-Sananikone E, et al. Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. Cell. 2012;149(3):656-70.
25. Son J, Lyssiotis CA, Ying H, Wang X, Hua S, Ligorio M, et al. Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Nature. 2013;496(7443):101-5.
26. Hardie RA, van Dam E, Cowley M, Han TL, Balaban S, Pajic M, et al. Mitochondrial mutations and metabolic adaptation in pancreatic cancer. Cancer Metab. 2017;5:2.
27. Witkiewicz AK, McMillan EA, Balaji U, Baek G, Lin WC, Mansour J, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun. 2015;6:6744.
28. Efeyan A, Ortega-Molina A, Velasco-Miguel S, Herranz D, Vassilev LT, Serrano M. Induction of p53-dependent senescence by the MDM2 antagonist nutlin-3a in mouse cells of fibroblast origin. Cancer Res. 2007;67(15):7350-7.
29. Andreeff M, Kelly KR, Yee K, Assouline S, Strair R, Popplewell L, et al. Results of the Phase I Trial of RG7112, a Small-Molecule MDM2 Antagonist in Leukemia. Clin Cancer Res. 2016;22(4):868-76.
30. Saiki AY, Caenepeel S, Cosgrove E, Su C, Boedigheimer M, Oliner JD. Identifying the determinants of response to MDM2 inhibition. Oncotarget. 2015;6(10):7701-12.
31. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015.
32. Micheel CM CD, Gao J, Maurer I, Miller C, Shaw KR, Levy MA, Schultz N, on behalf of the AACR Project GENIE Consortium. Clinical actionability and clinical trial matching for GENIE patient genotypes using My Cancer Genome, Personalized Cancer Therapy, and OncoKB. AACR Annual Meeting 2017, Washington DC ed.

A heuristic method for simulating open-data of arbitrary complexity that can be used to compare and evaluate machine learning methods*

Jason H. Moore, Maksim Shestov, Peter Schmitt, Randal S. Olson
Institute for Biomedical Informatics, University of Pennsylvania, D202 Richards Building, 3700 Hamilton Walk, Philadelphia, PA 19104
Email: jhmoore@upenn.edu

A central challenge of developing and evaluating artificial intelligence and machine learning methods for regression and classification is access to data that illuminates the strengths and weaknesses of different methods. Open data plays an important role in this process by making it easy for computational researchers to easily access real data for this purpose. Genomics has in some examples taken a leading role in the open data effort starting with DNA microarrays. While real data from experimental and observational studies is necessary for developing computational methods it is not sufficient. This is because it is not possible to know what the ground truth is in real data. This must be accompanied by simulated data where that balance between signal and noise is known and can be directly evaluated. Unfortunately, there is a lack of methods and software for simulating data with the kind of complexity found in real biological and biomedical systems. We present here the Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) method and prototype software for simulating complex biological and biomedical data. Further, we introduce new methods for developing simulation models that generate data that specifically allows discrimination between different machine learning methods.

Keywords: simulation, machine learning, open data.

1. Introduction

Simulation plays an important role in the development of computational and statistical methods because the ground truth is known. This allows the power and false-positive rate of methods to be evaluated and compared. Further, simulation allows features of the data such as size and complexity to be varied to evaluate method robustness. An important criticism of simulation is that the models used may not fully represent the complexity of both the noise structure and signal in the data. This is because the nature of the true signals in data derived from experimental or observational studies of biological or biomedical systems is usually not known. This is particularly true in genetics and genomics where we have barely scratched the surface of measuring all the components that might influence phenotypic variation and, for those measures that exist, have focused our analysis primarily on univariate effects. This sparse slice of possible etiological factors and simplistic analytical approaches have yielded a number of simulation methods that are based parametric statistical methods such as logistic regression or probabilistic methods such as penetrance functions. The goal of the present study is to develop heuristic methods for the discovery of complex biological systems models that can be used to simulate more realistic data. We focus here on the simulation of

* This work is supported by National Institutes of Health grants LM012601, AI116794, and DK112217.

genotypic and phenotypic data in samples derived from human populations that can facilitate the development of methods for genetic association studies and precision medicine.

Methods for simulating genetic data in human population-based studies fall into two general categories. The first set of methods are focused on generating patterns of genetic variation that might be found in human populations. The goal here is to approximate the allele and genotype frequencies that are expected in a human population and the correlation structure of the variants as shaped by selection and recombination. Forward-time simulators of human populations such as simuPOP [1-2], GenomeSIMLA [3-5] and SFS_CODE [6] can be used for this purpose. Genetic data simulated is this way can then be used to simulate phenotypes using a statistical or computational model. Simple additive effects can be simulated using a linear regression model or more complex genetic effects such as gene-gene interactions can be simulated using methods and software such as Epi2Loc [7] or GAMETES [8]. Although useful, these tools don't explicitly build their models using a framework that approximates the hierarchical complexity of biological systems. It is our working hypothesis that the simulation of data using biologically-realistic genotype-phenotype relationships will improve method development by more closely mimicking the hierarchical complexity of human health or model systems.

To address this concern, we previously introduced the Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) method and prototype software for simulating complex biological and biomedical data [9]. This approach combines a biological hierarchy, a flexible mathematical framework, a liability threshold model for defining disease endpoints, and a simple stochastic search strategy for identifying high-order gene-gene interaction models of disease susceptibility. HIBACHI allows the explicit definition of a biological framework for the propagation of genetic effects organized in gene regulatory regions, coding regions, noncoding regions, and interacting components such as genes for transcription factors and other regulatory sequences such as microRNAs or long noncoding RNAs. The product at the gene level is a quantitative trait (e.g. protein) that can then be used with a liability model to simulate disease status. Alternatively, multiple traits can be simulated from different sets of genetic variation at the gene level and then combined with additional functions to simulate a higher-order physiologic trait at the level of a pathway or system. Traits from multiple systems could be used to simulate one or more anatomical traits. HIBACHI provides the flexibility to map multiple genotypes to multiple phenotypes through hierarchical biological systems of any complexity. It is our working hypothesis that data simulated in this manner will be more biologically realistic and thus more useful for the evaluation of computational and statistical methods.

The prototype HIBACHI algorithm and software developed by Moore et al. [9] used a fixed biological architecture with mathematical functions that connected genotype with phenotype. The goal of the present study was to extend HIBACHI to include heuristic identification of both the wiring of the biological model and the mathematical functions that create the genotype to phenotype relationship. Because of the extensive size of the search space we chose to use stochastic search for model discovery. Further, due to the modular nature of HIBACHI models we selected genetic programming (GP) as an initial search algorithm because of its flexible representation of models as expression trees and because it inherently explores combinations of model subcomponents through its recombination operator. We describe the stochastic search engine using GP in detail and then

provide an example application that focuses on the discovery of genetic models that produce data that explicitly differentiate the performance of different machine learning algorithms. The results demonstrate the usefulness of HIBACHI with GP as a stochastic search engine for the evaluation of machine learning methods. Methods and software like this will play an important role in the post-genomics data science era focused on understanding the complexity of genotype-phenotype relationships with the end goal of precision medicine.

2. Methods

There are five components to our Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) simulation method. The first is the metaphor for the hierarchical biological framework that transmits information from genotype at the DNA sequence level through biomolecular interactions at the gene, cell, and pathway levels to a clinical endpoint. The second is the mathematical framework that generates the genotype to phenotype relationship or pattern. The third is the liability threshold model that is used to define disease status. The fourth is the genetic programming (GP) methods for the discovery of high-order models. The final component is an open-source python-based software package distributed via GitHub for simulating multiple data sets. We describe each of these in turn. The first three components are descriptions included in the work by Moore et al. [9] and repeated with some minor updates here for completeness. The last two describe new components to the method and software.

2.1. *A biology-based framework for simulation of complex biological systems*

The goal of this component is to provide a biological framework or scaffold to serve as a metaphor for genetic variants and their phenotypic relationships propagated through a hierarchical set of mathematical functions. The prototype for HIBACHI developed by Moore et al. [9] used a fixed architecture that was based on genetic effects at the gene level. We describe this framework first and then discuss how this relates to the new approach that uses GP to discover both the wiring diagram and the mathematical functions. The initial HIBACHI framework (see Figure 1) started with protein-coding gene (i.e. mRNA gene) with a single non-synonymous genetic variant that is assumed to change an amino acid. Upstream of the mRNA gene is a promoter with a single regulatory variant and an enhancer with a single regulatory variant. Also included in our initial framework are two genes that code for transcription factors that bind to the regulatory region. We included a protein-coding variant in the gene that codes for each transcription factor. We also included a single variant in a microRNA gene that participates in post-translational regulation. In total, this structure allowed for six genetic variants (coded 0, 1, 2) all influencing a protein product as a quantitative trait. In addition, we included an environmental factor (coded -2, -1, 0, 1, 2) to allow for non-genetic variation in the phenotypic values. It is important to note that this particular biological framework was a preliminary proof of concept. The goal of the present study is to allow this framework to vary as part of the search for models meeting certain objectives using GP. The metaphor still holds but the new GP-based systems allows for much greater flexibility in the size and shape of the models being generated. Other metaphors such as electronic health record (EHR) data could also be used here.

2.2. *A mathematical framework for simulation of complex patterns*

The goal of this component is to provide a flexible mathematical framework for combining features to produce an endpoint. Using the genetics metaphor, genotypic and non-genotypic values to produce phenotypic values. Each biology-based locus feeds into a mathematical function whose result is carried forward to the next function. For example, one transcription factor locus combines with the enhancer locus through a function whose result then combines with the second transcription factor. The result of this operation combines with the locus at the promoter. This result combines with the coding variant in the gene. This result combines with the microRNA locus. This result combines with the environmental factor to produce a protein product. Thus, the protein expression value is dependent on mathematical functions of the six loci and the environmental factor. This produces a distribution with several to millions of possible phenotypic values for most combinations of functions that can then be used with the liability threshold model described below to generate disease status. Cases and controls can then be sampled from this distribution or the continuous output can be used directly as a quantitative trait.

For each run the user can specify a set of mathematical function to use to build the models. Examples of basic functions include addition, subtraction, multiplication, division, modulus, and modulus-2. Logical functions include greater than, less than, AND, OR, and XOR. Bitwise functions include bitwise AND, bitwise OR, and bitwise XOR. Unary functions include absolute value, NOT, factorial, left and right. Large functions include power, log, permute, and choose. Miscellaneous functions include minimum and maximum.

2.3. *A liability threshold model for biology-based simulation*

We use a liability threshold model to simulate disease from the distribution of phenotypic values generated from the genotypic values and mathematical functions as described above. The user can select the liability threshold to achieve a particular disease prevalence. More details about the liability model are provided by Moore et al. [9].

2.4. *Genetic programming for model representation and stochastic search*

We selected genetic programming (GP) as our stochastic search engine for several reasons. First, GP uses binary expression trees that are a convenient data structure for representing HIBACHI models. This makes the models very easy to manipulate and evaluate computationally. Flexible representation is a known advantage of GP [10] Second, GP uses a recombination operator that explicitly swaps subcomponents of expression trees to generate variability in the solutions as part of its iterative search process. This is appealing because HIBACHI models are hierarchical in nature with modular subcomponents representing different biological processes such as transcription factor binding, miRNA regulation of transcription, etc. The ability to mix and match these genomic modules facilitates the development of new models that meet a simulation objective. An introduction to GP is provided by Poli et al. [11] in an open-access book for those seeking additional details of the method.

A key to GP search is the fitness function that specifies the value or quality of a particular set of mathematical functions and their wiring (i.e. the model) represented as an expression tree.

For the evaluation of this approach, we used the performance of pairs of machine learning algorithms as the primary fitness criteria (see below). In addition, we also use as an objective the complexity of the model with the idea that simpler models that meet the data objective are better (i.e. to be minimized). We balance these different objectives using Pareto optimization. We discuss below an application of HIBACHI to the discovery of models that can be used for evaluating machine learning methods.

2.5. *An open-source HIBACHI software package*

The HIBACHI software presented here was programmed entirely in Python. The stochastic search elements used the open-source Distributed Evolutionary Algorithms in Python (DEAP) framework available on GitHub. We used the Python-based scikit-learn machine learning library to carry out all analyses (scikit-learn.org). HIACBHI is available as open-source on GitHub (github.com/epistasislab/hibachi).

2.6. *Discovery and evaluation of HIBACHI models for evaluating machine learning methods*

Our goal is to use HIBACHI to discover models that generate data for which one machine learning algorithms perform better compared to another. We chose to focus on logistic regression, decision trees, and random forests as three commonly used methods. All machine learning was implemented using the open-source scikit-learn library in Python. For each model and dataset generated by HIBACHI we used pairs of machine learning algorithms with default settings to perform the analysis. The balanced accuracy of each analysis was reported and used as multiple objectives through Pareto optimization to maximize the performance of one method while minimizing the performance of the second method. All combinations of methods were evaluated. The difference in performance between the two methods is also combined with a fitness objective that attempts to minimize the complexity of the models in terms of the number of mathematical functions that are used.

3. Results

The HIBACHI method and software allows for simulation of complex datasets with specific properties indicated in the multi-objective fitness function of the genetic programming (GP) algorithm used for model representation and stochastic search. In order to simulate datasets with varying machine learning performance we used a three-objective fitness function, first objective tries to increase the difference in performance of two machine learning methods, second objective optimizes the performance of one of the two methods, and the third objective decreases the complexity of the underlying mathematical function that generates dataset labels.

Performance of machine learning methods was measured by the average area under the curve of the testing data ROC curve of 10 random 80/20 train/test data splits. The simulated datasets consist of 2000 samples with 10 features per sample with an even 50/50 split of cases vs. controls. Each feature is an integer from a set of {0, 1, 2} which corresponds to an encoding of genetic variants. GP algorithm was ran for 1000 generations with 1000 individuals evaluated at each generation.

Increasing generation number would only marginally improve results due to a strong drop off in fitness improvements past 200 generations.

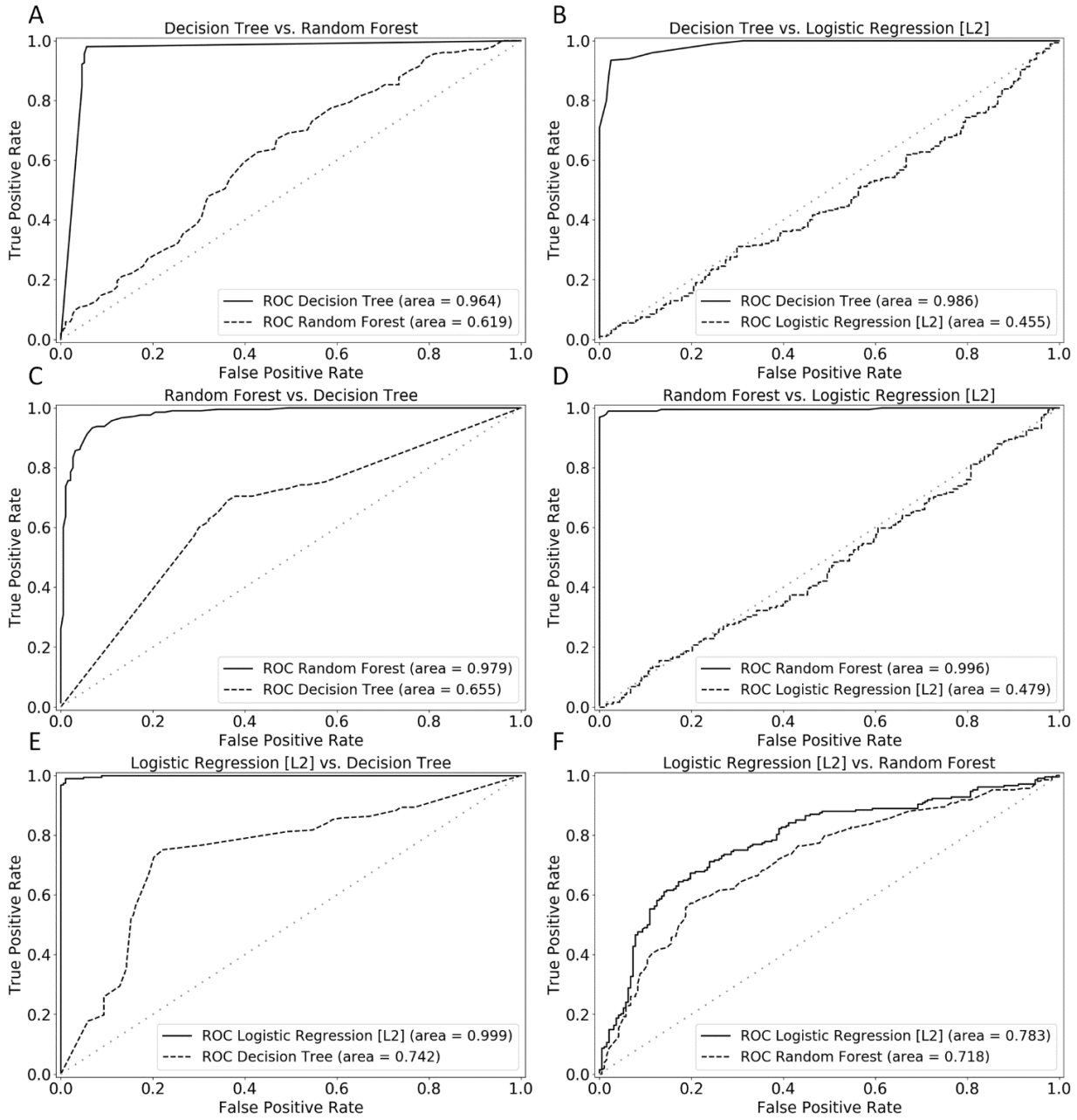


Fig. 1. ROC curves of datasets generated using HIBACHI with a three-objective fitness function. (A) ROC curve Decision Tree vs. Random Forest. (B) ROC curve Decision Tree vs. Logistic Regression. (C) ROC curve Random Forest vs. Decision Tree. (D) ROC curve Random Forest vs Logistic Regression. (E) ROC curve Logistic Regression vs. Decision Tree. (F) ROC curve Logistic Regression vs. Random Forest

We were able to generate synthetic datasets that have tailored performance for any given machine learning methods. Figure 1 shows the ROC curves of datasets generated with HIBACHI using 3 different machine learning methods that include a linear method – L2 penalized logistic regression with default parameters, non-linear method – Decision Tree with default parameters, no

depth limit, and a non-linear ensemble method – Random Forest with default parameters, no depth limit, and 100 estimators. The performance of non-linear vs. linear methods (Figure 1B&D) is as expected, once higher order interactions are introduced via XOR or other non-linear operator in the underlying mathematical function the performance drops to random performance. Figure 2 shows the function tree that gives a perfect performance for the decision tree algorithm and random performance for logistic regression. The reverse where we try to optimize a linear method vs. non-linear (Figure 1E&F), the performance difference is not as significant and the non-linear method never approaching random performance. Most of the operators in the linear vs. non-linear mathematical function are addition and subtraction. Mathematical function of the non-linear vs ensemble methods (Figure 1A) shows a similar performance to linear vs. non-linear method optimizations with no unique operators.

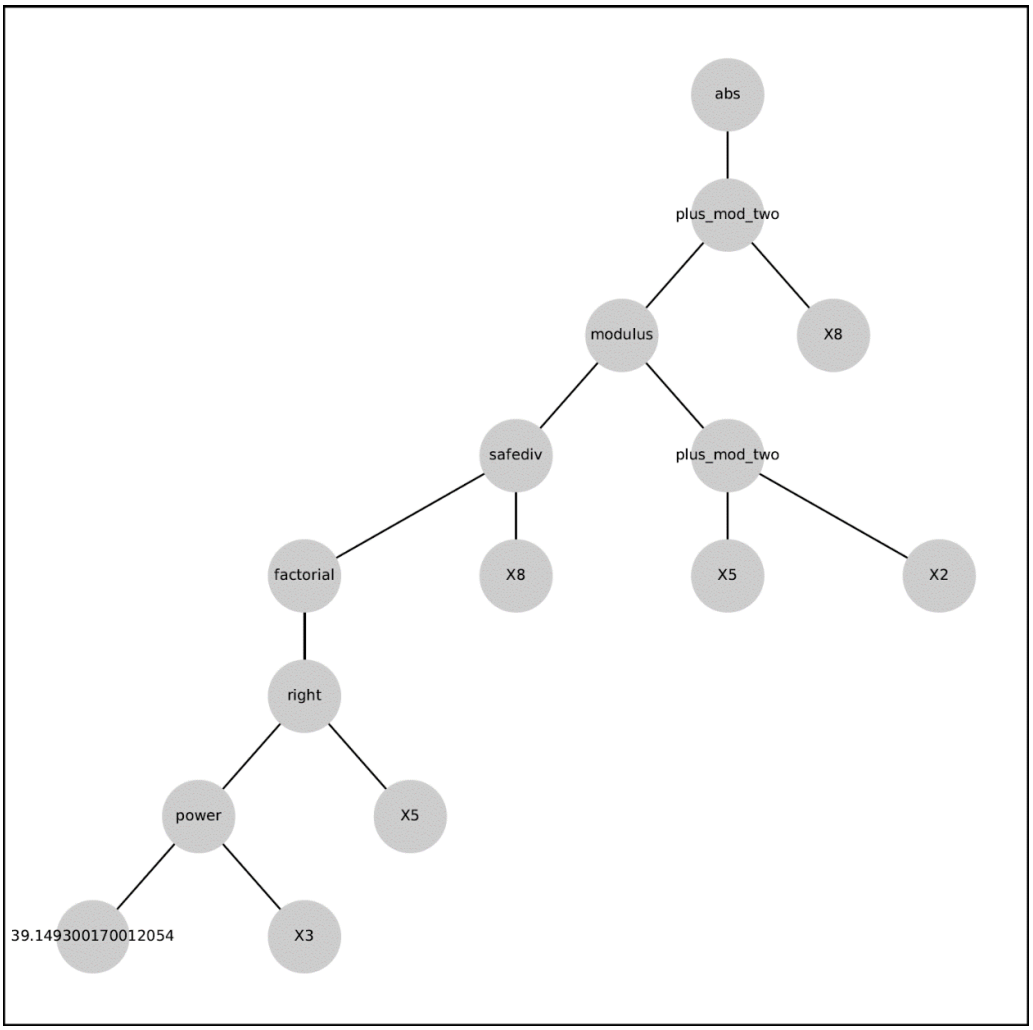


Fig. 2. Mathematical function tree that generates that optimizes decision tree performance (test set ROC AUC of ~1) and minimizes logistic regression performance (test set ROC AUC of ~0.5). X# corresponds to one of the 10 features.

4. Discussion

The generation and open sharing of simulated data is an important part of the artificial intelligence and machine learning development and evaluation process. Unfortunately, there is a lack of methods and software for generating data with the complexity that we often observe in biological and biomedical systems. To address this concern, we previously developed the Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) method and prototype software for simulating complex biological and biomedical data [9]. This approach combines a biological hierarchy, a flexible mathematical framework, a liability threshold model for defining disease endpoints, and a simple stochastic search strategy for identifying high-order gene-gene interaction models of disease susceptibility. We present here an extension of HIBACHI that improves its flexibility for generating models of different complexity and for generating models that can be used to evaluate and compare machine learning algorithms.
We also introduce a new Python-based software package as open-source.

The results demonstrate that it is possible to use HIBACHI to discover mathematical models that generate data for which logistic regression, decision trees, and random forests perform differently. As expected, it was quite easy to simulate data for which decision trees and random forests perform dramatically better than logistic regression and difficult to simulate data for which logistic regression does better than random forests. HIBACHI was able to generate data that revealed significant performance differences between other method contrasts. These results demonstrate the value of HIBACHI for using simulation to develop and test machine learning methods.

There are several possible directions for future studies with HIBACHI. First, HIBACHI models can easily be integrated together to build hierarchical models that might more closely reflect the hierarchy of real systems. For example, multiple HIBACHI models could be generated one for each gene in a genetic system with the output representing the continuous distribution of a protein product. Protein products could be combined with additional functions to produce the output of a biochemical system. Biochemical system output could be combined with additional functions to produce the output from a physiological system. These could be combined to produce a liability distribution for health and disease. Second, HIBACHI could be used to generate electronic health record (EHR) data to facilitate the development of methods for the rapidly growing field of clinical informatics. Good method for simulating EHR data are not available. Third, HIBACHI could be used to generate comprehensive sets of open-access data of differing size, shape, and complexity to serve as benchmark suites for method development. Finally, other fitness functions could be added to allow HIBACHI to be used for any number of simulation problems such as simulating data that resembles the patterns a real dataset. This opens the door to using HIBACHI for generating pseudo-simulated data resembling real data thus overcoming privacy and security concerns.

An important limitation of this approach is that it does not explicitly define an effect size for the simulated signal. Although not explicit, this could be overcome by making an effect size estimate part of the fitness function that is used to generate models through stochastic search. For example, you could evaluate models based on how close the subsequent machine learning results are to a target classification or regression error. Smaller errors would generate stronger signals and higher errors would generate weaker signals. An additional limitation is that the method is computationally

complex due the nature of the stochastic search algorithms. Fortunately, this approach can be easily parallelized for cloud and cluster computing. Despite these limitations, HIBACHI provides a vehicle for the flexible simulation of open data for research in the biological and biomedical sciences.

5. Acknowledgments

We would like to thank Mr. Peter Andrews and Dr. Jeff Kiralis for their work on the prototype HIBACHI method and software that led to the new methods presented in this paper. We would also like to thank the Genomics and Computational Biology (GCB) graduate group at the University of Pennsylvania for their generous support of Mr. Maksim Shestov.

References

1. B. Peng and C. I. Amos, *Bioinformatics*. **24(11)**, 1408 (2008)
2. B. Peng and M. Kimmel, *Bioinformatics*. **21(18)**, 3686 (2005)
3. S. Dudek, A. A. Motsinger, D. Velez, S. M. Williams and M. D. Ritchie, *Pacific Symposium Biocomputing*. 499 (2006)
4. T. L. Edwards, W. S. Bush, S. D. Turner, S. M. Dudek, E. S. Torstenson, M. Schmidt, E. Martin and M. D. Ritchie, *Lecture Notes in Computer Science*, **4973**, 24 (2008).
5. M. D. Ritchie and W. S. Bush, *Adv Genet*, **72**, 1 (2010)
6. R. D. Hernandez, *Bioinformatics*. **24(23)**, 2786 (2008)
7. R. K. Walters, C. Laurin and G. H. Lubke, *Twin Res Hum Genet*. **17(4)**, 272 (2014)
8. R. J. Urbanowicz, J Kiralis, N. A. Sinnott-armstrong, T. Heberling, J. M. Fisher and J. H. Moore, *BioData Min*. **5(1)**, 16 (2012)
9. J. H. Moore, R. Amos, J. Kiralis and P. C. Andrews, *Genet Epidemiol*. **39(1)**, 254 (2015)
10. D. Ashlock, Evolutionary Computation for Modeling and Optimization. New York, NY: Springer Science Business Media, Inc. Print (2006)
11. R. Poli, W. B. Langdon, N. F. McPhee and J. R. Koza, A field guide to genetic programming. S. l.: Lulu Press. Print (2008)

Identifying natural health product and dietary supplement information within adverse event reporting systems

Vivekanand Sharma and Indra Neil Sarkar
Center for Biomedical Informatics, Brown University
Providence, RI 02912, USA
Email: vivekanand_sharma@brown.edu and neil_sarkar@brown.edu

Data on safety and efficacy issues associated with natural health products and dietary supplements (NHP&S) remains largely cloistered within domain specific databases or embedded within general biomedical data sources. A major challenge in leveraging analytic approaches on such data is due to the inefficient ability to retrieve relevant data, which includes a general lack of interoperability among related sources. This study developed a thesaurus of NHP&S ingredient terms that can be used by existing biomedical natural language processing (NLP) tools for extracting information of interest. This process was evaluated relative to intervention name strings sampled from the United States Food and Drug Administration Adverse Event Reporting System (FAERS). A use case was used to demonstrate the potential to utilize FAERS for monitoring NHP&S adverse events. The results from this study provide insights on approaches for identifying additional knowledge from extant repositories of knowledge, and potentially as information that can be included into larger curation efforts.

Keywords: Natural Health Products; Dietary and Herbal Supplements; Adverse Event Detection; Terminology Mapping; Natural Language Processing.

1. Introduction

The biomedical community has benefitted from continuous development and improvement of automated methods for knowledge acquisition from heterogeneous data sources. A fundamental requirement for such tasks includes identification of entities of interest and their resolution to standard terminologies[1]. The process of converting unstructured free text fields from data into structured format creates opportunities to attain actionable knowledge by designing analytic enquiries. The heterogeneity of data from different sources poses challenges when seeking to perform comprehensive, multi-source analyses. Previous studies have demonstrated the utility of interlinked data from multiple sources to identify potential new knowledge[2–4]. The growing amounts of biomedical data from multiple sources suggest that an essential prerequisite for biomedical knowledge discovery will be the potential to leverage terminology resources for facilitating efficient indexing and subsequent retrieval. The biomedical domain is equipped with standard vocabularies from several sources that are used to facilitate access, retrieval and analysis of data from disparate data and knowledge sources. For example, the Unified Medical Language System (UMLS) Metathesaurus, maintained by the National Library of Medicine (NLM), is a repository of over one million biomedical concepts from more than 100 sources[5].

To support standardization and integration of available information about drugs and health related outcomes, the Observational Health Data Sciences and Informatics (OHDSI) workgroup was established with the goal of developing an open-source standardized knowledge base[6]. The

most significant utility of such a knowledge base is its ability to facilitate rigorous and accurate assessment of relationships between drugs and Health Outcomes of Interest (HOI). The Adverse Event Open Learning through Universal Standardization (AEOLUS) is a major product of the OHDSI workgroup, designed as a resource for the biomedical community[7]. AEOLUS consists of a standardized representation of FAERS[8] data, including normalization of drug names and health outcomes from the adverse event reports and precomputed common statistics. Use of FAERS data requires pre-processing, cleaning, and standardization, which presents a challenge for researchers intending to attain insights from adverse event reports. AEOLUS directly addresses this challenge, and reduces the requisite time and effort required to pursue research that utilize FAERS data.

Given the resources and initiatives in the biomedical and observational data realms to support a range of analyses, investigation of drug-HOI signals shows tremendous promise. However, the limited potential to investigate similar efficacy and safety issues related to dietary and herbal supplements (DHS) is generally due to lack of such resources. Although generally considered safe, there is evidence of DHS causing physical and economic harm[9]. An estimate of DHS-related adverse events suggest that they are associated with approximately 23,000 emergency department visits per year[10]. The incidents are higher for groups where the use of DHS is prevalent (e.g., among Navy and Marine Corps personnel, 22% of DHS users reported one or more adverse effects[11]). Such statistics underscore the need for systematic studies and the evaluation of available documentation in literature or reports associated with DHS use. The utility of existing biomedical vocabularies has been evaluated in the context of DHS, showing that UMLS generally, and, more specifically MeSH, SNOMED-CT, RxNorm, and NDF-RT include only 54%, 40%, 32%, 22%, and 14% of supplement concepts respectively[12]. Lack of robust acquisition of supplement documentation from electronic health records resulting from the gap between supplement and standard terminologies has also been highlighted[13]. Wang *et al.* found that only 14.67%, 19.65%, and 12.88% of ingredient terms from the Dietary Supplement Label Database (DSLD)[14] were mapped by UMLS, RxNorm, and NDF-RT, respectively[15]. The issue of less than 100% drug name mapping coverage in AEOLUS is noted by Banda *et al.* and attributed to those not found in RxNorm which include non-prescription products among other reasons resulting in unmapped records[7].

This study examined the potential and utility of creating a list of terms and concepts from ten sources that provide coverage for ingredients from Natural Health Products (NHP) and DHS. Using this as a resource, a custom thesaurus was built and used by the MetaMap NLP tool[16] to identify name strings (e.g., prescription drug, recreational substance, natural products and dietary supplements) found in FAERS. The system was specifically evaluated for its ability to recognize mentions of Natural Health Products and Supplements (NHP&S). The results from this study reveal challenges and opportunities in the development of an NHP&S terminology resource for automating acquisition of relevant information. The insights gained from this study may serve as motivation for improvement of the NHP&S thesaurus as well as its use for identifying and mapping relevant information in knowledge sources.

2. Materials and Methods

The goal of this study was to build a thesaurus of terms indicating NHP and DHS ingredients that could be integrated with extant biomedical Natural Language Processing (NLP) tools to facilitate acquisition of domain relevant information. Terms were identified from ten sources that included biomedical terminology sources as well as sources aimed at providing NHP&S information for healthcare providers and the general public. The terms were organized into concepts and a custom thesaurus that was subsequently used by the MetaMap[16] NLP tool to identify NHP&S concepts in FAERS. The utility of the NHP&S thesaurus was evaluated on randomly sampled intervention name strings. The NHP&S thesaurus was then used to process intervention name strings from FAERS to identify reports of adverse events associated with ingredients from NHP&S. A general overview of the approach is graphically depicted in Figure 1.

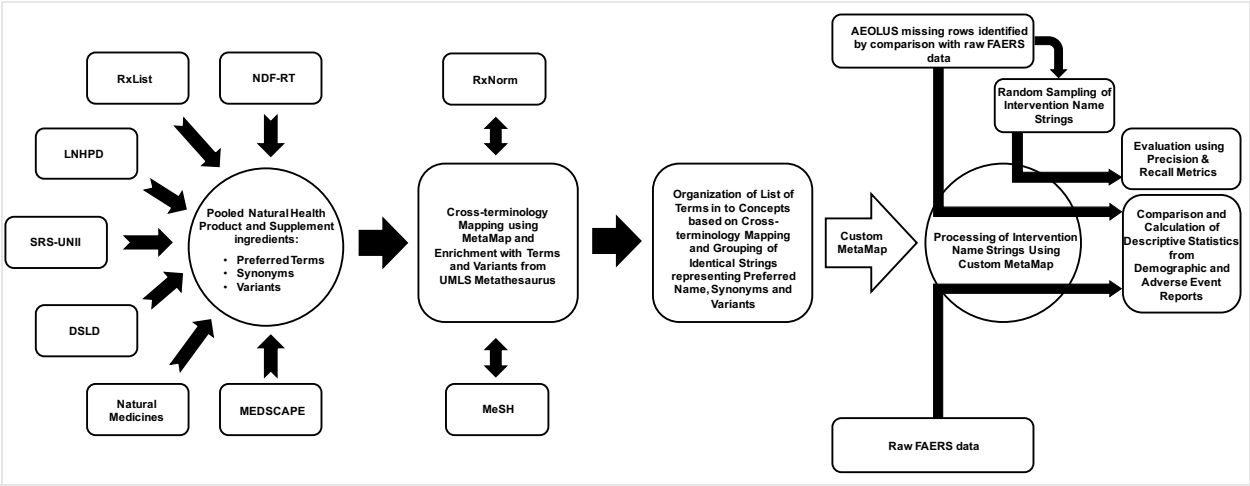


Figure 1. Overview of approach followed in this study.

2.1. *Identification of sources and extraction of NHP&S terms*

The goal of this step was to incorporate NHP&S terms from sources that offer reliable and comprehensive coverage of relevant terms, synonyms, and variants. The supplement ingredient terms and their synonyms were extracted from the databases shown in Table 1 (except RxNorm and MeSH, that was included after mapping). When available, the term variants were also extracted. Source identifiers were retained throughout the process; pseudo-identifiers were assigned for those terms that did not have an explicit identifier.

2.2. *Cross-terminology mapping of term strings from different sources*

The list of ingredient terms, synonyms, and variants from different sources were processed using the MetaMap NLP tool, which maps term strings to concepts included within UMLS Metathesaurus. From the MetaMap output (MMO), the mapped concepts, scores, semantic types, and terminology source(s) were extracted. The resulting list was filtered to retain those concepts that were identified with a perfect score of 1000. The filtering step included retaining concepts of semantic types associated with supplement ingredients[12]: (1) Plant (*plnt*); (2) Pharmacologic

Substance (*phsu*); (3) Organic Chemical (*orch*); (4) Food (*food*); (5) Biologically Active Substance (*bacs*); (6) Element, Ion, or Isotope (*elii*); and (7) Vitamin (*vita*). If the terminology source list included NDF-RT, RxNorm, or MeSH, it was recorded along with source identifiers.

Table 1. Sources selected for compiling NHP&S ingredient term list.

Source	Description
LNHPD[17]	*Licensed Natural Health Product Database*: Contains information about NHPs that have been issued a product licence after quality, safety and efficacy assessment by Health Canada. This database provides medicinal and non-medicinal ingredient information for a variety of NHPs.
DSLD[14]	*Dietary Supplement Label Database*: This database of full label information is a result of collaboration between the Office of Dietary Supplement (ODS) and the NLM to serve as an educational and research tool for students, healthcare providers, and the public.
SRS-UNII[18]	*Substance Registration System - Unique Ingredient Identifier*: This resource provides unique ingredient identifiers for substances in drugs, biologics, foods, and devices. From among the list of ingredients, those that had taxonomic links (other than viruses) were retained.
RxList[19]	*RxList* is a resource that offers pharmacological information on drugs and supplements. As a part of the WebMD network, the content is updated with recent articles and data from reliable sources such as pharmacists, physicians, FDA, etc. The "supplements" section was used to gather listed terms.
Natural Medicines[20]	*Natural Medicines*: This resource combines features and functionality from two of the major natural medicine databases, Natural Standard and Natural Medicines Comprehensive Database. The section "Food, Herbs & Supplement" was used to gather terms of interest.
Medscape[21]	*Medscape*: In addition to prescription drugs, this resource contains information related to herbals and supplements categorized by therapeutic classes are also provided. The section "Herbals and Supplements" was used to gather study relevant terms.
NDF-RT[22]	*National Drug File - Reference Terminology*: This resource is a formal representation of a drug list that include ingredients and provides hierarchical drug classification. The categories include "Herbs/Alternative Therapies" which was relevant for this study.
RxNorm[23]	*RxNorm* provides normalized names for drugs and its links to several other drug vocabularies used in pharmacy management. In addition to prescription drugs, it also includes food and dietary supplements among other types of interventions.
MeSH[24]	*Medical Subject Headings*: This is a controlled vocabulary maintained by the NLM for indexing biomedical artifacts (e.g., biomedical literature). It includes a range of terms including those used for drugs and herbs.
UMLS[5]	*Unified Medical Language System* maintained by the NLM provides a unified repository for over one million inter-related biomedical concepts from more than 100 sources. MetaMap was used to map the term strings from sources listed above except RxNorm and MeSH. For the identified UMLS concept list all synonymous terms and variants were extracted. Mappings to RxNorm, MeSH and NDF-RT were included in the final thesaurus to equip this resource with the ability to provide references to those sources when processing text.

2.3. *Grouping term strings and custom thesaurus*

Following MetaMap processing and identification of similar strings from across different sources, the list was enriched by extracting preferred terms, synonyms and variants from RxNorm, MeSH, as well as the UMLS Metathesaurus more generally (RxNorm: 4386 strings; MeSH:2826 strings; and Other sources: 5439 strings). Term strings indicating same ingredient across multiple sources

were grouped together. In addition to the entries mapped by MetaMap, the final dictionary included the unmapped terms which constitutes the major portion of the entries. Unique identifiers were assigned to indicate unique strings, variants, and concepts. This list was filtered against prescription drug list from National Drug Code Directory (NDC)[25] to remove any such ingredient name strings. This dataset was organized into tables for use with MetaMap Data File Builder suite (2016) to create a custom thesaurus that could be used with MetaMap ("Custom MetaMap").

2.4. *Sampling and evaluation*

The evaluation assessed the ability to recognize NHP&S strings as well as correctly eliminating non-NHP&S strings. A pool of unique strings was generated from missing rows of AEOLUS (mAEOLUS) by comparison with raw FAERS data files for years 2004-14. For example, ISR number 7811738 is missing in AEOLUS which contains NHP&S terms such as Red yeast rice, Fish oil, and Vitamin B6. Statistically significant random samples (95% Confidence Level at 4% Margin of Error) were selected and manually annotated as '*NHP&S*' or '*Non-NHP&S*'. To be more inclusive of NHP&S containing strings and to make evaluation more robust we performed the evaluation twice by creating two types of sample sets: (1) The strings were grouped separately according to year and random samples were picked without replacement; and (2) From the entire set of unique strings from the dataset, random samples were selected without replacement in ten iterations. The sampled strings were processed using Custom MetaMap and mappings with a score of 1000 (for MetaMap algorithm and scoring criteria refer to article by Aronson[26]) were retained to provide a more stringent evaluation and comparative performance in terms of our pipeline's ability to extract NHP&S records from FAERS when compared to those that were missed in AEOLUS as a result of inadequate coverage. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) was scored based on the ability to distinguish between NHP&S and non-NHP&S related terms and on whether a given intervention name string mapped to NHP&S term (correctly or incorrectly). The annotation was performed by an annotator whose expertise is in the area of natural health products and dietary supplements (VS) under the supervision of another subject-matter expert in biomedicine (INS). Evaluation was assessed based on the standard metrics of Precision and Recall. Year-wise evaluation was performed on the first set of samples and mean and standard deviation was calculated. For the second type of sample, Precision and Recall were calculated on the pooled data from ten iterations.

2.5. *Identification and summarization of Adverse Drug Event (ADE) reports*

Intervention name strings from AEOLUS missing rows (mAEOLUS) and raw FAERS data were processed separately using Custom MetaMap. The use of FAERS rows that were missing from AEOLUS was used to test the hypothesis that, due to inadequate coverage of existing biomedical terminologies, there is loss of information related to NHP&S. The NHP&S annotated strings were mapped back to ADE reports and a comparative examination was done. A basic comparison of NHP&S and non-NHP&S reports was performed by isolating relevant demographic and adverse event data. The stratified counts were normalized with the total counts for a given group (NHP&S or non-NHP&S) and used for comparison. The System Organ Class (SOCs) associated with

Preferred Terms for ADEs (which are encoded using MedDRA[27]) were identified and a comparative summary at the level of SOC was calculated using normalized counts.

3. Result

3.1. *Identification of sources and extraction of NHP&S terms*

Compilation of strings was initially from seven sources: LNHPD, DSLD, SRS-UNII, RxList, Natural Medicines, Medscape, and NDF-RT. Cross-terminology mapping resulted in enrichment of terms from RxNorm, MeSH, as well as UMLS. The final groupings resulted in 81,680 concepts encompassing 320,579 strings. The counts shown in Table 2 are based on the sum of entry terms, scientific names, synonyms, vernaculars as well as variants. The Source IDs indicate the preferred term used to list a given NHP&S within a given source; String IDs are additional synonyms or variants. The total of Source ID counts for individual source was more than that in the final thesaurus due to overlapping terms from different sources. The counts of overlapping terms among the sources in listed in Supplemental Table 2.

Table 2. Counts of ingredient name strings extracted from included sources.

Source		Source IDs	String IDs
LNHPD		6,108	9,359
DSLD		43,093	43,093
SRS-UNII		15,492	165,786
RxList		11,967	33,783
Natural Medicines		1,208	1,208
Medscape		193	1,248
UMLS	NDF-RT	4,179	11,273
	RxNorm	4,386	32,111
	MeSH	2,826	20,230
	Other	5,439	17,901

3.2. *Evaluation of Custom MetaMap*

The average number of distinct intervention name strings organized by year (2004-14) was 53,717.91±6796.87, ranging between 43,827 and 64,301. The sample size for random sampling from each year ranged between 592 and 595, with a mean of 593.45±0.89. The processing of sampled strings with Custom MetaMap resulted in mean precision and recall values of 0.94±0.01 and 0.72±0.08, respectively (F-score: 0.81±0.05). The total number of TP, FP, TN, and FN were 606, 39, 5640, and 241 respectively. Statistics for each individual year is provided in Supplemental Table 1. The second set was selected from 342,859 distinct intervention name strings from all years. The randomly selected sample size was 5990 gathered in ten iterations. Custom MetaMap processing of this sample resulted in a precision and recall of 0.93 and 0.66, respectively (F-score: 0.77). The total number of TP, FP, TN, and FN from this sample were 557, 40, 5102, and 291 respectively. A summary of the evaluation scores is listed in Table 3.

Table 3. Evaluation of Custom MetaMap on sampled intervention name strings from FAERS.

		Precision	Recall	F-score
Sampled year-wise	Range	0.92-0.97	0.57-0.85	0.71-0.90
	Mean	0.94±0.01	0.72±0.08	0.81±0.05
Sampled in 10 iterations then pooled		0.93	0.66	0.77

3.3. *Comparison of NHP&S mapped FAERS total and mAEOLUS*

Results from comparison of NHP&S mapped mAEOLUS with FAERS revealed that on average 39.11±11.37% more ADE records were retrieved using Custom MetaMap integrated with ingredient dictionary when compared to using OHDSI vocabulary alone. The numbers were comparatively lower for years 2013 (18.14%) and 2014 (13.27%). Figure 2 indicates the year-wise comparison of NHP&S associated ADE records retrieved from mAEOLUS and FAERS.

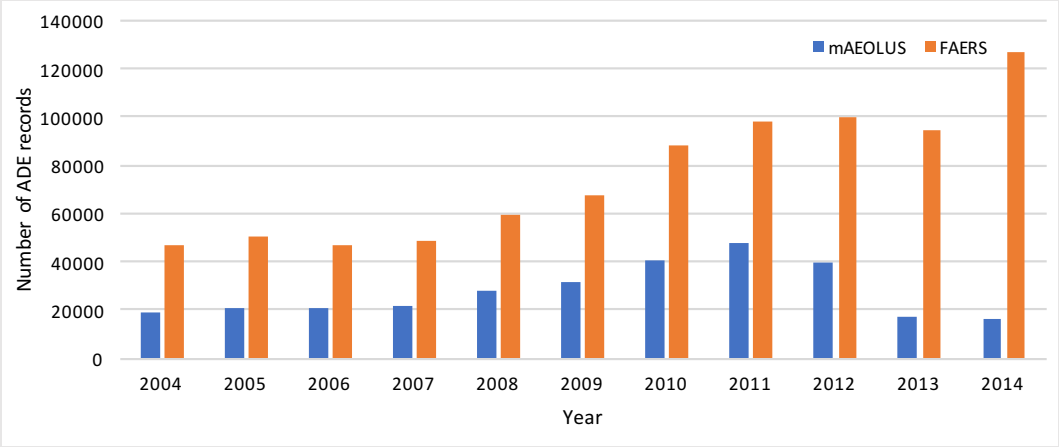


Figure 2. Comparison of ADE records identified from mAEOLUS and FAERS.

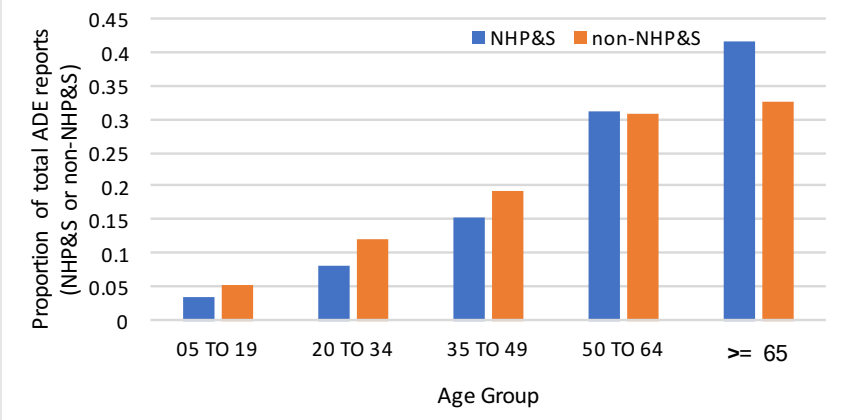


Figure 3. Comparison of NHP&S and non-NHP&S ADE report proportions stratified by age groups.

3.4. *Summary of NHP&S related reports in FAERS*

The NHP&S associated ADE records retrieved comprises of an average of 13.93±1.61% of total ADE records in FAERS every year from 2004 to 2014. The proportion of non-NHP&S ADE reports was higher in all age groups below 65. However, among the population with age group greater than or equal to 65, the proportion of NHP&S related reports was higher than non-NHP&S related reports (Figure 3). Figure 4 indicates the comparison of normalized values of NHP&S and non-NHP&S associated ADE report counts organized by SOCs. The top five SOC categories where the proportion of NHP&S is higher than non-NHP&S related ADE report counts were: (1) *Injury poisoning and procedural complications* (Inj&P); (2) *Gastrointestinal disorders* (Gastr); (3) *Infections and infestations* (Infec); (4) *Product issues* (Prod); and (5) *Metabolism and nutrition disorders* (Metab).

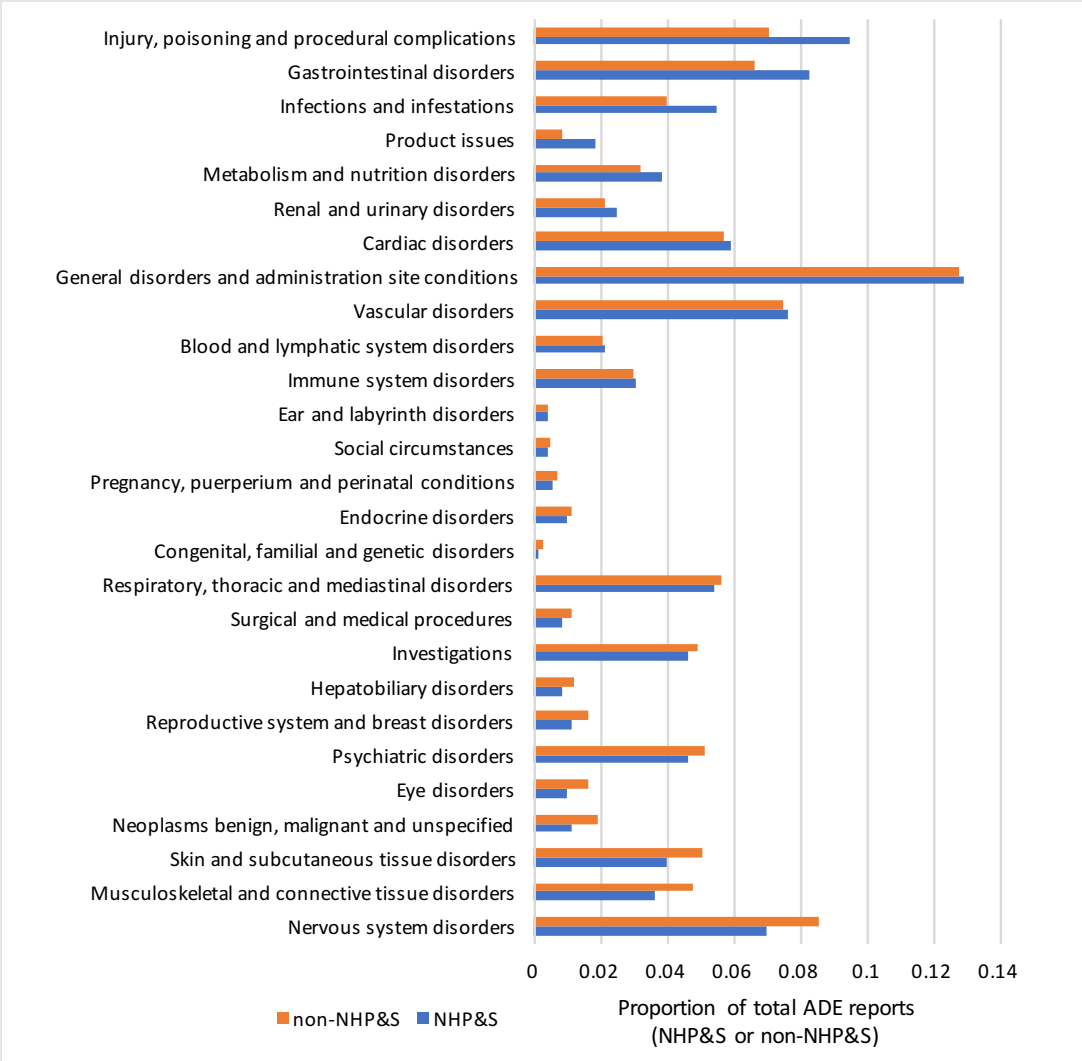


Figure 4. Comparison of proportions of NHP&S and non-NHP&S reports grouped by SOC.

4. Discussion

The biomedical domain is equipped with rich and multiple vocabulary sources and tools and techniques for concept recognition from unstructured data fields. Such resources play a key role in automation of cataloguing, indexing and retrieval of information of interest. Use of tools and techniques that provide scalable solution to analyze large amounts of data aid the discovery and generation of actionable outputs. Although there is significant amount of data publicly available in biomedical domain, the pipeline to attain insights from such data suffers from multiple hurdles. This in turn restricts the multi-disciplinary access and sharing of such data to those interested in translational research. The impediments in making the data accessible lies in data extraction, cleaning, standardization, and integration spanning multiple sources. Having these steps performed effectively may potentially facilitate design and execution of extensive data analysis plans. Community efforts such as the OHDSI focus on catering to such needs of researchers in biomedical community. However, the realm of NHP&S research lacks such resources which present a hurdle in pursuing data-driven investigations. This study explored the feasibility and utility of creating an NHP&S ingredient term thesaurus that could be leveraged by existing NLP tools for identifying relevant information embedded within biomedical knowledge sources.

The constructed NHP&S thesaurus for this study was a compilation of natural health products and dietary supplement ingredients from sources that have been either curated using evidence-based information (e.g., RxList, Medscape, Natural Medicines), reviewed by experts from FDA and the United States Pharmacopeia (e.g., SRS-UNII), issued a product license (e.g., LNHPD), or are/were available in the U.S. market (DSLD). A major challenge among the ingredient terminology sources is the lack of coverage of full set of synonyms, scientific names, and vernacular ("common") names[28]. Ambiguity of scientific names is another challenging aspect which requires close attention. Because many natural products are based on organism names, future work will include identifying natural product ingredient source organism names and gathering complete list of accepted scientific names, synonyms, and vernacular names. Similarly, for chemical dietary supplement ingredients, accepted IUPAC names, commonly used names, and abbreviations need to be included. In addition to the ingredient names, having commonly used product names in the thesaurus may improve the recall.

The results from evaluation suggests the need for development of NLP systems with enhanced mapping ability. The underreported nature adverse events related to DS, with only one in 100 being reported to FDA[29], accentuates the need for tools and approaches with higher sensitivity. Such tasks could benefit from the recent advancements being made in entity recognition from text sources, such as deep learning methods such as long-short term memory (LSTM)[30] or approaches combined with statistical word embeddings (LSTM-CRF)[31]. The comparison of results from mapping mAEOLUS with raw FAERS data with custom MetaMap shows recovery of additional ADE reports that were otherwise missed, potentially due to inadequate NHP&S ingredient term coverage within the current OHDSI vocabulary. The lower numbers for 2013 (18.14%) and 2014 (13.27%) could be due to less NHP&S records in missing AEOLUS rows or those already in the OHDSI vocabulary. Future work will be focused on grouping similar interventions associated with NHP&S. The incomplete grouping of entry terms within current thesaurus reflects higher number

of concepts. Efforts to expand this study will focus on manual curation to use relations to group terms into an ontological structure. Such step will result in fewer number of actual concepts/entries representing a compact NHP&S collection and will allow efficient categorization of their respective adverse events for calculation of signal disproportionality statistics. Such complete thesaurus of NHP&S would enable retrieval of relevant information from a variety of sources such as biomedical literature, clinical notes, online health forums, and social media. In addition to retrieval and dissemination of data in a standardized form, this effort will promote interoperability among traditionally disconnected data sources leading to generation of insights from more comprehensive analysis of data with limited risk of information loss.

NHP&S ADE reports may be important for analysis and detection of adverse event signals, both in terms of direct effects as well as interactions with pharmaceutical drugs. The proportion of adverse events related to NHP&S was higher in age group greater than or equal to 65 (senior citizens) compared non-NHP&S (Figure 3). This finding is consistent with the findings reported to the Special Committee on Aging, U.S. Senate[32]. Grouped by the top hierarchical structure of MedDRA, SOC, the proportion of injury and poisoning (*Inj&P*) related reports were higher in NHP&S group compared to non-NHP&S (Figure 4). The results presented here demonstrate the added potential for leveraging existing biomedical knowledge sources, such as FAERS, as a source for NHP&S knowledge.

This study highlights several challenges and opportunities in development of vocabulary resource and terminology mapping approaches for fostering advanced analytic investigations. A glimpse of utility of such resources in studying FAERS reports makes the case for investing the required time and effort in further enhancement to this infrastructure. Community wide efforts are required in this domain to make data accessible in standardized form in order to scale up to the methodological advances as exists in biomedical domain focused on drug-HOI associations.

5. Conclusion

This study developed a new NHP&S thesaurus for supporting the processing, identification, and standardization of relevant NHP&S data from existing digital resources. The application of the NHP&S thesaurus enabled a greater than 39% improvement in identifying NHP&S adverse events from the FAERS dataset. Such promising results suggest that there may be systematic approaches for identifying crucial NHP&S knowledge from existing biomedical data sources, and thus support overall curation efforts from complementary initiatives to develop community resources. Supplementary data are at: https://sites.google.com/a/brown.edu/phytokb/psb2018

6. Acknowledgement

This study was funded in part by NIH grants U54GM115677 and R01LM011963.

References

1. Kocbek S, Groza T. Building a dictionary of lexical variants for human phenotype descriptors. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics; 2016. p. 186–90.

2. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L. Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc. 2011 Jul;18(4):354–7.
3. Altman RB. Translational bioinformatics: linking the molecular world to the clinical world. Clin Pharmacol Ther. 2012 Jun;91(6):994–1000.
4. Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical language information and knowledge resources: GO and UMLS. Pac Symp Biocomput. 2003;439–50.
5. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267–70.
6. Boyce RD, Ryan PB, Norén GN, Schuemie MJ, Reich C, Duke J, et al. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. 2014 Aug;37(8):557–67.
7. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016;3:160026.
8. Evaluation CFD, Research. Drug Approvals and Databases - FDA Adverse Event Reporting System (FAERS). [cited 2017 Jul 18]; Available from: https://www.fda.gov/drugs/informationondrugs/ucm135151.htm
9. Office USGA. Health Products for Seniors: “Anti-Aging” Products Pose Potential for Physical and Economic Harm. 2001 Sep 10 [cited 2017 Jul 7];(GAO-01-1129). Available from: http://www.gao.gov/products/gao-01-1129
10. Geller AI, Shehab N, Weidle NJ, Lovegrove MC, Wolpert BJ, Timbo BB, et al. Emergency Department Visits for Adverse Events Related to Dietary Supplements. N Engl J Med. 2015 Oct 15;373(16):1531–40.
11. Knapik JJ, Trone DW, Austin KG, Steelman RA, Farina EK, Lieberman HR. Prevalence, Adverse Events, and Factors Associated with Dietary Supplement and Nutritional Supplement Use by US Navy and Marine Corps Personnel. J Acad Nutr Diet. 2016 Sep;116(9):1423–42.
12. Manohar N, Adam TJ, Pakhomov SV, Melton GB, Zhang R. Evaluation of Herbal and Dietary Supplement Resource Term Coverage. Stud Health Technol Inform. 2015;216:785–9.
13. Zhang R, Manohar N, Arsoniadis E, Wang Y, Adam TJ, Pakhomov SV, et al. Evaluating Term Coverage of Herbal and Dietary Supplements in Electronic Health Records. AMIA Annu Symp Proc. 2015 Nov 5;2015:1361–70.
14. Dietary Supplement Label Database (DSLD) [Internet]. [cited 2017 Jul 18]. Available from: https://dsld.nlm.nih.gov/dsld/index.jsp
15. Wang Y, Adam TJ, Zhang R. Term Coverage of Dietary Supplements Ingredients in Product Labels. AMIA Annu Symp Proc. 2016;2016:2053–61.
16. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May;17(3):229–36.
17. Licensed Natural Health Products Database (LNHPD) - Canada.ca [Internet]. Health Canada. 2007 [cited 2017 Jul 18]. Available from: https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/applications-submissions/product-licensing/licensed-natural-health-products-database.html

18. SRS-UNII. Substance Registration System - Unique Ingredient Identifier (UNII) [Internet]. Office of the Commissioner, Center for Biologics Evaluation and Research, Center for Veterinary Medicine, Center for Devices and Radiological Health, Center for Food Safety and Applied Nutrition, Center for Drug Evaluation and Research, National Center for Toxicological Research; [cited 2017 Jul 18]. Available from: https://www.fda.gov/forindustry/datastandards/substanceregistrationsystem-uniqueingredientidentifierunii/
19. RxList - The Internet Drug Index for prescription drugs, medications and pill identifier [Internet]. [cited 2017 Jul 18]. Available from: http://www.rxlist.com/script/main/hp.asp
20. Welcome to the Natural Medicines Research Collaboration [Internet]. [cited 2017 Jul 18]. Available from: https://naturalmedicines.therapeuticresearch.com/
21. Latest Medical News, Clinical Trials, Guidelines – Today on Medscape [Internet]. [cited 2017 Jul 18]. Available from: http://www.medscape.com/
22. Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, et al. Categorical information in pharmaceutical terminologies. AMIA Annu Symp Proc. 2006;116–20.
23. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof. 2005;7(5):17–23.
24. Medical Subject Headings - Home Page. 1999 Sep 1 [cited 2017 Jul 18]; Available from: https://www.nlm.nih.gov/mesh/
25. Evaluation CFD, Research. Drug Approvals and Databases - National Drug Code Directory. [cited 2017 Jul 19]; Available from: https://www.fda.gov/drugs/informationondrugs/ucm142438.htm
26. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17–21.
27. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999 Feb;20(2):109–17.
28. Sharma V, Sarkar IN. Leveraging biodiversity knowledge for potential phyto-therapeutic applications. J Am Med Inform Assoc. 2013 Jul;20(4):668–79.
29. US Government Accountability Report [GAO]. Dietary Supplements: FDA Should Take Further Actions to Improve Oversight and Consumer Understanding. Vol. GAO-09-250. 2009.
30. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak. 2017 Jul 5;17(Suppl 2):67.
31. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017 Jul 15;33(14):i37–48.
32. Office USGA. Health Products for Seniors: Potential Harm From “Anti-Aging” Products. 2001 Sep 10 [cited 2017 Jul 20];(GAO-01-1139T). Available from: http://www.gao.gov/products/GAO-01-1139T

Best practices and lessons learned from reuse of 4 patient-derived metabolomics datasets in Alzheimer's disease*†

Jessica D. Tenenbaum
Department of Biostatistics & Bioinformatics, Duke University, Box 2721
Durham, NC 27710, USA
Email: jessie.tenenbaum@duke.edu

Colette Blach
Duke Molecular Physiology Institute, Duke University, Box 104775
Durham, NC 27701, USA
Email: colette.blach@duke.edu

The importance of open data has been increasingly recognized in recent years. Although the sharing and reuse of clinical data for translational research lags behind best practices in biological science, a number of patient-derived datasets exist and have been published enabling translational research spanning multiple scales from molecular to organ level, and from patients to populations. In seeking to replicate metabolomic biomarker results in Alzheimer's disease our team identified three independent cohorts in which to compare findings. Accessing the datasets associated with these cohorts, understanding their content and provenance, and comparing variables between studies was a valuable exercise in exploring the principles of open data in practice. It also helped inform steps taken to make the original datasets available for use by other researchers. In this paper we describe best practices and lessons learned in attempting to identify, access, understand, and analyze these additional datasets to advance research reproducibility, as well as steps taken to facilitate sharing of our own data.

Keywords: FAIR; Open Data; Data Sharing; World Scientific Publishing.

1. Background & Introduction

The importance of data sharing and reuse is increasingly recognized across the biomedical research landscape. Also receiving increased attention are the challenges of adhering to best practices in data sharing. In many cases, researchers and even data managers are not properly incentivized to put in the up-front time and effort required to make data discoverable, comprehensible, and interoperable. Even when projects do plan ahead for data sharing by incorporating the required effort into a budget and hiring experienced informatics personnel, it is not always obvious how best to present data resources to facilitate discovery and uptake by others.

* This work is supported by 1RF1AG051550-01 and UL1TR001117

1.1. *The FAIR guiding principles*

Recognizing the urgent need to improve infrastructure for scholarly reuse of data, a group of stakeholders came together to develop what they referred to as "FAIR guiding principles", with FAIR as an acronym for Findable, Accessible, Interoperable, and Reusable.[1] These principles are meant to serve as guidelines and desiderata for good data stewardship. They are intended to enhance reusability of data, particularly from the machine perspective, enabling "computational agents" to identify, retrieve and analyze relevant datasets. A resource is 'F' (findable) if it has a globally unique and persistent identifier paired with rich metadata and is indexed in a searchable resource. 'A' (accessible) means that both data and metadata are retrievable using a standard, open protocol that allows for authentication as needed. The 'I' (interoperable) criteria relate to use of standards for knowledge representation. Finally, in order to be considered 'R' (reusable), a resource must have clearly defined and documented provenance and rules for usage.

The authors of the FAIR guiding principles make two important points that are relevant to the exercise described here: first, humans and machines face different challenges in the discovery and retrieval of relevant datasets. Humans have an intuitive sense of semantics and are able to interpret contextual clues such as icons, page structure, and narrative text. Machines lack these skills, but are far superior in scale and speed. In an ideal world, a resource enables discovery and re-use by both human and machine "stakeholders". Second, the FAIR authors assert that an optimal state in which computers are able to fully "understand" and operate on a digital object will likely rarely be achieved. Our intent in this work is not to fault any existing data resources, producers, or curators for in any way falling short of this theoretic optimal state. Rather, we seek to highlight ways in which existing datasets, all of which were made available before the FAIR guidelines were published, already adhere to these principles, and provide practical suggestions for how data producers going forward can make resources findable, accessible, interoperable, and reusable for both machines and humans.

1.2. *The Alzheimer's Disease Metabolomics Consortium*

The Alzheimer's Disease Metabolomics Consortium (ADMC- https://sites.duke.edu/adnimetab/) is a large, inter-institutional consortium that brings together centers of excellence of metabolomics, informatics and modeling to work collaboratively with Alzheimer's Disease experts to elucidate the molecular mechanisms of etiology and progression in AD. ADMC uses a systems approach in which metabolomics data are used to inform and complement genomics, proteomics, and neuroimaging data to provide novel insights about disease mechanisms.

The ADMC generated metabolomics data in collaboration with the Alzheimer's Disease Neuroimaging Initiative (ADNI) on the ADNI-1 cohort (see Section 2.1. below). These data were analyzed to identify peripheral metabolic changes in AD patients and correlate them with cerebrospinal fluid pathology markers, imaging features, and cognitive performance. Desiring to validate findings in independent cohorts, we identified other extent sample collections and/or datasets for which similar clinical and molecular data had been collected, or could be generated prospectively (Figure 1). In this paper we assess the degree to which these datasets already adhere to FAIR criteria and identify additional desiderata for best practices in data sharing, especially for

human users. Note that all of the datasets included here were discovered through distinctly human mechanisms: prior knowledge, networking, and past first-hand experience.

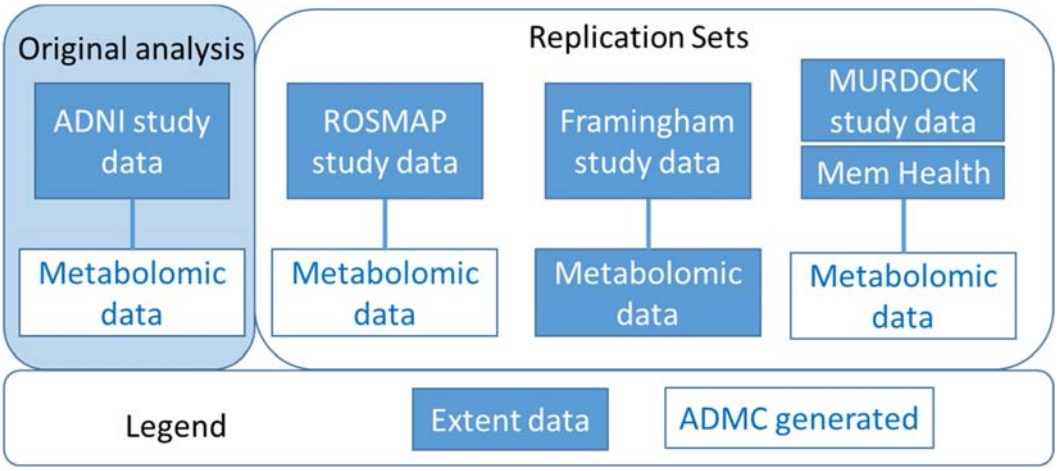


Figure 1: Metabolomic profiling was performed on the ADNI-1 cohort. The resulting metabolomic dataset was combined with clinical data collected on the ADNI-1 cohort, including AD-related markers and cognitive tests, to identify biomarkers in AD. Three additional cohorts were identified for which either metabolomic data had been collected (Framingham) or biospecimens were available (MURDOCK and ROSMAP). The ADMC performed metabolomic profiling on serum samples from ROSMAP and MURDOCK. Analysis of these datasets is ongoing.

2. Methods

2.1. *Datasets*

Three cohorts were identified for use in validation of original findings (Table 1). In both the original analysis of the ADNI-1 cohort[2] and replication in the additional datasets, analysis required metabolomic data, demographics, and clinical data, e.g. cognitive tests, changes in AD status, and APOE genotype.

The ADNI-1 cohort on which the original analysis was performed is part of the Alzheimer's Disease Neuroimaging Initiative and comprises 200 normal controls, 400 individuals with MCI, and 200 subjects with mild AD. Metabolomics data were generated on baseline serum samples using the AbsoluteIDQ®-p180 kit (Biocrates AG).[3]

The Framingham heart study was initiated in 1948 to identify risk factors for heart disease, beginning with 5200 adult men and women from the town of Framingham, MA. In 1971 a second-generation "offspring" cohort was enrolled, consisting of 5,100 of the original participants' adult children and their spouses.[4] The offspring cohort had their second examination 8 years after enrollment, and subsequent visits approximately every 4 years after that, including imaging, cognitive assays, etc. On their fifth visit, blood was drawn and used to perform metabolomic profiling using a liquid chromatography / mass spectrometry (LC/MS) platform.[5] They did not the Biocrates p180 platform, however there was overlap in the specific metabolites measured including a number of amino acids, lysophosphatidylcholines, and sphingomyelins.

Table 1. Overview of datasets included in evaluation.

Dataset	Full name	Study URL	Data URL
ADNI	Alzheimer's disease neuroimaging initiative	http://adni.loni.usc.edu/	http://adni.loni.usc.edu/data-samples/access-data/
Framingham	Framingham Heart Study	https://www.framinghamheartstudy.org/	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?id=phs000007
ROSMAP	Religious Orders Study and Memory and Aging Project	https://www.synapse.org/#!Synapse:syn3219045	https://www.radc.rush.edu/
MURDOCK	Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis	https://www.murdock-study.com/	https://www.murdock-study.com/services/data-dictionary/

The Religious Orders Study (ROS) and the Memory and Aging Project (MAP) are both longitudinal cohort studies of aging and Alzheimer's disease (AD) run from Rush University. ROS enrolled individuals from more than 40 groups of religious orders (nuns, priests, brothers) across the United States for longitudinal clinical analysis and brain donation.[6] MAP was designed to complement the ROS study by using a similar structure and design as ROS, but enrolling participants with a wider range of life experiences and socioeconomic status.[7] The entire ROSMAP cohort consists of approximately 3000 participants. The ADMC has performed mass-spectrometry-based metabolomic profiling on both serum and post-mortem brain samples for a subset of the ROSMAP cohort using the AbsoluteIDQ®-p180 kit from Biocrates Life Sciences.

Finally, the MURDOCK Study is not an open dataset but rather a community-based longitudinal registry and biorepository based in Kannapolis, NC and run by Duke University with more than 12,000 participants enrolled.[8] A number of prospective disease-specific "sub-studies" have been initiated from this registry, including a memory health study with approximately 800 participants. Blood and urine samples were collected at baseline enrollment along with demographic and clinical information. MURDOCK participants consent to give researchers access to their electronic health records for future study, and follow-up questionnaires are collected annually to ascertain longitudinal health status from the patient perspective. For the memory health study, participants were given assessments of cognitive status at enrollment and in a follow-up visit two

years later. Metabolomic profiling was performed on baseline serum samples using the AbsoluteIDQ®-p180 kit.

2.2. *Data governance*

ADNI has a relatively straightforward process for applying for access. One must agree to an online Data Use Agreement and fill out a form that includes one's institutional affiliation and a description of the proposed use of the data. Annual status updates are requested via email, and failure to provide them results in access being rescinded.

Access to the Framingham data involves a more complex process. The Framingham data are stored in dbGaP. In order to request access, the applicant must have an approved IRB protocol for data analysis from their home institution. An application is then required that describes the proposed use of the data as well as a data management plan to keep data secure. Notably, the principal investigator's signature is not sufficient. Rather, an institutional signing authority is required to be involved, as well as an IT Director who has institutional (not just departmental) authority, e.g. the Chief Information Officer or Director of IT Security. A major hurdle for our inter-institutional consortium was the requirement that each institution obtain the data directly from dbGaP rather than access the data through our secure file share. Statistical collaborators at other institutions were thus required to obtain their own respective IRB protocol approval and apply for access through dbGaP including a named signing authority and IT contact. Even using Duke's protocol as a basis, this slowed things down considerably.

For the ROSMAP and MURDOCK studies, each has a process in place for a would-be collaborator to fill out a proposal for use of data and/or samples. A signed DUA is required between the source institution and each collaborating institution, as well as a material transfer agreement (MTA) where applicable. For both studies, the collaborator must then identify which specific variables are needed. MURDOCK additionally requires a data sharing document that specifies the mechanism of the data exchange.

3. Results

3.1. *FAIR Assessment*

We attempted to assess each dataset's adherence to the FAIR guiding principles. Note that we did not rely solely on machine-readable data and metadata particularly for the 'F', 'A', and 'R' criteria, but took into account resource owners' efforts to make datasets findable, accessible, and reusable for humans as well. The overall scores are provided in Table 2, with descriptions provided below.

We assessed each resource on a scale from 1 to 5 with 1 signifying no adherence at all and 5 connoting perfect adherence to the principles. By definition, since we were able to re-use each dataset to some degree, none of them received a score of 1. Conversely, none of them received a perfect 5 in any of the four areas. A formal analysis enumerating each sub-criteria is beyond the scope of this review, but specific examples of how the different datasets demonstrated the guiding principles are described in the following sections, along with some areas for improvement.

Table 2. Scoring of compliance with FAIR principles for each dataset. Legend: 1- no adherence; 2- minimal evidence of adherence; 3- some adherence; 4- good adherence; 5- follows principles to the letter. The MURDOCK Study is not included here because it is not an open data set but rather a registry and biorepository for collaborative research.

Dataset	Findable	Accessible	Interoperable	Reusable
ADNI	3	3	2	4
Framingham	4	3	2	4
ROSMAP	4	2	2	4

3.1.1. *Alzheimer's Disease Neuroimaging Initiative ADNI-1 Cohort*

ADNI is indexed in the Neuroscience Information Framework (NIF) as a resource, though not as a dataset *per se* ('F'). Access to ADNI data generally requires log-in to the Laboratory of Neuro Imaging (LONI) Image and Data Archive (IDA)[9] and manual navigation through a web interface to identify the files of interest ('A'). Given that ADNI is a complex study in its second decade and involves a complicated protocol to collect clinical, genomic, demographic, imaging, and cognitive data on multiple sub-cohorts, the available data spans hundreds of files and thousands of variables. This can be challenging to navigate, particularly for researchers new to the study. ADNI mitigates these challenges through extensive documentation and data dictionaries ('R'). ADNI has data dictionaries for each data file and a single consolidated dictionary in .csv format that enables searching for terms and filtering by topic. ADNI also has a merged file containing the most important variables. A major strength of ADNI is that all data files are available not only as .csv but also as packages for R, SPSS, SAS, and Stata ('A', 'I'). LONI also has tools for visualization of the population by different parameters (Figure 2).

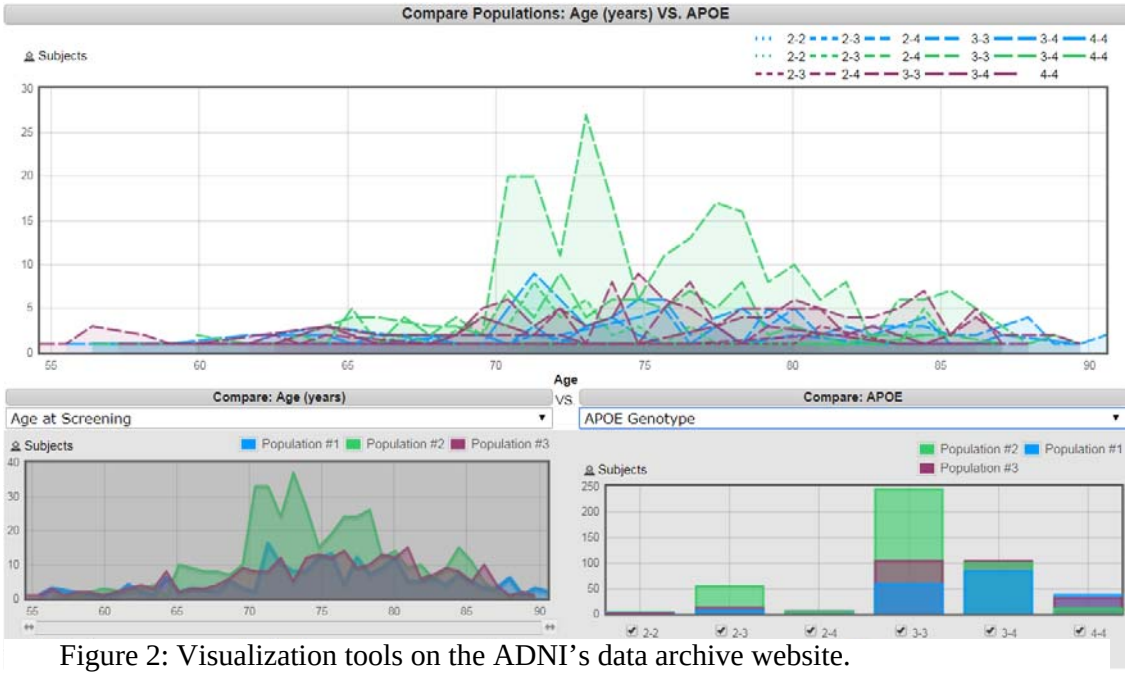


Figure 2: Visualization tools on the ADNI's data archive website.

The extensive existing clinical dataset for the ADNI-1 cohort was collected under the NIA-funded Alzheimer's Disease Neuroimaging Initiative. All ADNI-1 data must follow ADNI data governance rules, which include the stipulation that data be distributed only through the Laboratory of Neuro Imaging (LONI) Image and Data Archive (IDA).[9] Metabolomic data were generated through the NIA-funded Accelerating Medicines Partnership - Alzheimer's Disease (AMP-AD), which has a different set of data governance rules, including the requirement that all data be accessible through Sage Bionetworks' Synapse platform.

In order to comply with AMP-AD rules that require access to datasets through Sage Bionetworks' Synapse platform[10], a project was created in Synapse with digital object identifiers (DOIs) that point to the relevant permanent URLs in LONI's IDA ('F'). Importantly, these DOIs can be versioned as the underlying data files are updated, e.g. when additional clinical data are collected. Though two separate logins are required, one for Synapse and one for LONI, the hand-off is otherwise transparent and selecting "download" from the Synapse interface enables a user to download the appropriate file through the open http protocol directly after LONI authentication ('A').

3.1.2. *Framingham*

The study itself is indexed in both DataMed and NIF and has a permanent, versioned accession number in dbGaP ('F'). Data are downloadable from dbGaP through http once permission is obtained ('A'). Documentation is extensive, including annotated codebooks, procedures, variable statistics and publications ('R'). Again unsurprisingly for such a large, complex, and long-running study the documentation can be overwhelming, particularly for someone new to the study. The complexity is partially mitigated by search tools on the Framingham web site for variables. File names include some amount of metadata, with documentation to help the user understand shorthand naming conventions. However, individual data dictionaries for respective files contain headers that describe file content clearly and in detail. Another helpful resource for understanding Framingham data is a spreadsheet listing all of the different data files along with what cohorts they apply to and what types of data they contain ('R') (Figure 3). Finally, although dbGaP does not make explicit use of the Data Use Ontology (DUO), which has been adopted by the Global Alliance for Genomics and Health (GA4GH, or the Global Alliance) to code consent information, it does reference concepts that appear in the DUO such as not-for-profit use only ('F').

Nature of Data	Data Set Name	Cohort						Time Interval
		Gen 1 Original	Gen 2 Offspring	Gen 2 New Offspring	Gen 3 Generation 3	Omni I	Omni 2	
Year of Birth in 5 Year Intervals	birthyr_alls	●	●		●			Exam 1
Initial Neuropsych Battery Assoc with Brain MRI	neuropsych_2005s	●	●					thru 2005
SOE (Sequence of Events)	soe_2007s	●	●	●	●			thru 2007
Survival File & Follow-up CVD Events	srv_2007s	●	●					thru 2007

Nature of Data	Interview / Physical Exam / Examiner's Opinion																							
	Original data collectio			Validated (though medical records) and/or Reviewed and/or Derived																				
	Clinic Exam	Neuropsychology	Health Status Update	Menopause	Foot Study	MMSE, CERAD/Stroop	Rheumatic Heart	ICD Codes	Dementia	Atrial Fibrillation	Exam Dates, Age, Sex	Food Frequency with Der	Cancer	Diabetes	Cardiovascular	Survival	Endpoints:	Social Network	Stroke Related	Medications	Neuropsychology Brain MRI, S	Bone Related
Year of Birth in 5 Year Intervals											●											
Initial Neuropsych Battery Assoc with Brain MRI																					●	
SOE (Sequence of Events)																	●					
Survival File & Follow-up CVD Events																●						

Figure 3: Framingham spreadsheet of data files and related metadata.

3.1.3. *ROSMAP*

The Rush Alzheimer's Disease Center (RADC) has developed an elegant and user friendly "Research Resource Sharing Hub" designed to enable non-RADC investigators to navigate the complex set of data and biospecimens available for sharing (Figure 4) ('F'). This website provides extensive documentation, the ability to generate reports on numbers of research participants matching specific criteria broken down by demographics (Figure 6), and the ability to submit a request for data and/or biospecimens ('A'). Once our data request was approved, the Rush team extracted the required data and shared it via Dropbox.

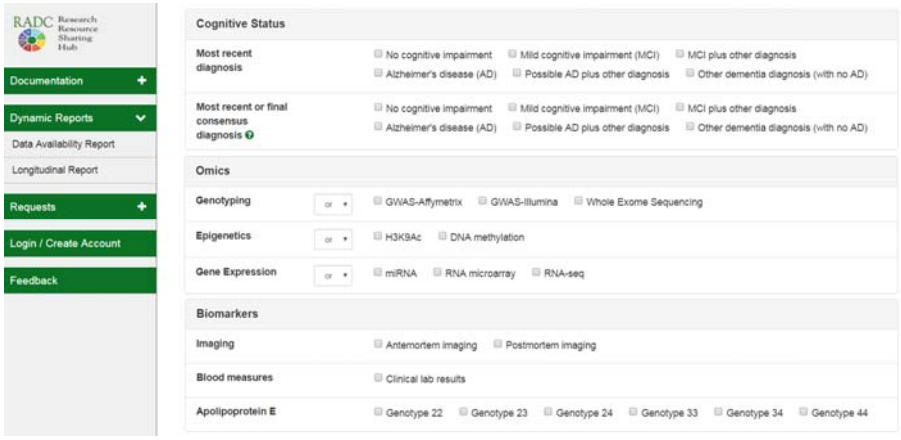


Figure 4: A screenshot (edited) of Rush's Research Resource Sharing Hub, enabling users to query for available data for research participants who meet specific criteria.

3.1.4. *MURDOCK*

The MURDOCK Study is not an open dataset but rather a registry and biorepository intended to facilitate cohort identification and collaborative sub-studies. Thus, in contrast with the datasets described above, the MURDOCK Study currently has only five forms and hundreds of data elements compared to the many thousands found in Framingham or ADNI. The main MURDOCK Study website provides a link to an online data dictionary documenting the different data

Self-Reported Diseases	Percent of Total Cohort
Coronary Artery Disease	7.3%
Cancer	23.6%
Diabetes	18.4%
High Cholesterol	44.5%
Osteoarthritis	23.9%
Depression	28.3%
Other Mental Illness	5.2%

Calculated BMI	Percent of Total Cohort
Underweight (<18.5)	1.1%
Normal (18.5-24.9)	26.9%
Overweight (25.0-29.9)	33.0%
Obese (>29.9)	37.1%
No BMI Recorded	1.9%

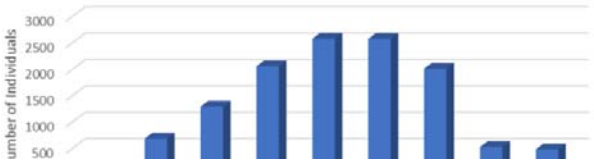
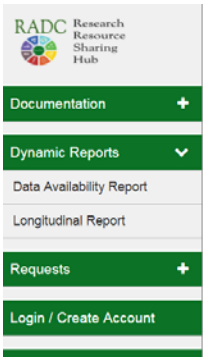


Figure 5: Self-reported clinical history, BMI, and age in the MURDOCK Registry found on the public facing MURDOCK Study website.



Ethnicity by sex	Female	Male	Total
Latino	136	41	177
Non-Latino	2263	846	3109
Unknown	0	0	0
Total	2399	887	3286

Race by sex	Female	Male	Total
White	2195	849	3044
Black/African-American	177	32	209
Asian/Pacific-Islander	14	4	18
Other	13	2	15
Total	2399	887	3286

Education	Total participants	Mean education (years)	Sample standard deviation (years)	Participants without college	Percent without college
All	3286	16.2	3.8	667	20.3%
Female	2399	15.9	3.6	511	21.3%
Male	887	16.9	4.2	156	17.6%
White	3044	16.3	3.7	600	19.7%
Black/African-American	209	15.3	3.8	55	26.3%
Asian/Pacific-Islander	18	14.1	4.6	7	38.9%
Other	15	13.8	6.3	5	33.3%

Figure 6: Tabular results of a query of a Rush Research Resource Sharing Hub query for frequency data of ROSMAP participants.

elements collected at the enrollment and follow-up stages of the study (‘R’). The website also gives a human readable overview of some demographic data and self-reported clinical history for a number of common diseases (‘R’) (Figure 5). It also provides information regarding the cognitive tests performed in the memory health study: attention and concentration, executive functions, memory, language, visual skills, conceptual thinking, calculations, and orientation. Once the data transfer is approved, the MURDOCK team extracts the specified field data and shares it using Box or ftp (‘A’).

3.2. *Common challenges across datasets*

3.2.1. *Metadata summarization and complexity*

Critical for every project was high level documentation to acquaint collaborators with study design and available data domains. Graphical overviews with linked details tend to be more informative and user-friendly than text-based summaries. In some cases, collection protocols were represented graphically; along with their corresponding naming conventions and file names. The best overviews included metrics for the data sets such as counts of different sample types, data types, etc. Metadata describing the processes, data files, fields and coded values were available from each of the projects and essential for data re-use. In almost all cases, metadata was largely human-readable and not computable or queryable (ROSMAP being the notable exception- see Figure 4).

Though none of the datasets described here were shared through metabolomics-specific repositories with computable metadata, progress has been made in establishing standards for metadata for metabolomic datasets. For example, EMBL-EBI (European Molecular Biology Laboratory- European Bioinformatics Institute)’s Metabolights data repository requires ISA-tab formatted metadata and provides a preconfigured downloadable ISACreator template. Use of ISA tools and the ISA standard does have some associated learning curve, but our team was able to make the ADNI1 p180 dataset ISA-compliant with significant help from knowledgeable curators for a recently accepted “data descriptor” (*Nat Sci Data, in press.*). According to a reviewer of this manuscript and documentation on GitHub (https://github.com/ISA-tools/isa-api), there exists a script, biocrates2isatab.py, that enables seamless conversion of Biocrates data to ISA-tab format,

however we were unable to locate the script itself- perhaps it is not yet publicly available. Certainly the use of such tools and standards will help to ensure FAIR datasets moving forward.

Other important sources of study metadata are data dictionaries for each domain. Data dictionaries can take on different levels of rigor and utility. Ideally a data dictionary is provided in a tabular format so that it is searchable for specific terms, browsable to get a feel for the different data domains and variables included, and filterable by topic. If the data files themselves do not use standard identifiers for variables, the data dictionary may facilitate mapping variables to existing standards, e.g. mapping local identifiers for metabolites to standard identifiers such as INCHI Key or ChEBI ID. Although some variables may seem obvious enough not to need descriptive text, contextual information is often helpful, e.g. TimeStamp might be described as "Time stamp for blood draw" rather than "Time stamp."

An additional local use case was the ability to query and filter based on status of metabolomics assays, e.g. which biospecimens had been assayed on a specific platform, and connecting that information to clinical and demographic data. A tool with i2b2-like graphical querying functionality would enable a PI or researcher to assess how many participants had both metabolomic and imaging data, and a diagnosis of AD.

3.2.2. *Data concept mapping across projects*

Notable progress by the Metabolomics Standards Initiative and the Coordination of Standards in Metabolomics (COSMOS) initiative.[11, 12] But as with many biological domains other than genomics, adoption of metabolomic data standards has been slow. Metabolomic data itself adds a layer of complexity in that some observations of molecular species may be ambiguous, for example lacking the ability to differentiate between two molecules with the same atomic composition but with double bonds between different carbon atoms. It is therefore not possible in some cases to assign a specific identifier to a given experimental value, since the value actually represents *species A* OR *species B*. Since the same Biocrates kit was used for three of the four datasets, mapping of metabolites for those three sets is trivial. Mapping and some manual review are needed to map the overlapping species between the p180 kit and the LCMS platform results for the Framingham study. For example, lysophosphatidylcholine (carbon:double bond = 16:0), is referred to as "C16_0_LPC" and "lysoPC a C16:0" in Framingham and the Biocrates kit respectively. Analysis has not yet been performed to determine consistency among the Biocrates datasets, nor comparability between Biocrates and the other LCMS platform, but this will be an important finding for future attempts to compare across metabolomic datasets.

In all observed cases, studies defined their own data elements rather than using existing concepts from existing terminologies such as SNOMED CT, LOINC, or PhenX. This resulted in some cases of significant semantic differences in variables of the same name, for example 'APOE' as a genotype vs. continuous variables representing RNA expression. Increased use of commonly accepted standards will increase interoperability of datasets moving forward.

Also related to interoperability, categorizing diagnoses was not consistent across studies and different protocols were used for consensus diagnosis. Although different assessments were used to evaluate cognitive impairment, they were each established, validated, standardized instruments. It

was therefore possible to establish equivalent concepts across projects with input from clinical experts.

In all cases, no matter how detailed the codebooks or project descriptions, there was always some need to ask for assistance from the data owners and to document this additional information for the analysts. This included, for example, additional information conveyed within a variable name e.g. single letter codes within variable names identifying brain regions.

3.2.3. *Versioning and data provenance*

Reproducible research requires the ability to track different versions of data as well as data provenance. Data sources can change for many different reasons, either because an error was discovered, or because additional data have become available. The Framingham study does a particularly good job versioning the data available in dbGaP, clearly identifying later releases of data for download after an embargo period, and dividing the data into two different groups based on participant consent. (One group consented to use for all research; the other consented for research use only by nonprofit entities.) For the ADNI-1 cohort, LONI has a policy that file names should not change so that researchers can always find the file they had previously downloaded. In addition, in order to adhere to DOI requirements for the AMP-AD project, LONI has enabled explicit versioning of data files within the IDA.

4. Conclusions

Based on our experience exploring publicly available datasets to validate translational findings we would add to the FAIR guiding principles the following best practices, particularly to enable data discovery and reuse by human beings: 1. Provide user-friendly metadata in the form of a graphical overview of data, sample types, instruments used at timepoints and counts; 2. Provide a data dictionary that is both browsable and searchable; and 3. Use common data elements wherever possible for data collection, whether from clinical terminologies or molecular databases.

It is easy for a group to become so familiar with their own data that they lose perspective on how it will be seen and interpreted by others. This exercise has helped inform our own work to make our data FAIR for other researchers, though we are not yet where we wish to be. Understanding of data sharing use cases, a well-formed plan, and dedicated resources are needed to enable adherence to FAIR principles.

It is encouraging to see that real effort is being devoted to making scholarly data available for re-use. A decade ago, it would have been difficult to obtain even a single dataset for validation. Our experience with the three cohorts described above suggests that although we have a long way to go before data are FAIR for computational agents, significant progress is being made to make data resources findable, accessible, and reusable by human agents. Our experience also suggest that, as with clinical data, we have a long way to go before data are truly interoperable. The obstacles are largely not technical ones. Education in the issues described here, as well as the will and the resources through aligned incentives, will ensure that we continue to make progress toward a FAIRer research data landscape.

5. Acknowledgments

The authors would like to thank our contacts at the four described studies for their support in understanding the rich and complex datasets: Michael Donohue, Karen Crawford, and Arthur Toga from ADNI; Lauren Silva, Honghuang Lin, Chunyu Liu, and Rhoda Au from the Framingham Study, Debra Fleischman, Gregory Klein, and John Gibbons from ROSMAP, and Brenda Plassman, Lawrence Whitley and Heather MacDonald from the MURDOCK Memory Health Study. We also thank the reviewers for their helpful comments.

6. References

1. Wilkinson, M.D., et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 2016. **3**.
2. Toledo, J.B., et al., Metabolic network failures in Alzheimer's disease-A biochemical road map. Alzheimers Dement, 2017.
3. St. John-Williams L, et al., Targeted metabolomics and medication classification data from participants in the ADNI1 cohort. . Nat Sci Data, 2017(Accepted).
4. Kannel, W.B., et al., An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol, 1979. **110**(3): p. 281-90.
5. Ho, J.E., et al., Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. PLoS One, 2016. **11**(2): p. e0148361.
6. Bennett, D.A., et al., *Overview and findings from the religious orders study.* Curr Alzheimer Res, 2012. **9**(6): p. 628-45.
7. Bennett, D.A., et al., *Overview and findings from the rush Memory and Aging Project.* Curr Alzheimer Res, 2012. **9**(6): p. 646-63.
8. Tenenbaum, J.D., et al., The MURDOCK Study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records. Am J Transl Res, 2012. **4**(3): p. 291-301.
9. Crawford, K.L., S.C. Neu, and A.W. Toga, *The image and data archive at the laboratory of neuro imaging.* Neuroimage, 2016. **124**: p. 1080-1083.
10. Omberg, L., et al., Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nature genetics, 2013. **45**(10): p. 1121-1126.
11. Salek, R.M., et al., COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics, 2015. **11**(6): p. 1587-1597.
12. Members, M.S.I.B., et al., *The metabolomics standards initiative.* Nat Biotechnol, 2007. **25**(8): p. 846-8.

Democratizing data science through data science training

John Darrell Van Horn[1], Lily Fierro[2], Jeana Kamdar[1], Jonathan Gordon[2], Crystal Stewart[1], Avnish Bhattrai[1], Sumiko Abe[1], Xiaoxiao Lei[1], Caroline O'Driscoll[1], Aakanchha Sinha[2], Priyambada Jain[2] , Gully Burns[2], Kristina Lerman[2], José Luis Ambite[2]

[1]*USC Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, 2025 Zonal Avenue, SHN, Los Angeles, CA 90033, Phone: 323-442-7246*

[2]*Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA*

jvanhorn@usc.edu, [jeana.kamdar, crystal.stewart, avnish.bhattrai, Sumiko.abe, xiaolei, caroline.odriscoll] @ini.usc.edu

[ambite, lfierro, jgordon, burns, lerman, priyambj]@isi.edu

The biomedical sciences have experienced an explosion of data which promises to overwhelm many current practitioners. Without easy access to data science training resources, biomedical researchers may find themselves unable to wrangle their own datasets. In 2014, to address the challenges posed such a data onslaught, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative. To this end, the BD2K Training Coordinating Center (TCC; bigdatau.org) was funded to facilitate both in-person and online learning, and open up the concepts of data science to the widest possible audience. Here, we describe the activities of the BD2K TCC and its focus on the construction of the Educational Resource Discovery Index (ERuDIte), which identifies, collects, describes, and organizes online data science materials from BD2K awardees, open online courses, and videos from scientific lectures and tutorials. ERuDIte now indexes over 9,500 resources. Given the richness of online training materials and the constant evolution of biomedical data science, computational methods applying information retrieval, natural language processing, and machine learning techniques are required - in effect, using data science to inform training in data science. In so doing, the TCC seeks to democratize novel insights and discoveries brought forth via large-scale data science training.

Keywords: Education, Metadata, Data Collection, Information Storage and Retrieval, Pattern Recognition, Automated, Classification

INTRODUCTION

Biomedical research has rapidly become a principal focal point for innovation and creativity in modern data management and analysis techniques - often spanning computational, statistical and mathematical disciplines being applied in the biological sciences to extract maximal utility from large-scale data (*1*, *2*). However, the rapidity with which data acquisition is occurring across biomedical research (*3*), often in lock-step with advances in technology, means that even what might have once been considered having a small-science focus finding themselves facing "big data" challenges (*4*, *5*).

To meet the essential data and computing demands of today's research biomedical ecosystem, the NIH has made an unprecedented investment in data science research and training through its Big Data to Knowledge (BD2K; https://datascience.nih.gov) program (*6*). Through a series of career development awards (K01 awards), institutional training awards

(T32/T15), a variety of data science training research awards (R25), and the training components of a dozen Centers of Excellence (U54), the NIH has placed a premium on the development of a new generation of biomedical data science professionals (*7*). These efforts systematically produce unique training materials seeking to introduce researchers to everything from the basics of databasing to data mining, machine learning, and the in-depth examinations of applied biomedical analytics in the investigation of health as well as disease. To catalyze and support all these efforts, the NIH established the BD2K Training Coordinating Center (TCC; http://www.bigdatau.org) with the stated aim to provide biomedical researchers with the tools to negotiate the complex landscape of data science education for biomedical researchers.

1.1 The BD2K Training Coordinating Center

Given the constantly developing analytical needs in biomedical science, varied training modalities are ideal for catering to the differing learning styles of busy researchers. Online training programs like massive open online courses (MOOCs) tend to focus on complete, end-to-end curricula in much the same manner as a traditional university course. Indeed, in some instances universities have adopted the MOOC model to develop entire degree programs. In contrast, scientific seminar and conference presentations tend to provide a narrower scope of content in highly specific domains. Hands-on training programs, done in-person or via the internet, on the other hand, often offer opportunities to directly learn the steps involved with some process, software tool, or analytical approach and then the chance to apply these concepts directly to an example data. While each of these models has their relative advantages and disadvantages, it is often the combination of these which provides the maximal utility for learning. That is, providing the foundational basis of data techniques applied to biomedical research challenges, the step-by-step understanding of computational processes, as well as the focused research rationale for why such measurements are being made. Spanning these levels of understanding help to reinforce a deeper appreciation for what data are telling one and what they represent about underlying biological systems. Consequently, the TCC provides multiple levels of training content, materials, and opportunities for online, as well as, in-person learning.

1.1.1 In-Person Training Activities and Workshops

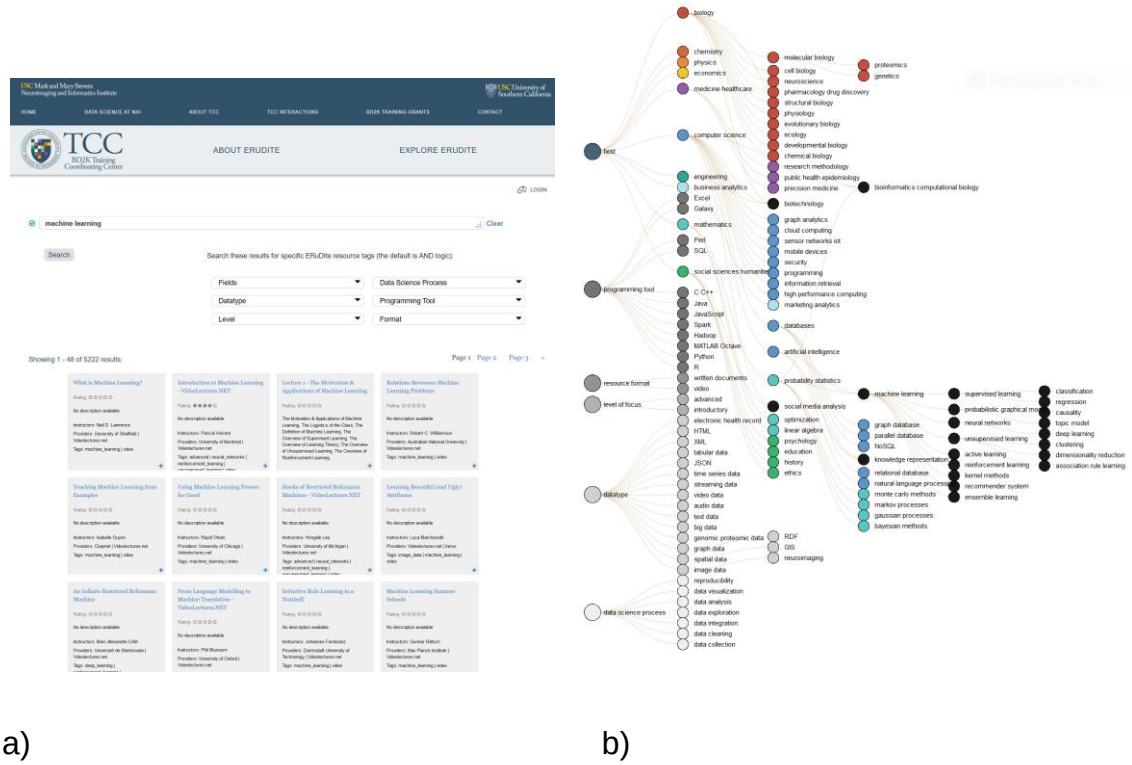
The TCC has created two specific programs for in-person training and learning which encourage mentorship and collaboration in biomedical data science. First, we developed the Data Science Rotations for Advancing Discovery (RoAD-Trip) program to foster new collaborations among junior biomedical researchers and senior-level data scientists to address the challenge of translating complex data into new knowledge (http://www.bigdatau.org/roadtrip). This program seeks to promote the careers of young biomedical scientists, engage established data scientists, and encourage the development of joint biomedical data science projects which are suitable as new NIH grant proposals.

Second, we have also organized an annual Data Science Innovation Lab, representing another example of fostering new interdisciplinary collaborations among quantitative and biomedical researchers to address data science challenges. This five-day residential workshop is supported by the NIH and the National Science Foundation. With the aid of professional facilitators or mentors, the accepted participants form teams that seek to solve specific data

science challenges. In 2016, the Data Science Innovation Lab discussed mobile health and the challenges arising from the use of wearable or ambient sensors. The most recent Data Science Innovation Lab, held in June 2017 with over 30 attendees, put its focus on understanding of the microbiome and the big data derived from microbiota (http://www.bigdatau.org/innovationlab2017).

1.1.2 Online Training Portal and Subject-Specific Training Search Tools

The TCC currently has two initiatives on creating educational video content. In an effort to expand on the general knowledge of Big Data, the TCC alongside the USC School of Cinematic Arts, created the short film: "Big Data: Biomedicine" which focuses on the science of big data and its implications for the future of biomedical research (https://youtu.be/F6CI7jXHGWg). Then, in collaboration with the BD2K Centers-Coordination



a) b)

Figure 1: a) TCC ERuDIte faceted search page. b) One representation of the ERuDIte "Knowledge Map".

Center (BD2KCCC) and the NIH Office of Data Science, the TCC has developed a weekly webinar series entitledThe BD2K Guide to the Fundamentals of Data Science Series (http://www.bigdatau.org/data-science-seminars). This series consists of lectures from experts across the country covering the basics of data management, representation, computation, statistical inference, data modeling, and other topics relevant to "big data" in biomedicine. These seminar videos are recorded and uploaded to YouTube, while we also include these videos (along with any other archived learning materials from BD2K centers) on the TCC website for discovery and adding to personal educational plans (see below).

1.1.3 The Educational Resource Discovery Index

A major effort of the TCC has been the creation of a sophisticated database of high-quality training materials available from different portals across internet called the Educational Resource Discovery Index (ERuDIte). Through ERuDIte, the TCC provides user-friendly access to a rich catalog of training resources and assembled learning materials. Specific selections of resources permit collections of educational topics tailored to a user's current and hoped-for knowledge and learning goals. In what follows in Sections 2 and 3, we present the details of our ERuDIte system and a brief summary of how we have applied data science to the organization of data science training materials in this unique platform.

2. The TCC Website and ERuDIte

The TCC's http://www.bigdatau.org website and ERuDIte have been specifically developed to provide an online platform for fostering and supporting self-directed learning in data science topics as they relate to biomedical research challenges. Users can search ERuDIte using faceted search over several dimensions (cf. Section 3.3), which describe different aspects of the learning resources, to identify those relevant to their training needs. The portal also provides summary visualizations of the ERuDIte catalog contents (http://www.bigdatau.org/statistics). In addition, learners can create individual profiles and then receive access to personalized learning features which allow them to monitor which training resources they have completed and to modify individual learning plans as needed.

Given the breadth and depth of data science training resources, the task of collecting and curating relevant, high-quality learning materials is not a trivial matter and requires a combination of manual and automatic approaches. Data science includes methods from and applications to multiple and diverse fields. As a result, researchers interested in learning about the techniques of data science can be faced with a range of MOOCs about databases, a practical tutorial on a blog which illustrates Matlab processing scripts, an online textbook describing the basics of Bayesian learning using the R programming language, or a video covering the latest advances in deep learning. Such resources provide value for learning, but each may have a different quality, time commitment, and relevance for specific training goals.

In the next section, we discuss how we are building ERuDIte using many of the same data science techniques we seek to teach. Specifically, we describe our approaches to 1) identifying high-quality learning resources using both manual and automatic techniques, 2) develop standard schemas and ontologies to describe the resources, 3) automatically assigning rich descriptors to each resource, and 4) provide access to these learning resources.

3. Building the Educational Resource Discovery Index (ERuDIte)

ERuDIte uses techniques from knowledge representation, data modeling, natural language processing, information retrieval, and machine learning to discover, integrate, describe, and organize resources, making ERuDIte a system that uses data science techniques to teach data science. As a result, the multiple components of ERuDIte follow core steps in the data science process: resource identification and extraction (Sect. 3.1), resource description and

integration (Sect. 3.2) and automatic modeling (Sect. 3.3).

Provider	Types	Total	With Descriptions	With Transcripts	With Additional Text + Slides
BD2K	*Video / Written*	515	461	264	59
edX	*Course / Video*	92	91	70	54
Coursera	*Course / Video*	77	77	54	56
Udacity	*Course / Video*	17	17	17	0
Videolectures.net	*Video*	8078	5741	165	4596
YouTube	*Video*	410	356	252	0
ELIXIR	*Course / Written*	235	48	0	0
Bioconductor	*Course / Written*	5	2	0	0
Cornell Virtual Workshop	*Course / Written*	38	19	0	0
NIH	*Video*	1	1	0	0
OHBM	*Video / Written*	78	6	0	51
TOTAL		**9,546**	**6,819**	**822**	**4,816**

Table 1: Summary of data currently collected in ERuDIte (as of Sept. 2017), by source and by resource type.

3.1 Resource Identification and Aggregation Methods

To start our collection process, we first reviewed MOOCs, blogs, e-books, videos, websites, conference presentations and tutorials, and other relevant data science material available on the web (http://www.bigdatau.org/about_erudite). In selecting resources, we consider the reliability of the source provider, the didactic value, and overall quality of the resources. From high-quality sources, such as MOOCs or conference tutorials, we extracted the metadata about the resources using an automatic scraping framework (Sect. 3.1.1). For resources of mixed quality, such as YouTube, we developed automated quality identification techniques (Sect. 3.1.2).

3.1.1 Automatic Website Scraping

Collecting relevant material is essential to the value and success of ERuDIte. As a result, during the initial stages of development, we focused our attention on gathering high

quality sources that contained collections of individual resources. This included MOOC sites such as edX (https://www.edx.org/), Coursera (https://www.coursera.org/), and Udacity (https://www.udacity.com/) in addition to the sites of other BD2K centers creating their own training materials (https://commonfund.nih.gov/bd2k).

A few unstructured sources with high-quality resources required completely manual attention, but overall, we focused our early identification efforts on structured sources that would allow us to gather resource data in a semi-automated way. Coursera and Udacity provided rich APIs, but the majority of the other resources required scraping. To streamline the scraping procedure, which demands individual source customization, we created a framework with website-specific modules using the popular Python packages BeautifulSoup and Dryscrape. The framework provides tools to handle dynamic JavaScript pages and to structure, extract, and export the collected data. The scraping framework is then packaged as a Docker image loaded with all the dependencies, and we store the image in a central repository. This allows for parallel development where multiple members of the team can extend the framework as needed without having to manage local software package installation and updates. Consequently, using this framework, we were able to identify and gather large collections quickly. As of September 27, 2017, ERuDIte contains over 9,500 resources. Table 1, above, provides details of the current collection of indexed resources.

3.1.2 Automated Quality Assessment

To expand our resource collection beyond our manually curated sources, we are developing techniques to identify high-quality learning resources from large open collections, such as YouTube. In this section, we describe how information extraction and machine learning techniques are applied to assess the quality of data science videos in YouTube and include these resources into ERuDIte.

Searching for the phrase “data science” on YouTube yields over 190,000 videos (and over 19 million if “data science” is not constrained to be a phrase). However, the amount of relevant and pedagogically valuable videos is a fraction of this number. To filter down the results, we trained a classifier to assess quality using video metadata, such as upload date, views, and “likes”, as well as extracted text, such as title, description, and automatic transcripts.

We use a set of concepts relevant to data science (specifically, from the Data Science Domain from the ontology described in Sect. 3.2.2) to search across YouTube. The search queries include concept names, sometimes with additional clarification terms, for example: “bioinformatics”, (“data science” AND “python”), or ((“data science” OR “machine learning”) AND “regression”). Sixty-two such queries were conducted and the resulting metadata was obtained from those videos and playlists appearing in the first 20 pages of results from YouTube for each query, which yields a dataset of 41,605 videos (35,235 unique). We then manually annotated 986 videos, sampled from across the different pages of results for different queries. These were judged on a scale of 0–4, where 0 is a video that is completely unhelpful as a resource for learning about data science, while 4 is most helpful. These are scored with resources labeled 0–1 considered low-quality and 2+ considered good-quality. This provided us with a roughly balanced data set of 417 low-quality videos and 569 high-quality videos. Finally, using k-fold cross-validation, a logistic regression classifier was trained using a variety of features, including the video text and metadata. The classifier achieves precision of 0.79,

recall of 0.85, and a F_1 score of 0.82. This performance is sufficient to select promising videos from YouTube for human curation. As training data size increases, we expect that the automatic classification quality will approach human levels of agreement and minimize human effort.

3.2 Resource Description Schema and ERuDIte Integration

With resources originating from different sites and creators, we have developed a metadata standard (Sect 3.2.1) to unify the structure of the ERuDIte index and to provide as much information as possible to learners to select relevant resources. We have also developed an ontology with 6 hierarchical dimensions to further describe the learning resources (Sect 3.2.2). These metadata are organized in the ERuDIte database (Sect 3.2.3), and openly shared with the community using the JSON-LD Linked Data standard (Sect 3.2.4).

3.2.1 ERuDIte's Learning Resource Metadata Standard

To design the metadata standard for learning resources in ERuDIte, we reviewed previous standards, including Dublin Core, Learning Resource Metadata Initiative (LRMI), IEEE's Learning Object Metadata (LOM), eXchanging Course Related Information (XCRI), and Metadata for Learning Opportunities (MLO), as well as emerging standards such as Bioschemas.org and the CreativeWork and Course schemas from the Schema.org vocabularies. The key classes of our standard are CreativeWork (used for learning resources), Person (for instructors or material creators), and Organization (for affiliations and resource providers).

We are collaborating with ELIXIR, a large European project organizing life science data, as well as other international organizations (e.g. Goblet, from England, H3Africa, from South Africa, and CSIRO, from Australia), to converge to a common standard for learning resources. In coordination with these groups, we have adopted schema.org vocabularies, defining additional properties when critically needed. Schema.org has the support of major search engines (such as Google, Bing) which facilitates discovery and dissemination of resources indexed in ERuDIte. A white paper on our joint efforts is due in the autumn of 2017.

3.2.2 The Data Science Education Ontology

To further describe the contents of learning resources, we have created the Data Science Education Ontology (DSEO), based on the Python machine learning package scikit-learn (*8*). The DSEO is specifically organized as a SKOS vocabulary since the flexible "broaderTransitive" property from SKOS best captures the subtle relationships between our concepts. The DSEO is publicly available at http://bioportal.bioontology.org/ontologies/DSEO. Specifically, the DSEO has six hierarchical dimensions, each describing a different facet of a learning resource:

- *Data Science Process* (8 concepts): What stages of the data science process will this resource help me with understanding?
- *Domain* (74 concepts): What field of study does this resource focus on?
- *Datatype* (18 concepts): What types of biomedical data are addressed in the resource?
- *Programming Tool* (13 concepts): What programming tools are being used in or taught by this resource?

- *Resource Format* (2 concepts): In what manner is this resource presented?
- *Resource Depth* (2 concepts): How advanced is this resource? At what experience level is it pitched?

3.2.3 Resource Database

To store, integrate, and efficiently query ERuDIte's resource data and metadata we have adopted the use of a relational database for storing the direct output of our scrapers. We then define a set of database views to integrate data across sources and map source tables to a relational implementation of our metadata standard. This enables us to flexibly extend the metadata schema without modifying collected data. For efficiency, we create a materialized view with appropriate keys and indices which joins standard schema views to form a composite table that powers the resource detail pages on the BD2K TCC website. Additionally, we also generate an Elasticsearch (elastic.co) index over the metadata to power the faceted search of ERuDIte.

3.2.4 Linked Data Representations

To disseminate the resources indexed in ERuDIte as broadly as possible, we embed structured metadata for each resource expressed in JSON-LD (https://json-ld.org/), in addition to our standard schema, on each learning resource webpage. Sharing resource metadata as openly as possible through embedded, concise JSON-LD has several benefits: 1) it complies with the goals of the Semantic Web and Linked Data communities to make the data available on the web. This is, information is not only human readable, but also readable by machines, and to allow for additional content about web-objects to be accrued in a distributed fashion (*9*); 2) web-based search engines, such as Google, are encouraging the use of the JSON-LD structured data format for webpages, since it aids in their own indexing and the representations of webpage content. This enhanced indexing facilitates discovery of resources by users outside of ERuDIte. We have used our previous work on schema mapping (*10*) to conveniently translate our relational schema of the resources metadata into JSON-LD format and have successfully applied this approach here, as well. What is more, in order to maximize the reuse of ERuDIte, we have licensed bigdatau.org website content and ERuDIte schema under a Creative Commons Attribution-Non-Commercial-ShareAlike (CC BY-NC-SA) license (https://bigdatau.ini.usc.edu/about_erudite).

3.3 Automated Concept Tagging for ERuDIte

Concept modeling has formed an integral element in the construction of ERuDIte. However, *manually* tagging the thousands of resources in ERuDIte with concepts from video and other content would be time- and cost-prohibitive. Thus, we have developed *automated labeling methods*, based on machine learning, natural language processing and information retrieval techniques, to efficiently tag the growing collection of ERuDIte learning resources.

In order to evaluate this concept modeling framework, we created a "gold standard" consisting of 726 manually-curated resources (data science courses from Coursera, Udacity, edX, Cornell's Virtual Workshop, and videos from Videolectures.net and YouTube) labeled

with the appropriate tags from each DSEO dimension. We randomly select 581 resources (~80%) for training and cross validation, and left aside 145 resources for testing. In previous experiments (*11*) we predicted all concepts across all dimensions; however, upon investigation, classifier performance increased per concept tag if trained on a per dimension basis. We then created fixed fold assignments for the training set of resources, and conducted five-fold cross-validation grid searches over the hyper-parameters defined for each classifier. The hyper-parameter grid included value ranges for parameters specific to the vectorization of each resource and to the classifier method itself. Averaged F_1 scores weighted by the support available in each fold were employed to select the best classifier with the best hyper-parameter combination. Predicted tags for the 145 resources in the test set were then obtained.

Dimension	**Classifier Type**	**F_1**	**F_1 (support >= 5)**	**F_1 (support >= 10)**	**F_1 (support >= 15)**
Domain	*Logistic Regression*	0.762	0.778	0.793	**0.807**
Resource Depth	*SVM*	**0.778**	**0.778**	**0.778**	**0.778**
Resource Format	*SVM*	**0.989**	**0.989**	**0.989**	**0.989**
Data Science Process	*Logistic Regression*	**0.705**	**0.704**	**0.704**	**0.704**
Programming Tool	*Logistic Regression*	0.533	0.538	0.537	**0.555**
Datatype	*Logistic Regression*	0.481	0.488	**0.492**	**0.492**

Table 2: F_1 classification scores on ERuDIte test set of resources, by classifier method, for the overall number of tags as well as those with at least a given level of support (from at least 5-to-15 training resources being present). A manuscript providing additional mathematical details on the validation of the ERuDIte autotagging and curation system is in preparation.

Table 2 briefly summarizes the performance of the best classifiers along each dimension. Some dimensions are clearly more difficult to resolve than are others. Particularly, our routine struggles with classifying the type of programming tools and datatypes dimensions. We believe that this can be explained by simply having fewer resources are tagged with concepts for these two dimensions. With greater numbers of exemplars along these axes, the better our classification will become. This suspicion has been born out and will be the topic of a subsequent research article from our team. Briefly, but not surprisingly, we observed that classifier performance is markedly improved on tags where there are at least fifteen resources labeled with that tag in the training set. A more comprehensive and detailed article on the validation of ERuDIte and the automated resource tagging approach is in preparation.

3.4 Further Work

3.4.1 Community Validation and Ongoing System Re-Training

The performance of our currently classifiers for ERuDIte resource identification (Sect.

3.1.2) and labeling (Sect. 3.3) has shown promising results (F1~0.8), but are not yet sufficiently accurate for automatically including their results directly into ERuDIte. We plan to leverage our automated classifier system to propose resources and tags to domain expert curators. This community-driven approach will aid in reviewing our classifier predictions thereby ensuring the inclusion of high quality resources and tagging. We have also developed a web-based curation interface for use in accelerating the ERuDIte resource curation process. In addition to validating the predicted tags, reviewers can also propose concepts which do not currently exist in our DSEO vocabulary. As additional curated resources and tags are included, we will systematically retrain our classifiers and expect their performance to improve.

3.4.2 Discovery of Training Pre-requisites

To enable personalized learning plans, automatically inferring which data science concepts are presented in each resource and what other concepts are prerequisites for these is an important step. For example, if a learner is interested in a course on *Machine Learning in Matlab* but her user profile does not indicate experience in mathematics, ERuDIte might recommend with a resource on *Probability*. Indeed, there are a number of methods to predict the underlying concepts present in a set of training resources, with topic modeling approaches such as Latent Dirichlet Allocation (LDA) being a commonly employed approach (*12*). Another approach worth exploring is one that exploits naturally occurring sequential data. Given such a set of sequential data, where each entry is associated with a distribution of concepts, weights are accrued for how likely a concept is to occur *in advance of* another

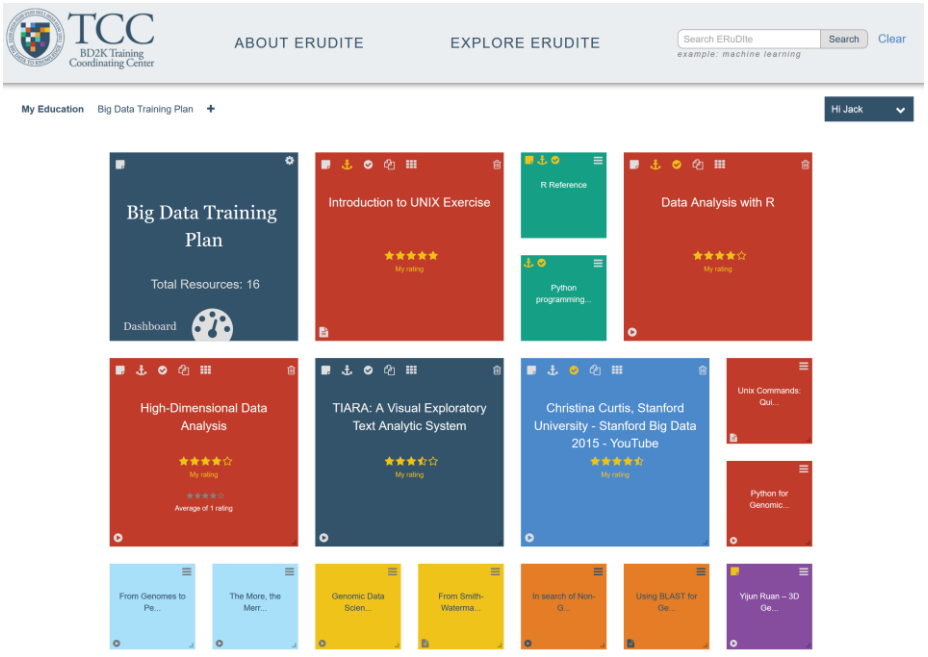


Figure 2: An example of an ERuDIte "training plan" comprised of resources indexed in the ERuDIte database. Users select, gather, arrange, color-code, rate, and then can utilize the educational resources in a prescribed, pre-requisite order or in any order they wish.

concept, and a net effect score can be computed by subtracting the weight for the converse of directionality, e.g. a concept occurring *after* another concept. For instance, the scraping of

textbook tables of contents and course syllabi from the Web is one such means of thematic concepts which occur in a regularized order. For example, given the chapter title "Unsupervised Learning: Clustering and Dimensionality Reduction", an informed algorithm would select the top five (albeit imperfect) results but with diminishing weights: derived via dimensionality reduction, k-means clustering, fuzzy clustering, machine learning, or sparse dictionary learning, to comprise the semantic relevance vector for the topic set (*13*). By so doing, we envision an evidence-based means of curriculum development.

3.4.3 Personalization of Training Resources

The TCC website is also designed to collect usage data which will allow us to develop personalized and custom learning experiences. First, registered learners can create their own profiles and detail their prior knowledge and their interests, which we can then use to align with our resource concepts to make recommendations and to present search results. Second, registered learners can create educational plans that include resources that they want to review and complete. These can be used as sources of dependency relationships between resources, and also to prevent suggesting resources already known to the user. Third, we have implemented user monitoring, which allows us to understand the sequence of resource browsing activity in order to drive recommendations as users explore ERuDIte. Presently, ERuDIte offers resource-specific recommendations based upon a semantic similarity search on a single resource's title. However, as TCC website and ERuDIte use grows, personalized recommendations via methods such as collaborative filtering are fully expected.

4. CONCLUSION

The BD2K TCC seeks to promote and support biomedical data science learning with a multi-pronged approach. We routinely organize in-person training events to engage researchers in data science learning through applied projects in biomedicine. These are used to actively create online learning materials with a potential to guide learners through data science concepts used in current biomedical research. Moreover, we use data science techniques to collect and organize learning resources widely available on the internet in order to help self-directed learners easily maneuver through the data science landscape - creating an online space where the knowledge and skills being taught are those being used in the learning ecosystem itself.

As the field of biomedicine increasingly demands multi-disciplinary skills and data science knowledge, resources for easily available constant learning are sorely needed. Data science is now providing support to a broad range of translational scientists (*14*) who require skills in data management, analytics, and visualization in order to better understand the richness and nuances of the data they are collecting. We believe that the work of the TCC and its emphasis on ERuDIte will contain the materials that not only establish the fundamentals of data science but also track the interest levels associated with new methods and techniques (*15*). Finally, even though the TCC is primarily focused on biomedical data science, we expect that the materials and approaches we provide will help the scientific community at large – helping to democratize training in data science to the widest possible audience.

ACKNOWLEDGEMENTS

This work is supported by an NIH Big Data to Knowledge (BD2K) award to the BD2K Training Coordinating Center (U24 ES026465; http://www.bigdatau.org). Further information on the NIH BD2K program, its contributing NIH institutes, and all of its supported projects may be found via the NIH Data Science Office website (https://datascience.nih.gov/). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Institutes of Health.The authors would also like to express their gratitude to the staff of the USC Mark and Mary Stevens Neuroimaging and Informatics Institute at the USC Keck School of Medicine.The authors declare no competing financial interests.

REFERENCES

1. R. Margolis *et al.*, The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* **21**, 957-958 (2014).
2. N. R. Adam, R. Wieder, D. Ghosh, Data science, learning, and applications to biomedical and health sciences. *Ann N Y Acad Sci* **1387**, 5-11 (2017).
3. J. D. Van Horn, A. W. Toga, Human neuroimaging as a "Big Data" science. *Brain Imaging Behav* **8**, 323-331 (2014).
4. G. J. Rinkus, Sparsey: event recognition via deep hierarchical sparse distributed codes. *Front Comput Neurosci* **8**, 160 (2014).
5. T. Althoff *et al.*, Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336-339 (2017).
6. P. E. Bourne *et al.*, The NIH Big Data to Knowledge (BD2K) initiative. *J Am Med Inform Assoc* **22**, 1114 (2015).
7. L. X. Garmire *et al.*, THE TRAINING OF NEXT GENERATION DATA SCIENTISTS IN BIOMEDICINE. *Pac Symp Biocomput* **22**, 640-645 (2016).
8. F. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825−2830 (2011).
9. M. Taheriyan, C. A. Knoblock, P. Szekely, J. L. Ambite, in *Proceedings of the 2012 ESWC Confernce on Linked APIs for the Semantic Web Workshop (LAPIS)*. (2012).
10. J. L. Ambite *et al.*, paper presented at the Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 2017.
11. J. L. Ambite *et al.*, in *Procs. 26th Intl Conf on World Wide Web*. (ACM, New York, NY, USA, 2017), vol. WWW '17 Companion.
12. E. Gabrilovich, S. Markovitch, in *Proceedings of International Joint Conference on Artificial Intelligence*. (Hyderabad, India, 2007).
13. J. Gordon, L. Zhu, A. Galstyan, P. Natarajan, G. A. P. C. Burns, in *Proceedings of the Association for Computational Linguistics (ACL)* (2016).
14. R. D. Jackson, S. Gabriel, A. Pariser, P. Feig, Training the Translational Scientist. *Science Translational Medicine* **2**, 63mr62 (2010).
15. M. C. Dunn, P. E. Bourne, Building the biomedical data science workforce. *PLOS Biology* **15**, e2003082 (2017).

IMAGING GENOMICS

HENG HUANG

Department of Electrical and Computer Engineering, University of Pittsburgh
Benedum Hall 3700 O'Hara Street, Pittsburgh, PA 15261
E-mail: heng.huang@pitt.edu

LI SHEN

Center for Neuroimaging, Department of Radiology and Imaging Sciences
Center for Computational Biology and Bioinformatics
Indiana University School of Medicine
355 West 16th Street Suite 4100, Indianapolis, IN 46202
E-mail: shenli@iu.edu

PAUL M. THOMPSON

Imaging Genetics Center, Stevens Institute for Neuroimaging & Informatics
4676 Admiralty Way, Marina del Rey, CA 90232
E-mail: pthomp@usc.edu

KUN HUANG

Department of Medicine, Indiana University School of Medicine
340 W 10th St #6200, Indianapolis, IN 46202
E-mail: kunhuang@iu.edu

JUNZHOU HUANG

Department of Computer Science and Engineering, University of Texas at Arlington
500 UTA Blvd., Arlington, Texas, 76019
E-mail: jzhuang@exchange.uta.edu

LIN YANG

Department of Biomedical Engineering, University of Florida
1275 Center Drive, Biomedical Sciences Building JG-56, Gainesville, FL, 32611
E-mail: lin.yang@bme.ufl.edu

1. Introduction

As an emerging research field, the goal of Imaging Genomics is to study the integrative high-throughput imaging (such as histopathological images in cancer research, MRI and PET images in brain study) and omics (such as SNP, DNA sequence, RNA expression, methylation, epigenetic markers, proteomics, and metabolomics) data, which created new opportunities for exploring relationships between genotypes, phenotypes, and clinical outcomes using quantitative methods. Imaging Genomics research holds great promise for precision medicine to better understand diseases, from genetic and genomic determinants to the complex interplay of phenotypic traits. The unprecedented scale and complexity of the Imaging Genomics data have presented critical computational bottlenecks requiring new biomedical data science tools. The

technological advance in this field is urgently needed and has the potential to significantly contribute to multiple national health priority areas such as the BRAIN Initiative,[1] the Precision Medicine Initiative,[2] and the BIGDATA Initiative.[3]

The objective of this Imaging Genomics Session at PSB 2018 is to encourage discussion on fundamental concepts, novel methods and innovative applications. We hope that this session will become a forum for researchers to exchange ideas, data, and software, in order to speed up the development of innovative technologies for hypothesis testing and data-driven discovery in Imaging Genomics.

2. Overview of Contributions

Our session includes seven accepted papers covering a variety of the subjects in the imaging genomics field. The papers address the imageing genomics research questions from genotype-phenotype association studies of complex brain disorder and cancerous diseases to survival analysis and disease characterization. The computational methods range from the convolutional neural network methods to the low-rank based multi-task learning model. The large-scale imaging and omics data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, The Cancer Genome Atlas (TCGA), the Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) Consortium, and The Cancer Imaging Archive (TCIA) have been analyzed in the accepted papers.

Miller *et al.* analyzed the rare variants from whole-genome sequencing from the ADNI cohort and identified several genes as significantly associated with the imaging phenotype using only synonymous variants that affected codon frequency. Their study showed that the codon bias may play a role in Alzheimer's Disease and that it can be used to improve detection power in rare variant association analysis. Huang *et al.* developed a deep survival learning model to predict patients' survival outcomes by integrating the multi-dimensional TCGA data. In order to take the advantage of both histopathological image information and molecular profiles from imaging-omics data, an integrative pipeline based on deep learning model was created. The unsupervised training and supervised fine tuning processes are combined to conduct survival prediction using a limited number of patient samples. Chidester *et al.* introduced a Discriminative Bag-of-Nuclear-Words (DBoNW) method to predict genomic markers using imaging features, which addresses the challenge of summarizing histopathological images by representing nuclei with learned discriminative nuclear codewords. A reliable patch-based nuclear segmentation scheme using convolutional neural networks was also developed to extract the nuclear features. Huo *et al.* proposed a new multi-task learning model to analyze the associations between single nucleotide polymorphisms (SNPs) and quantitative traits (imaging measures). The low-rank structure in the new multi-task learning model is beneficial to uncover the correlation between genetic variations and imaging phenotypes, such that the candidate genes or loci which is relevant to the biological etiology of the disease can be identified. Adhikari *et al.* developed a multi-site resting state functional MRI (rsfMRI) analysis pipeline to allow research groups around the world to process rsfMRI scans in a harmonized way, to extract consistent and quantitative measurements of connectivity and to perform coordinated statistical tests. The ENIGMA-rsfMRI analysis pipeline was used to ver-

ify and replicate the assertion that there is moderately strong genetic influence on the resting state signal. Han and Kamdar used a bi-directional convolutional recurrent neural network model to predict the methylation state of the MGMT regulatory regions in Glioblastoma Multiforme (GBM) patients via their brain MRI scans collected from TCIA combined with methylation data from TCGA. Finally, Srivastava *et al.* studied the coherent "trans-omics" features that characterize varied clinical cohorts across multiple sources of TCGA data for more descriptive and robust disease characterization. The results showed that the histology images outperformed molecular features while predicting cancer stages, transcriptomics held superior discriminatory power for ER-Status and PAM50 subtypes, and there exist a few cases where all data modalities exhibited comparable performance.

References

1. M. McCarthy. US to launch major brain research initiative. *BMJ*, 346:f2156, 2013.
2. F. S. Collins and H. Varmus. A new initiative on precision medicine. *N Engl J Med*, 372(9):793–5, 2015.
3. L. Ohno-Machado. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. *J Am Med Inform Assoc*, 21(2):193, 2014.

Heritability estimates on resting state fMRI data using ENIGMA analysis pipeline

Bhim M. Adhikari
Maryland Psychiatry Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA
Email: badhikari@mprc.umaryland.edu

Neda Jahanshad
Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, Marina del Rey, CA, USA
Email: neda.jahanshad@usc.edu

Dinesh Shukla
Maryland Psychiatry Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA
Email: dshukla@mprc.umaryland.edu

David C. Glahn
Department of Psychiatry, Yale University, School of Medicine, New Haven, CT, USA
Email: david.glahn@yale.edu

John Blangero
Genomics Computing Center, University of Texas at Rio Grande Valley
Email: john.blangero@UTRGV.edu

Richard C. Reynolds
National Institute of Mental Health, Bethesda, MD, USA
Email: reynoldr@mail.nih.gov

Robert W. Cox
National Institute of Mental Health, Bethesda, MD, USA
Email: robertcox@mail.nih.gov

Els Fieremans
Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, NY, USA
Email: els.fieremans@nyumc.org

Jelle Veraart
Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, NY, USA
Email: Jelle.Veraart@nyumc.org

Dmitry S. Novikov
Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, NY, USA
Email: Dmitry.Novikov@nyumc.org
Thomas E. Nichols

Department of Statistics, University of Warwick, Coventry, CV47AL, UK
Email: t.e.nichols@warwick.ac.uk

L. Elliot Hong
Maryland Psychiatry Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA
Email: ehong@mprc.umaryland.edu

Paul M. Thompson
Imaging Genetics Center, Stevens Neuroimaging & Informatics Institute, Keck School of Medicine of USC, Marina del Rey, CA, USA
Email: pthomp@usc.edu

Peter Kochunov
Maryland Psychiatry Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA
Email: pkochunov@mprc.umaryland.edu

Big data initiatives such as the Enhancing NeuroImaging Genetics through Meta-Analysis consortium (ENIGMA), combine data collected by independent studies worldwide to achieve more generalizable estimates of effect sizes and more reliable and reproducible outcomes. Such efforts require harmonized image analyses protocols to extract phenotypes consistently. This harmonization is particularly challenging for resting state fMRI due to the wide variability of acquisition protocols and scanner platforms; this leads to site-to-site variance in quality, resolution and temporal signal-to-noise ratio (tSNR). An effective harmonization should provide optimal measures for data of different qualities. We developed a multi-site rsfMRI analysis pipeline to allow research groups around the world to process rsfMRI scans in a harmonized way, to extract consistent and quantitative measurements of connectivity and to perform coordinated statistical tests. We used the single-modality ENIGMA rsfMRI preprocessing pipeline based on model-free Marchenko-Pastur PCA based denoising to verify and replicate resting state network heritability estimates. We analyzed two independent cohorts, GOBS (Genetics of Brain Structure) and HCP (the Human Connectome Project), which collected data using conventional and connectomics oriented fMRI protocols, respectively. We used seed-based connectivity and dual-regression approaches to show that the rsfMRI signal is consistently heritable across twenty major functional network measures. Heritability values of 20-40% were observed across both cohorts.

Key words: functional connectivity, heritable, seed-based connectivity

1. Introduction

Resting state functional MRI (rsfMRI) studies investigate large-amplitude, spontaneous low-frequency fluctuations in the fMRI signal that are temporally correlated across functionally related brain areas [1-4]. It is the basis for a powerful method to evaluate temporal correlations

of low-frequency blood oxygenation level-dependent fluctuations across brain regions in the absence of a task or stimulus [2, 4]. Genetic analyses on rsfMRI phenotypes are challenged by limited statistical power. One way to address this is by pooling data from multiple cohorts or studies. The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) consortium has developed an rsfMRI analysis pipeline to perform consistent analysis and extraction of resting state connectivity measures across data collected using diverse protocols [7]. Here, we demonstrate the utility of this pipeline by replicating findings of significant heritability in the default mode network in (A) the original Genetics of Brain Structure (GOBS) cohort [10], and (B) data from the Human Connectome Project (HCP; [11]); we also aim to demonstrate consistent additive genetic contribution to intersubject variance in other intrinsic brain networks. The motivation for this work is that, ultimately, it should be possible to discover genetic variants that reliably affect brain function, but a key milestone in this quest is to establish that the metrics targeted are heritable, i.e., individual genetic variance accounts for a significant proportion of their variation across subjects. This paper is one key step in this process.

The ENIGMA rsfMRI pipeline differs from existing pipelines in two notable respects. Many fMRI analysis pipelines require input from multiple imaging modalities. Commonly, a structural T1-weighted (T1w) MRI scan is required to regress out signal trends from cerebrospinal fluid (CSF) and cerebral white matter, and T1w data is used for anatomical registration to an atlas space [5, 6]. In the spirit of other ENIGMA pipelines, our pipeline for rsfMRI is a single-modality pipeline. It uses a deformable template created from 1,100 individual images provided by ENIGMA sites to incorporate shape distortions common to fMRI images [7]. The pipeline uses direct tissue classification of rsfMRI data to regress out some of the variance due to methodical (non-biological) factors. Both approaches avoid the potential pitfalls of site-to-site variance in T1w data and coregistration biases that may influence the rsfMRI phenotypes.

In another notable difference, we use a novel denoising technique based on the Marchenko-Pastur distribution [8] - fitted to the eigenvalues of a principal component analysis (MP-PCA) across space and time - to identify and remove the principal components originating due to thermal noise. We used this MP-PCA-based denoising technique to reduce signal fluctuations rooted in thermal noise and hence increase the tSNR without altering the spatial resolution. The ability to suppress thermal noise is based on data redundancy in the PCA domain, using universal properties of the eigenspectrum of random covariance matrices [8]. The bulk of the PCA eigenvalues arise due to noise and can be asymptotically represented by the universal MP distribution, in the limit of the large signal matrix size (voxels × time points). The Marchenko-Pastur parameterization allows us to identify noise-only components and to estimate the noise level in a local neighborhood based on the singular value decomposition of a signal matrix combining neighborhood voxels [9]. After removing noise-only components, the resulting images show enhanced SNR; the residual noise is contained in the remaining components, which cannot be further denoised by this method.

Here, we validate the ENIGMA-rsfMRI analysis pipeline by attempting to verify and replicate the assertion that there is moderately strong genetic influence on the resting state signal [10], and that significant heritability estimates are consistently found across independent cohorts. While the actual value of the heritability estimate may depend on cohort parameters such as their demographics, age, and environment – and presumably also on the image SNR – a key goal of a collaborative imaging genetics initiative is to identify brain metrics that are significantly heritable across cohorts, as a precursor to a more in-depth search for common variants associated with the trait, in this case brain function. Glahn and colleagues showed that individual variance in measures of connectivity within intrinsic brain networks - such as the default-mode network - is influenced by genetic factors: for some metrics, ~40% of the variance could be attributed to additive genetic factors [10]. The ENIGMA-rsfMRI pipeline was used to measure individual variations in the default mode network and perform heritability analysis in the same dataset used by Glahn and colleagues (N=334), using two complementary measurements of connectivity: region-based (also called "seed"-based) analysis and dual regression. We further expanded this analysis by demonstrating that similar heritability estimates can be obtained for other intrinsic brain networks and replicated the detection of significant heritability estimates in the data from a young adult sample collected by HCP (N=518). The detection and estimation of additive genetic effects may depend on the degree of relatedness across individuals underlying the sample structure. We tested heritability measurements computed from two commonly used familial study designs: GOBS subjects were recruited from an extended pedigree and HCP subjects were recruited from a twin/siblings registry. Similarity in the heritability measurements across two diverse cohorts would therefore be further evidence to support the suitability of the ENIGMA rsfMRI protocol and connectivity measurements for large-scale genetic analyses of cerebral functional connectivity.

2. Methods and Materials

2.1. Study subjects and imaging protocols:

Two rsfMRI datasets were analyzed (GOBS and HCP: acronyms are detailed below).

2.1.1. GOBS - Genetics of Brain Structure and Function study

Subjects: This sample comprised 334 (124 M/210 F, mean age: 47.9±13.2 years) Mexican-American individuals from 29 extended pedigrees (average family size = 9 people; range 5-32) who participated in the Genetics of Brain Structure and Function study. Individuals in this cohort have actively participated in genetics research for over 20 years and, were randomly selected from the community with the constraints that they are of Mexican-American ancestry, part of a large family, and live within the San Antonio, TX region. In this study, individuals were excluded for MRI contraindications, history of neurological illnesses, or stroke or other major neurological event. All participants provided written informed consent on forms approved by the institutional review board at the University of Texas Health Science Center San Antonio

(UTHSCSA).

Imaging: All imaging was performed at the Research Imaging Institute, UTHSCSA, on a Siemens 3 T Trio scanner using a multichannel phased array head coil. Whole-brain, resting-state functional imaging was performed using a gradient-echo echo planer imaging (EPI) sequence sensitive to the BOLD effect with the following parameters: TR=3000 ms, TE=30 ms, spatial resolution=1.72×1.72×3 mm^3, flip angle=90 degrees. The resting-state protocol included 43 slices acquired parallel to the sagittal plane containing the anterior and posterior commissures; scan time was 7.5 min.

2.1.2. HCP – Human Connectome Project

Subjects: We included rsfMRI data, 518 participants (240 M/278 F; mean age 28.7 ± 3.7 years) from the Human Connectome Project (HCP) dataset, released in March 2017. Participants were recruited from the Missouri Family and Twin Registry [11]. All HCP participants were from young adult sibships of average size 3–4 that include a monozygotic or dizygotic twin pair and (where available) their non-twin siblings. Subjects ranged in age from 22 to 37 years. This age range was chosen as it corresponds to a period after neurodevelopment is largely completed and before the typical age of onset of neurodegenerative changes. The inclusion and exclusion criteria are detailed elsewhere [11]. The HCP subjects are healthy young adults within a restricted age range and free from major psychiatric or neurological illnesses [12, 13]. All subjects provided written informed consent on forms approved by the Institutional Review Board of Washington University in St Louis.

Imaging: All HCP subjects are scanned on a customized Siemens 3 T "Connectome Skyra" scanner housed at Washington University in St. Louis, using a standard 32-channel Siemens receive head coil. RsfMRI data consisted of two runs in one session. Within a session, oblique axial acquisitions alternated between phase encoding in a right-to-left direction in one run and phase encoding in a left-to-right direction in the other run. Resting state images were collected using a gradient-echo echo planar imaging (EPI) sequence with the following parameters: TR=720 ms, TE=33.1 ms, flip angle=52 degrees, FOV=208×180 mm (RO×PE), matrix=104×90 (RO×PE), 2.0 mm isotropic voxels, 72 axial slices, multiband factor=8; scan time was 28.8 min.

2.2. Functional Image Analysis

The rsfMRI data processing was carried out using the ENIGMA resting state analysis pipeline implemented in the Analysis of Functional NeuroImages (AFNI) software [14]. ENIGMA developed a single-modality resting state analysis pipeline [7]. The ENIGMA pipeline is an extension of the conventional AFNI rsfMRI pipeline [14] (**Figure 1**). The first step is the application of principal components analysis (PCA)-based denoising [8, 9], to improve signal-to noise ratio (SNR) and temporal SNR (tSNR) properties of the time series data, with no loss of spatial resolution of the image and without the introduction of additional partial volume effects

[7]. This denoising approach is free from the limitations of the loss of spatial resolution of the image and introduction of additional partial volume effects that lead to complications in further quantitative analyses [8]. The MP-PCA approach does not alter the resting state network activation patterns, whereas spatial smoothing using a Gaussian kernel leads to partial voxel averaging, spreading the activations across gray and white matter regions and removing smaller nodes. Finally, the noise-maps produced by MP-PCA approach provide valuable information for quality control as deviations from the expected uniform or slowly varying in space pattern of thermal noise may indicate problems with the coil or other scanner hardware.

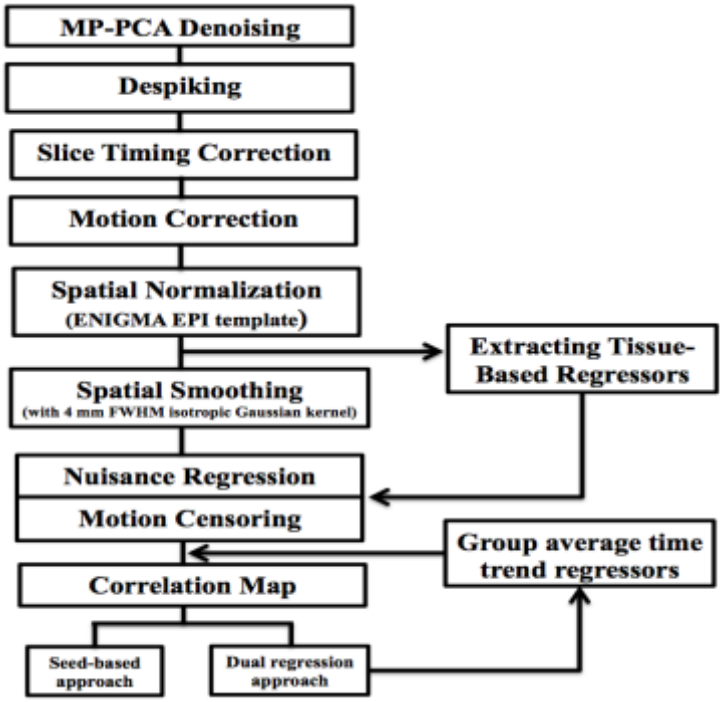


Fig. 1. Flowchart of ENIGMA rsfMRI analysis pipeline.

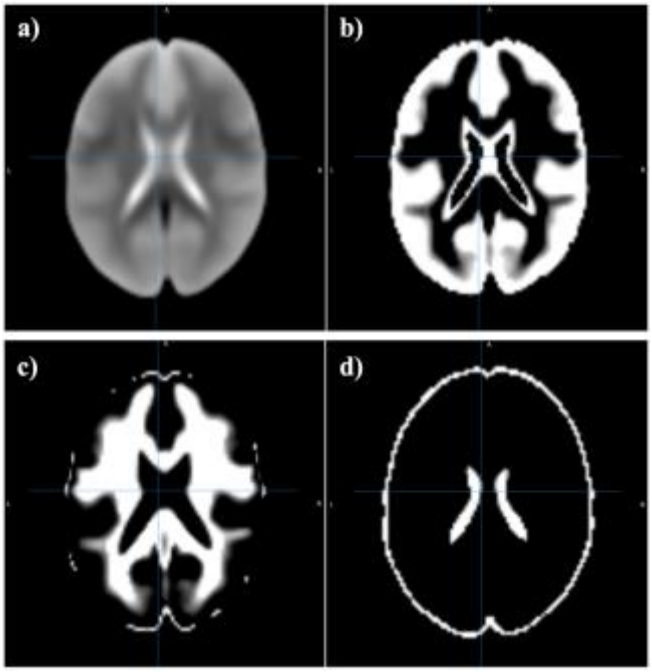


Fig. 2. ENIGMA EPI brain template (a) and segmented tissue classes (b-d) for gray matter, white matter and cerebrospinal fluid, respectively.

In the next step, supplementary data, if provided, is used for correction of spatial distortions associated with long-TE gradient echo imaging. Two available corrections are the gradient-echo 'fieldmap' or the reversed-gradient approach. In the next step, a transformation is computed registering the base volume to the ENIGMA EPI template (**Figure 2**) that was derived from 1,100 datasets corrected across 22 sites [7] to develop a spatial template and spatial atlas. This atlas has a dual purpose: it is used for regression of the global signal, and also offers a common anatomical spatial reference frame. Next, correction for head motion is performed by registering each functional volume to the volume with the minimum outlier fraction (suggesting it has little motion), where each transformation is

concatenated with the transformation to standard space, to avoid unnecessary interpolation. Nuisance variables such as the linear trend, 6 motion parameters (3 rotational and 3 translational directions), their 6 temporal derivatives (rate of change in rotational and translational motion) and time courses from the local white matter and cerebrospinal fluid (CSF) from lateral ventricles were modeled using multiple linear regression analysis, which were then removed as regressors of no interest. Time points with excessive motion (> 0.2 mm), estimated as the magnitude of displacement from one time point to the next, including neighboring time points and outlier voxels fraction (>0.1) were censored from statistical analysis. Images were spatially normalized to the ENIGMA EPI template in Montreal Neurological Institute (MNI) standard space for group analysis.

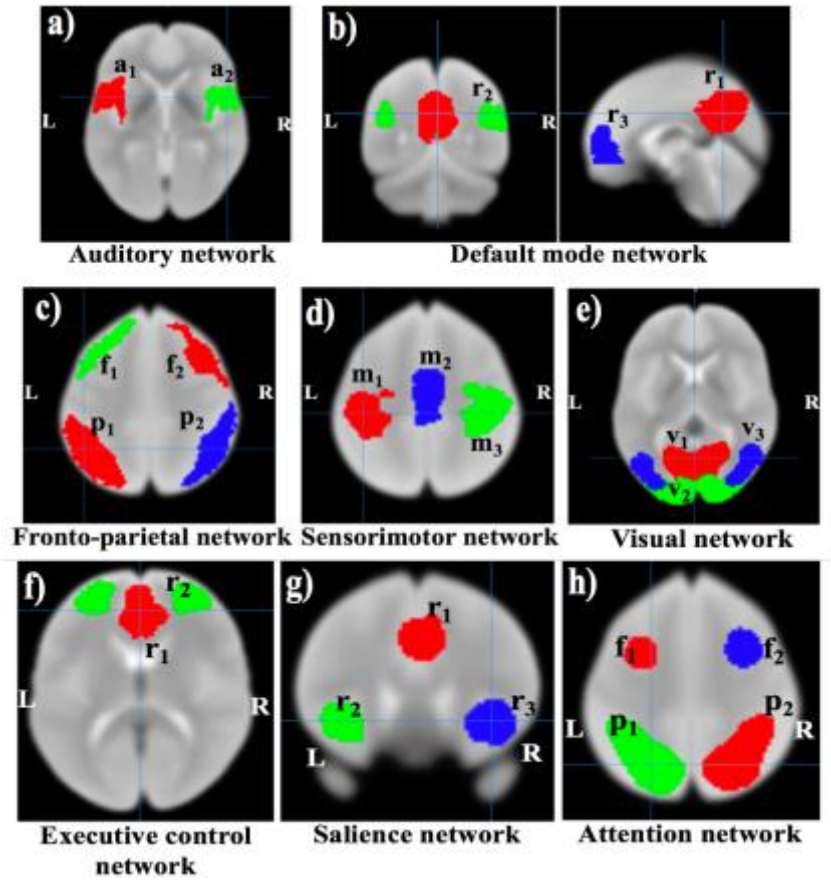


Fig.3. Resting state network template ROIs based on the BrainMap activation database and resting fMRI dataset [4]. Here, L=left, R=right, in (a) a_1=left primary and association auditory cortices, a_2=right primary and associated auditory cortices, in (b) r_1=posterior cingulate/precuneus, r_2=bilateral temporal-parietal regions and, r_3=ventromedial frontal cortex, in (c) f_1/f_2=left/right frontal area and p_1/p_2=left/right parietal area, in (d) m_1/m_3=left/right motor area and m_2=supplementary motor area, in (e) v_1=medial visual areas, v_2= occipital visual areas, and v_3=lateral visual areas, in (f) r_1=anterior cingulate cortex and r_2=bilateral medial frontal gyrus, in (g) r_1=anterior cingulate cortex and r_2/r_3=left/right insula, in (h) f_1/f_2=left/right middle frontal gyrus and p_1/p_2=left/right superior parietal lobule.

2.3. Functional connectivity analysis

Resting state network templates were defined based on the probabilistic regions of interest (ROIs) from 20-component analysis of the BrainMap activation database and resting fMRI dataset [4]. We defined the binary masks of the resting state template regions from auditory network (AN), default mode network (DMN), fronto-parietal network (FPN), sensorimotor network (SMN), visual network (VN), executive control network (ECN), salience network (SN), and attention network (AttN) (**Figure 3**). Mean time series were extracted from the seed regions of each network and connectivity maps corresponding to each seed region were obtained by assessing correlations along the time series for different regions. Next, Fisher's *r*-to-*z* transformations were applied

to obtain a normal distribution. We calculated seed-based functional connectivity values between seed regions in each network and performed heritability calculation. For the HCP dataset, heritability measures were calculated for all subjects (N=518) under consideration and, in a separate analysis, for subjects (N=481) with censored TRs less than 15% of the total TRs during the processing. Furthermore, we performed dual regression analysis for the default mode network template ROIs, and calculated the functional connectivity measures and hence heritability estimates on the GOBS dataset (N=334). (This same dataset was used in a prior study by David Glahn and colleagues [10]). In the case of dual regression, for the given network template ROIs, average single time series were computed from the preprocessed data for each subject, and hence we obtained the average time series from all subjects, and then averaged these to obtain an average time series that represents the group average time trend. The group average time trend was regressed out from each subject's data, before calculating the functional connectivity values.

2.4. Heritability estimation

For the heritability estimations, the variance components method was used, as implemented in the Sequential Oligogenic Linkage Analysis Routines (SOLAR) Eclipse software package (http://www.nitrc.org/projects/se_linux) [15]. SOLAR uses maximum likelihood variance decomposition methods, extensions of the strategy developed by Amos and colleagues [16]. The covariance matrix Ω for a pedigree is given by: $\Omega = 2\Phi\sigma_g^2 + I\sigma_e^2$, where σ_g^2 is the genetic variance due to the additive genetic factors, ϕ is the kinship matrix representing the pair-wise kinship coefficients among all individuals, σ_e^2 is the variance due to individual — unique environmental effects and measurement error, and I is an identity matrix (under the assumption that all environmental effects are uncorrelated among family members). Narrow sense heritability is defined as the fraction of phenotypic variance σ_p^2 attributable to additive genetic factors,

$$h^2 = \frac{{\sigma_g}^2}{{\sigma_p}^2} \quad (1)$$

The variance parameters are estimated by comparing the observed phenotypic covariance matrix with the covariance matrix predicted by kinship [15]. Significance of the heritability estimate is tested by comparing the likelihood of the model in which σ_g^2 is constrained to zero with that of a model in which σ_g^2 is estimated. Twice the difference between the log-likelihoods of these models yields a test statistic, which is asymptotically distributed as a 1/2:1/2 mixture of χ^2 variables with 1 degree-of-freedom and a point mass at zero. Prior to the heritability estimation, phenotype values from each dataset were adjusted for covariates including sex, age, age^2, age×sex interaction, and age^2×sex interaction. Inverse Gaussian transformation was also applied to ensure normality of the distribution. Outputs from SOLAR include the heritability estimate (h^2), the significance value (p), and the standard error for each trait (SE).

3. Results

3.1. Seed-based analysis. Heritability estimates for connectivity measurements extracted from the seed-based approach are summarized in **Table 1**. The default mode network (DMN) showed significant heritability for the connectivity from the posterior cingulate/precuneus to bilateral temporal-parietal regions (h^2=0.34±0.16, p=0.014) and ventromedial frontal cortex (h^2=0.35±0.17, p=0.014) respectively for the GOBS dataset. Replication analyses, in the HCP dataset, demonstrated significant heritability in the functional connectivity measures from all node pairs of the DMN. In the fronto-parietal network, we found significantly heritable functional connectivity in both datasets. Heritability estimates in other networks showed a similar pattern of genetic control in both GOBS and HCP with greater evidence for statistical significance (i.e., lower p-values) observed in HCP subjects. Heritability estimates were found to be improved by excluding subjects who had more than 15% of total TRs censored due to motion from the HCP dataset (HCP, N=481).

3.2. Dual regression analysis. To confirm the heritability estimates from seed based connectivity, a dual regression analysis was also conducted in GOBS. Here, connectivity values for the DMN ROIs were again significantly heritable from posterior cingulate/precuneus to bilateral temporal-parietal regions (h^2=0.31±0.17, p=0.027) and ventromedial frontal cortex (h^2=0.25±0.17, p=0.038) **(Table 2)**. The connection from bilateral temporal-parietal regions to posterior cingulate/precuneus was likewise significantly heritable (h^2=0.26±0.16, p=0.035).

Table 1. Heritability estimates for measures derived from resting state networks (RSNs). *Regions are based off of **Figure 3**. Bolded connections are significant at 5% FDR. #Estimated heritability, h^2 (SE). Abbreviations: GOBS= Genetic of Brain Structure and Function study, HCP=Human Connectome Project, DMN=default mode network, FPN=fronto-parietal network, SMN=sensorimotor network, VN=visual network, SN=salience network, AttN=attention network, ECN=executive control network, AN=auditory network.

Network		GOBS (N=334)		HCP (N= 518)		HCP (N=481)	
	Regions*	Heritability#	p-value	Heritability#	p-value	Heritability#	p-value
DMN	r_1-r_2	**0.34 (0.16)**	0.014	**0.27 (0.09)**	1.0×10^{-7}	**0.28 (0.09)**	1.0×10^{-7}
	r_2-r_3	0	0.500	**0.14 (0.09)**	0.008	**0.13 (0.09)**	0.014
	r_3-r_1	0.09 (0.15)	0.276	**0.15 (0.11)**	0.025	0.12 (0.11)	0.059
	r_2-r_1	0.09 (0.13)	0.244	**0.27 (0.09)**	4.6×10^{-8}	**0.25 (0.09)**	2.0×10^{-7}
	r_3-r_2	0	0.500	**0.23 (0.12)**	0.002	**0.21 (0.12)**	5.7×10^{-3}
	r_1-r_3	**0.35 (0.17)**	0.014	0.09 (0.1)	0.120	0.08 (0.09)	0.094
FPN	f_1-p_1	0.14 (0.14)	0.149	**0.16 (0.11)**	0.019	**0.16 (0.10)**	0.018
	p_1-f_1	0.13 (0.14)	0.169	**0.16 (0.11)**	0.018	**0.13 (0.11)**	0.044
	f_2-p_2	**0.31 (0.15)**	0.016	**0.19 (0.14)**	0.034	**0.26 (0.14)**	0.021
	p_2-f_2	**0.29 (0.15)**	0.025	**0.27 (0.14)**	0.042	**0.36 (0.13)**	0.009
SMN	m_1-m_2	0.09 (0.14)	0.255	**0.29 (0.15)**	0.017	**0.35 (0.14)**	0.006
	m_2-m_3	0	0.500	0.14 (0.12)	0.113	0.14 (0.13)	0.135
	m_3-m_1	**0.32 (0.20)**	0.041	**0.27 (0.14)**	0.009	**0.32 (0.14)**	0.007
	m_2-m_1	0.06 (0.12)	0.302	0	0.500	0	0.500
	m_3-m_2	0	0.500	0.15 (0.13)	0.108	**0.24 (0.14)**	0.044
	m_1-m_3	**0.32 (0.20)**	0.045	**0.25 (0.13)**	0.008	**0.32 (0.14)**	0.005

VN	v_1-v_2	0.210 (0.15)	0.062	**0.14 (0.09)**	0.021	**0.15 (0.09)**	0.019
	v_2-v_3	**0.36 (0.14)**	0.004	**0.17 (0.11)**	0.029	**0.20 (0.11)**	0.017
	v_3-v_1	0.12 (0.14)	0.168	0.03 (0.04)	0.191	0.05 (0.06)	0.145
	v_2-v_1	**0.32 (0.15)**	0.009	**0.13 (0.09)**	0.042	**0.18 (0.11)**	0.017
	v_3-v_2	0.13 (0.14)	0.161	**0.15 (0.09)**	0.040	**0.19 (0.11)**	0.017
	v_1-v_3	0.17 (0.14)	0.100	0.06 (0.05)	0.060	**0.09 (0.06)**	0.030
SN	r_1-r_2	0.20 (0.13)	0.062	0.07 (0.09)	0.121	0.06 (0.08)	0.166
	r_2-r_3	**0.24 (0.12)**	0.019	**0.25 (0.11)**	0.002	**0.28 (0.13)**	0.002
	r_3-r_1	0	0.500	**0.13 (0.08)**	0.005	**0.12 (0.08)**	0.013
	r_2-r_1	0.16 (0.12)	0.084	**0.20 (0.11)**	0.002	**0.17 (0.11)**	0.008
	r_3-r_2	**0.18 (0.12)**	0.049	**0.31 (0.12)**	3.8×10^{-4}	**0.32 (0.12)**	4.0×10^{-4}
	r_1-r_3	0	0.500	0.05 (0.06)	0.142	0.04 (0.06)	0.196
AttN	f_1-p_1	**0.20 (0.12)**	0.031	0.10 (0.15)	0.288	0.08 (0.19)	0.384
	p_1-f_1	**0.21 (0.12)**	0.024	0.08 (0.11)	0.213	0.05 (0.10)	0.274
	f_2-p_2	**0.32 (0.12)**	0.001	**0.27 (0.14)**	0.018	**0.32 (0.14)**	0.011
	p_2-f_2	**0.31 (0.12)**	0.002	**0.32 (0.14)**	0.005	**0.35 (0.14)**	0.004
ECN	r_1-r_2	0.17 (0.14)	0.088	**0.17 (0.11)**	0.023	**0.18 (0.12)**	0.015
	r_2-r_1	**0.23 (0.14)**	0.034	**0.23 (0.11)**	0.003	**0.22 (0.11)**	0.002
AN	a_1-a_2	0.12 (0.16)	0.209	0.05 (0.09)	0.275	0.03 (0.07)	0.336
	a_2-a_1	0.05 (0.14)	0.356	0.05 (0.08)	0.260	0.04 (0.08)	0.303

Table 2. Heritability estimates for measures derived from DMN (GOBS). * Regions are based off of **Figure 3.** Bolded connections are significant after multiple comparisons correction with FDR at q=5%. #Estimated heritability, h^2 (SE).

Network		Seed-based approach		Dual regression approach	
	Regions*	Heritability#	p-value	Heritability#	p-value
DMN	r_1-r_2	**0.34 (0.16)**	0.014	**0.30 (0.17)**	0.027
	r_2-r_3	0	0.500	0	0.500
	r_3-r_1	0.09 (0.15)	0.276	0.09 (0.11)	0.279
	r_2-r_1	0.09 (0.13)	0.244	**0.26 (0.16)**	0.035
	r_3-r_2	0	0.500	0	0.500
	r_1-r_3	**0.35 (0.17)**	0.014	**0.25 (0.17)**	0.038

3.3. Execution time. All analyses were carried out at the Center for High-Performance Computing at Washington University. The processing time of ENIGMA rsfMRI pipeline varied with complexity of the dataset. Dual regression analysis of GOBS resting data consisting of N=150 fMRI volume took about 30 min per subject/network on a modern linux server node. Analysis of HCP data of N=2400 fMRI volumes took about 6 hours per subject/network.

4. Discussion

We applied the ENIGMA rsfMRI pipeline to two datasets (GOBS and HCP), collected ten years apart, to demonstrate that we could consistently detect genetic influences on resting state connectivity. Building on prior work in individual cohorts, this experiment provides direct evidence that connectivity within the DMN and other intrinsic brain networks is influenced by genetic factors. We found between 20-40% of the intersubject variance in functional connectivity within functional networks was under genetic control. Our findings replicate previously reported heritability measurements in the GOBS cohort and extend this research by conducting harmonized analyses in the HCP subjects. The pattern of heritability was similar between two cohorts collected using very different imaging protocols and sample designs. Together, these findings strongly suggest that resting state connectivity is under a moderate genetic control and this heritability can be detected in, in terms of image acquisition, both legacy and state-of-the art samples. Establishing the consistency of the heritability of resting state functional connectivity provides critical information necessary before these measures can be appropriately used in genetic studies designed to identify or functionally characterize genes influencing measures of brain function. Showing reproducible and significant heritability is necessary before indices of default-mode functional connectivity can be considered as an intermediate phenotype or endophenotype for in-depth genetic analyses.

Prior works on rsfMRI analysis are generally multimodal and rely on spatial co-alignment of subject's structural (T1w) and rsfMRI data for regressions of global connectivity signals and ROI analyses in a common anatomical frame. The site-to-site variability in the quality of the T1w data and the variance in registration quality between T1w and rsfMRI images may influence the results of the overall rsfMRI analysis. To minimize these potential pitfalls, we have used the ENIGMA rsfMRI analysis pipeline - a unimodal analysis workflow that uses a deformable ENIGMA EPI template to serve the dual purpose of regression of the global signal and offering a common anatomical spatial reference frame. The use of this deformable template greatly improves registration for individual EPI images, including ventricular overlap, when compared to the standard ICBM-152 template [7]. This analysis pipeline also incorporates the MPPCA denoising algorithm, which helps to improve SNR/tSNR properties of the time series data [8,9], with no loss of spatial resolution of the image and without the introduction of additional partial volume effects [7]. In addition, it includes the two complementary measurements of connectivity: region-based analysis and dual regression, to test heritability estimations on functional connectivity measurements.

The ENIGMA rsfMRI pipeline is built using the NIH-supported software - AFNI - that is freely available to both non-commercial and commercial users. The use of free-license software opens ENIGMA collaboration to commercial entities such as pharmacological companies. It is a unimodal analysis workflow designed for consistent retrospective analyses of state-of-the-art and legacy data. The pipeline incorporates stringent quality assurance (QA) and quality control steps. It incorporates traditional QA measurements to detect and censor motion and other types artifacts that are detectable visually. It also uses novel analysis of the heterogeneity of the thermal noise within imaging volume to enable identification of more subtle artifacts such as time-and-space

related variability in the coil sensitivity profiles. The efforts to compare the performance of ENIGMA rsfMRI analysis pipeline across multiple cohorts with other rsfMRI analysis pipelines are ongoing.

Acknowledgments

Support was received from NIH grants U54EB020403, U01MH108148, 2R01EB015611, R01MH112180, R01DA027680, R01MH085646.

We are grateful to Human Connectome Project and the Center for High-Performance Computing at Washington University, University of Washington St. Louis for allowing the use of their computational facility for this project.

References

1. B. Biswal, F.Z. Yetkin, V.M. Haughton, et al., *Magn. Reson. Med.*, **34**: 537 (1995).
2. M.D. Fox, and M.E. Raichle, *Nat. Rev. Neurosci.*, **8**(9): 700 (2007).
3. D.S. Margulies, A.M. Clare Kelly, L.Q. Uddin, et al., *Neuroimage*, **37**: 579 (2007).
4. S.M. Smith, P.T. Fox, K.L. Miller, et al., *Proc Natl Acad Sci U S A*, **106**: 13040 (2009).
5. M. Jenkinson, C.F. Beckmann, T.E. Behrens, et al., *Neuroimage*, **62**: 782 (2012).
6. S.M. Smith, M. Jenkinson, M.W. Woolrich, et al., *NeuroImage*, **23**: S208 (2004).
7. B.M. Adhikari, N. Jahanshad, D.K. Shukla, et al., *Brain Imaging Behav*, (under review), (2017).
8. J. Veraart, D.S. Novikov, D. Christiaens, et al., *Neuroimage*, **142**: 394 (2016).
9. J. Veraart, E. Fieremans, and D.S. Novikov, *Magn. Reson. Med.*, **76**: 1582 (2016).
10. D.C. Glahn, A.M. Winkler, P. Kochunov, et al., *Proc Natl Acad Sci U S A*, **107**(3): 1223 (2010).
11. D.C. van Essen, S.M. Smith, D.M. Barch, et al., *Neuroimage*, **80**: 62 (2013).
12. E.L. Edens, A.L. Glowinski, M.L. Pergadia, et al., *J. Addict. Med.*, **4**(1): 55 (2010).
13. C.E. Sartor, V.V. McCutcheon, N.E. Pommer, et al., *Psychol Med*, **41**(7): 1497 (2011).
14. R.W. Cox, , *Comput Biomed Res*, **29**: 162 (1996).
15. L. Almasy, and J. Blangero, *Am. J. Hum. Genet.*, **62**: 1198 (1998).
16. C.I. Amos, *Am. J. Hum. Genet.* , **54**: 535 (1994).

Discriminative bag-of-cells for imaging-genomics

Benjamin Chidester
Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
E-mail: bchidest@cs.cmu.edu

Minh N. Do
Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA
E-mail: minhdo@illinois.edu

Jian Ma
Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
Email: jianma@cs.cmu.edu

Connecting genotypes to image phenotypes is crucial for a comprehensive understanding of cancer. To learn such connections, new machine learning approaches must be developed for the better integration of imaging and genomic data. Here we propose a novel approach called Discriminative Bag-of-Cells (DBC) for predicting genomic markers using imaging features, which addresses the challenge of summarizing histopathological images by representing cells with learned discriminative types, or codewords. We also developed a reliable and efficient patch-based nuclear segmentation scheme using convolutional neural networks from which nuclear and cellular features are extracted. Applying DBC on TCGA breast cancer samples to predict basal subtype status yielded a class-balanced accuracy of 70% on a separate test partition of 213 patients. As data sets of imaging and genomic data become increasingly available, we believe DBC will be a useful approach for screening histopathological images for genomic markers. Source code of nuclear segmentation and DBC are available at: https://github.com/bchidest/DBC.

Keywords: Imaging-genomics; Histopathological image analysis; Computational pathology

1. Introduction

Cancer is a genetic disease that develops from accumulated genomic alterations that disrupt normal cellular processes and give rise to phenotypic changes, such as cell size, shape, and structural relationship within a tumor.[1] High-throughput genomic approaches have revealed new understandings of the complexity of cancer. We now know that even patient samples from the same type of cancer may exhibit a high level of inter-tumor heterogeneity.[2] For example, breast cancer patients can be largely categorized into four molecular subtypes (luminal A and

B, HER2-enriched, and basal) with distinct prognosis.[3,4] Additionally, 'imaging-genomics' has been coined to refer to recent developments in leveraging new insights gained from genomics with traditional imaging of radiology or histopathology.[5–8] For histopathological images, features of cells and nuclei, which are used by pathologists to diagnose cancer, provide the most direct connection to the genomic signatures of a patient's tumor. Yet whole slide images (WSIs) contain tens of thousands of cells with diverse characteristics, which makes associating phenotype with genomics challenging. Previous works in imaging-genomics have sought to draw connections by first clustering nuclei and cells into types in an unsupervised fashion and then associating with genomic markers, such as gene expression.[5,7,8] Others have looked for connections with specific cell types, leveraging biological understanding, such as the affect of the cellularity of lymphocytes on copy number variation in tumors.[6] There remains a need for machine learning tools that can effectively capture the diversity of cellular phenotypes within and across tumors for high-throughput investigation of general image-genomic associations.

To address this need, we propose a novel, general framework for predicting arbitrary genomic markers from imaging features called Discriminative Bag-of-Cells (DBC). In this framework, cells within a histopathological image are grouped into types and the image is summarized succinctly by a histogram of cell types, and a classifier is learned from these histograms of types to predict genomic markers (e.g., mutation, gene expression, or molecular subtype). Our framework is inspired by the bag-of-words (BoW) approach, which has been successfully applied to document classification and image classification.[9] In addition to learning the BoW classifier, our method also learns the cell-type, or codeword, assignments in a discriminative fashion to find types that are more informative of the specific genomic marker. This avoids the trouble of unsupervised clustering of cellular features,[5,7,8] which does not optimize cell-type assignments for genomic markers of interest. Some works have also proposed a similar training of discriminative codebooks,[10,11] which inspired this particular enhancement for our BoW framework for histopathological images.

DBC has several advantages over other methods of high-throughput histopathological image analysis. A primary advantage is the interpretability of the learned model. From the learned cell types, we can distinguish what types of nuclei are important for classification. This allows us to trace back to the original nuclei in the sample to visualize how they are categorized and to learn which types of nuclei in the images are discriminative for specific genomic markers. Another advantage of DBC is that it can learn what heterogeneity within a tumor is informative of a particular genomic marker and what is not. For example, it is known that a tumor can be comprised of multiple molecular subtypes,[2] but there is still a need to understand what cellular phenotypes are unique to which subtype. DBC seeks to account for such diversity by learning how the mixing of cell types is informative of an assigned subtype.

Our framework relies upon extraction of nuclear features from histopathological images, from which the cell types are learned. For hematoxylin and eosin (H&E) histopathological imaging, due to the high degree of variation in sample acquisition, such as stain intensity and slice thickness, and the care required to produce quality samples, segmenting nuclei and extracting image features is often unreliable. Most studies have relied upon popular microscopy image analysis software, such as CellProfiler,[12] to perform image feature extraction,[13] but the

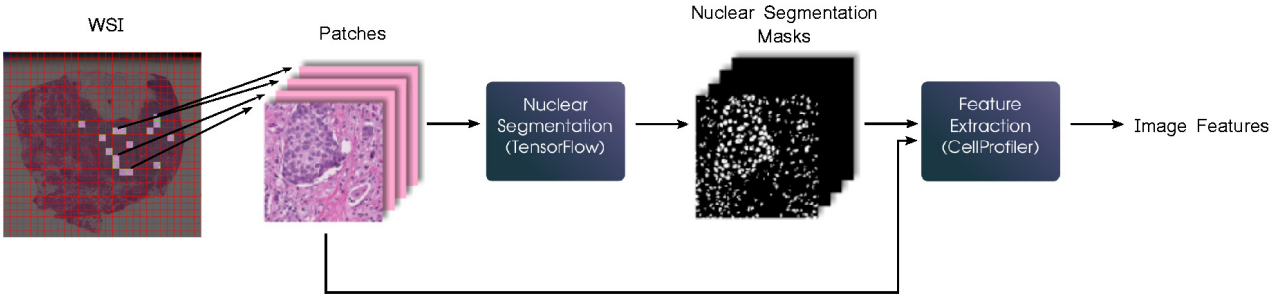


Fig. 1. Overview of our method for combined nuclear segmentation and feature extraction from WSI tiles.

commonly used available algorithms fail to generalize adequately to the significant variation of most large-scale histopathological image data sets. Recently, several methods for nuclear detection and segmentation have been developed using deep learning and convolutional neural networks (CNNs),[14–17] which have shown improvements over traditional methods. For our framework, we developed our own patch-based CNN for nuclear segmentation, which we optimized for computational efficiency and trained using additional image data from TCGA. Additionally, unlike CellProfiler, our patch-based CNN method requires no parameter tuning, which allows for easy analysis by non-specialists, which would help facilitate high-throughput imaging studies. The contribution of this work is to apply deep-learning-based nuclear segmentation in conjunction with nuclear feature extraction to show its utility for high-throughput histopathological image analysis, specifically for predicting genomic markers.

As a proof of principle, we applied our method to breast cancer patient samples from TCGA[4] to detect the basal subtype. For this task, the learned model achieved a class-balanced accuracy of 70% (i.e., sensitivity and specificity scores were both 0.7) on a separate test partition of 213 patients, of which 39 patients were of the basal subtype, which is a significant improvement over the standard method of summarizing images by the statistics of the distribution of cellular features. The algorithm also learned eight cell types relevant to the basal subtype. To our knowledge, this is the first work that attempts to predict molecular subtype solely from histopathological images. With improved nuclear segmentation and increased access to imaging-genomic data sets, DBC has the potential to be a useful tool for cancer screening of genomic markers.

2. Methods

2.1. *Overview*

To represent an H&E image as a histogram of nuclear words, we first extract nuclear and cellular features. These features are derived from the segmentation result of a patch-based CNN scheme. From the segmentation, CellProfiler computes cellular features of shape, texture, and color. The overall method for nuclear segmentation and feature extraction is shown in Fig. 1. With each cell in the image represented by its extracted feature vector, DBC jointly learns a cell-type assignment and a BoW classifier to predict a genomic marker of interest.

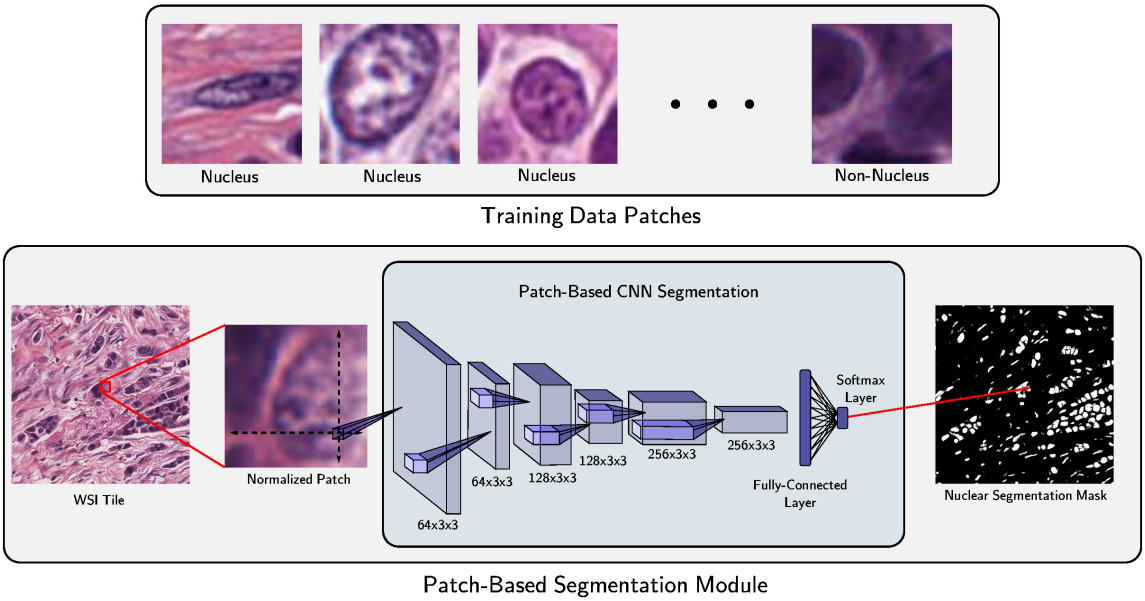


Fig. 2. Example nucleus and non-nucleus patches (top) from the training data for our CNN. Once trained, our CNN (bottom) classifies each local patch of a WSI tile individually, yielding a full nuclear segmentation mask for the tile.

2.2. *WSI Tiling*

Though DBC can be applied to any H&E image, whether it be a biopsy or a WSI, to process WSIs specifically, we divide them into non-overlapping tiles to be processed individually. From the WSIs, representative tiles of cancerous tissue regions that are free from artifacts, such as out-of-focus regions and tissue folds, are selected. Although this is performed mostly manually now, deep learning algorithms for segmenting cancerous regions of WSIs could be potentially incorporated.[18] For processing TCGA WSIs, we selected up to 15 representative tiles for each slide, with a tile size of 1000×1000 pixels.

2.3. *Nuclear Segmentation Module Based on CNN*

To achieve computationally efficient and reliable nuclear segmentation, we developed a patch-based CNN method that learns to detect nuclear pixels from the statistics of local patches. Similar to other recent patch-based CNN approaches,[17] the segmentation module produces a segmentation mask for an H&E image by generating, for each patch within the image, a binary label, indicating whether the center pixel of each patch belongs to a nucleus or not. Since nuclei only occupy a patch of a few pixels, the CNN needs to operate only on a small patch surrounding each pixel to produce its label. The training data for our network was imaged at 40× magnification, for which we chose the local patch size to be 51 × 51 pixels. Fig. 2 shows a diagram of the segmentation module operating on a WSI tile to produce a full nuclear segmentation mask.

Our network consists of six convolutional layers of $\{64, 64, 128, 128, 256, 256\}$ $3 \times 3 \times N$ filters, where N is the number of filters of the previous layer, and was implemented in TensorFlow.[19] At every other layer, starting at the second layer, the convolution operator is applied

at a stride of two to gradually taper the dimension of each subsequent layer towards the fully-connected layers near the output. The final two layers are a fully-connected layer of 50 nodes and a softmax output layer of two nodes, respectively. Before being fed through the network, each tile is unmixed into its hematoxylin and eosin stains and normalized[20] to mitigate stain variation and to reduce the last dimension N of the input layer filters from three to two.

For training our CNN, we used an available data set of segmented ER-positive epithelial nuclei,[17] supplemented with our own data set of nuclei patches from 68 TCGA-BRCA patients that included epithelial nuclei, stromal nuclei, and lymphocytes. For each TCGA-BRCA patient, we extracted several hundred sample patches of nuclei and non-nuclei, comprising 32,174 patches in total. The patients used in the data set represented a variety of tissue source sites to encourage the network to learn robustness to variation in sample acquisition. The pre-processing of each patch consisted of normalizing the stain,[20] separating the normalized hematoxylin and eosin images, and generating rotations at 90 degree increments, as well as horizontal and vertical flips of the images, to promote invariance to such manipulations, which naturally arise in H&E images. Several example patches collected for our data set are shown in Fig. 2.

Once the initial binary segmentation mask for each WSI tile is generated by the CNN, the mask and the corresponding H&E image are passed to CellProfiler to be further enhanced by smoothing the boundaries and separating clumped nuclei. Compared to relying solely upon CellProfiler for segmentation, the CNN-based method produces more reliable performance, as shown in Section 3.

2.4. *Nuclear and Cellular Feature Extraction*

After nuclear segmentation, we use CellProfiler to determine the boundaries of the cell corresponding to each nucleus. We consider only the pixels close to the nucleus boundary, but not overlapping with neighboring cells, as belonging to the cell. In our analysis, the distance we chose was 15 pixels. Features describing the shape, texture, and color of each nucleus and cell for each patient are then extracted through CellProfiler from the segmented masks and corresponding H&E images. From the nuclear and cellular segmentation boundaries computed with our CNN and CellProfiler's refinement steps, we extract a total of 219 image features for each cell-nucleus pair.

2.5. *Discriminative Bag-of-Cells*

Each image is then summarized by a histogram of the various cell types it contains, where the cell types are learned *discriminatively* for the specific genomic marker. We denote the extracted feature vector for a cell by $\mathbf{x} \in \mathbb{R}^d$, where d is the number of extracted features. For each sample, or patient, s, with $N_x(s)$ segmented cells, we extract a set of $N_x(s)$ cellular feature vectors $X_s = \{\mathbf{x}_i\}_{i=1}^{N_x(s)}$. We denote the genomic marker of interest by $y \in [0, 1, \ldots, K-1]$, where K is the number of possible states the marker can assume. Each patient then consists of a pair of cellular feature vectors and a marker: (X_s, y_s).

Our DBC framework consists of two learners: the cell-type assignment $f_x(\cdot)$ and the BoW classifier $f_b(\cdot)$. The cell-type assignment produces the cell-type representation $\mathbf{c}_i = f_x(\mathbf{x}_i) \in$

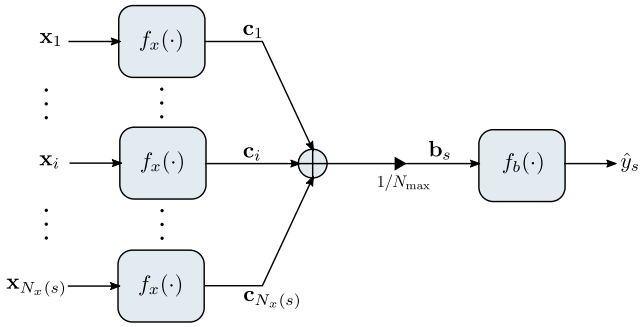


Fig. 3. Flow chart of DBC prediction on sample s.

$[0,1]^C$ (satisfying $\sum_j^C c_{i,j} = 1$), which acts like a soft one-hot-encoded vector assignment to one of C cell types, for each cell i. The cell-type representations are summed over all cells, producing the BoW representation $\mathbf{b}_s$, which is normalized by $N_{\max}$, the maximal number of cells associated with any sample. In practice, this was found to help the network to learn. The final output of the DBC is the predicted marker $\hat{y}_s = f_b(\mathbf{b}_s)$. The flowchart describing DBC is shown in Fig. 3.

In the standard BoW approach, the type, or codeword, assignment $f_x(\cdot)$ is learned independent of the subsequent BoW classifier, usually by a clustering algorithm such as K-means. The drawback of the standard approach is that the assignment to codewords may not be discriminative of the classification task. However, in DBC, by constructing both learners with neural networks, we can train the combined network end-to-end, allowing $f_x(\cdot)$ to be optimized for the discriminative task at hand. To enforce that the cell-type assignment $f_x(\cdot)$ act like a soft one-hot encoding, its output layer is a softmax layer of C nodes. The output layer of the BoW classifier $f_b(\cdot)$ is also a softmax layer, and the final predicted marker $\hat{y}$ is the index of the maximal value of the output softmax layer. The entire DBC framework was implemented in TensorFlow.[19]

3. Results

3.1. *Segmentation Evaluation*

We evaluated our patch-based CNN segmentation module on the UCSB Bio-Segmentation Benchmark.[21] The segmentations for this data set are pixel-wise binary masks of nuclei pixels. Boundaries between touching nuclei were not delineated on the masks, so in to order ensure objectivity and reproducibility, we inferred these boundaries automatically using CellProfiler's nuclei separation tool. Since the images were captured at a lower magnification, and not 40× magnification like the images for our training data, we resized them by a factor of 2 to approximate 40× magnification. We refer to the resulting separated nuclei masks as the gold standard. For our evaluation, a nucleus is considered a true positive if its center is within a distance of 12 pixels of the center of a nucleus in the gold standard mask. When a gold standard nucleus matches multiple predicted nuclei, its closest match is chosen.

The scores for our network, our CellProfiler pipeline, and the work of Janowczyk and Madabhushi[17] are shown in Table 1. The network of Ref. 17 produces a probability, which

Table 1. Comparative detection accuracy of our CNN, our CellProfiler pipeline, and the network Ref. 17 on UCSB breast cancer H&E images.

Algorithm	Precision	Recall	F1-Score
Our CNN	0.830	0.875	0.852
Janowczyk and Madabhushi[17]	0.850	0.850	0.850
CellProfiler	0.905	0.712	0.800

Table 2. Comparative segmentation accuracy of our CNN, our CellProfiler pipeline, and the network of Ref. 17 UCSB breast cancer H&E images.

Algorithm	DSC Mean	DSC Median	DSC STD	HD Mean	HD Median	HD STD	MAD Mean	MAD Median	MAD STD
Our CNN	0.72	0.80	0.20	4.24	3.00	3.83	1.82	1.26	1.43
Janowczyk and Madabhushi[17]	0.75	0.81	0.16	4.22	2.83	4.03	1.75	1.29	1.20
CellProfiler	0.76	0.82	0.18	4.62	2.83	4.70	1.84	1.24	1.64

allows for tuning by the user by varying the threshold to be applied to make a binary decision, so we evaluated thresholds ranging from 0 to 0.92 by increments of 0.04 and reported the the threshold with the best F1-score. It is expected that, assuming the training procedure can determine a good set of parameters, our network should perform comparably, if perhaps slightly better, than the network of Ref. 17, since we have an expanded data set with our own training patches from TCGA, and indeed, our algorithm performs comparably. What is of more importance is that compared to the results of CellProfiler, our approach shows a marked improvement.

To evaluate the quality of the resulting nuclear boundaries, we used the the Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean absolute distance (MAD), which were adopted from other works[16] for consistency and comparison. Since these metrics can only be applied to true positive nuclei, they do not capture the effects of false negatives and false positives, so a desired balance must be considered when comparing algorithms. A comparison of the results for each of these metrics is shown in Table 2. Again, our CNN and the network of Ref. 17 perform similarly. CellProfiler performs slightly better in terms of DSC, but worse in both HD and MAD, despite it having a significantly higher precision score, which indicates that it is more conservative in what it detects as nuclei. We observed on this data set that our patch-based CNN algorithm is able to detect and segment both stromal and epithelial nuclei, but that it struggles with nuclei with fainter hematoxylin stain, which could be a consequence of the difference in resolution between the 20×UCSB images and the 40×TCGA images on which it was trained.

Data sets such as TCGA pose a much greater challenge for segmentation since the acquisition procedures across the various tissue source cites are less controlled and prone to introduce significant variation. A prominent advantage of our network is that it has been trained on WSIs from TCGA-BRCA patients, increasing its robust to these variations. We

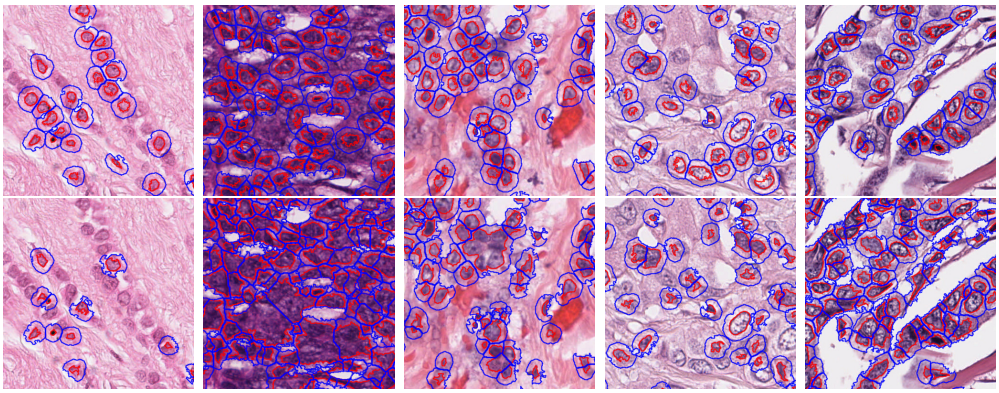


Fig. 4. Cell and nucleus segmentation results of our patch-based CNN approach (top row) and CellProfiler (bottom row) on several regions of diagnostic WSIs of the TCGA-BRCA data set. Nuclear boundaries are drawn in red and cellular boundaries in blue.

chose a small subset of 1000 × 1000 tiles from TCGA-BRCA WSIs with varying slide quality and hue on which to qualitatively compare our trained patch-based CNN approach and CellProfiler. The CellProfiler pipeline we tested consisted of adaptive thresholding to remove white background pixels, adaptive three class thresholding of the unmixed hematoxylin channel to segment nuclei, declumping of nuclei, and removal of overly large or small nuclei. We tuned the parameters to generalize the segmentation as best as possible across the test images, limiting the white pixel background threshold on the hematoxylin channel to lie between 0 to 0.3 and the adaptive three class threshold for nuclei segmentation to lie between 0.4 to 1. The yielded nuclear and cellular boundaries of both methods for regions of five images from the test set are shown in Fig. 4. Increasing the later threshold improved performance on darker images, but at the cost of missing most nuclei in lighter images. In contrast, our approach was able to detect and segment nuclei well despite the stark differences in intensity and hue. In particular, it was better able to avoid clumping large nuclei together, as seen in the two darker images. We also note that, unlike CellProfiler, our method required no parameter tuning for this test set, which is crucial for high-throughput analysis of large data sets of WSIs.

Computational cost is also an important consideration for high-throughput analysis. Segmenting each 1000×1000 tile using our approach required only an additional 5.7 seconds on average for processing with our patch-based CNN on a single Intel Xeon E5-1603 v3 (2.8GHz) CPU with 16GB of RAM and a single NVIDIA GeForce GTX 1080Ti GPU, which is a minimal overhead for the increased accuracy gained.

3.2. *Detecting BRCA Basal Subtype with Discriminative Bag-of-Cells*

We applied DBC on the extracted cellular features of 607 TCGA-BRCA patients to predict the status of the basal molecular subtype. The set of patients was split into 50% training, 15% validation, and 35% testing, which yielded a test set of 213 patients, of which 39 were of the basal subtype. To prevent overfitting to the training partition, during training, the accuracy of the model on the validation set was monitored and training was ended when this accuracy began to steadily decrease. Additionally, we removed all nuclear features with variance below

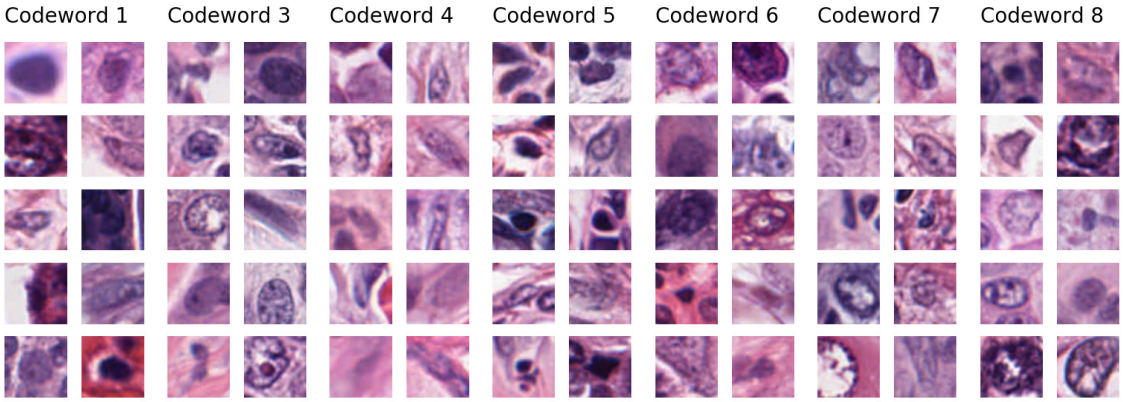


Fig. 5. Example nuclei from TCGA-BRCA patient WSIs for each of the eight learned codewords. Codeword two had no dominant example nuclei.

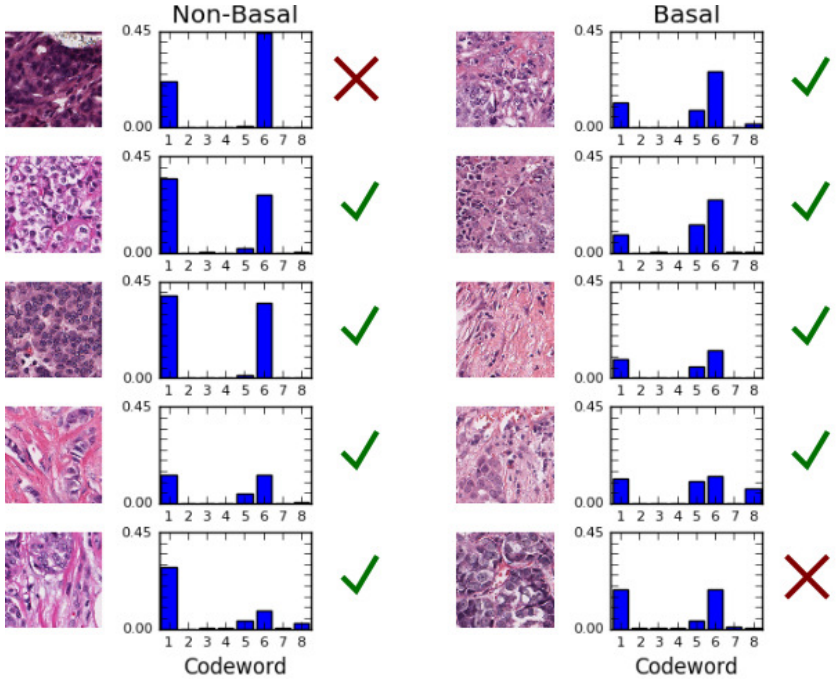


Fig. 6. BoW histograms for several example TCGA-BRCA patients. (left) Non-basal patients. (right) basal patients. Sample patches of the WSI for each patient are shown on the left of each of the two columns. The markings to the right of the BoW histograms indicate if the predicted subtype was correct.

0.001 across the training set to keep the model from fitting noise. For this particular prediction task, we trained a codeword assignment with two hidden layers of $\{25, 15\}$ nodes and a BoW classifier with just one hidden layer of six nodes. We found that a codebook of size $C = 8$ allowed for an accurate BoW classifier while minimizing overfitting. To help the model train, the nuclear features were normalized to a mean of 0.5 and a variance of 0.5.

Fig. 5 shows randomly selected example nuclei for the eight learned cell types by the codeword assignment of DBC. These are nuclei for which the maximal codeword index is the

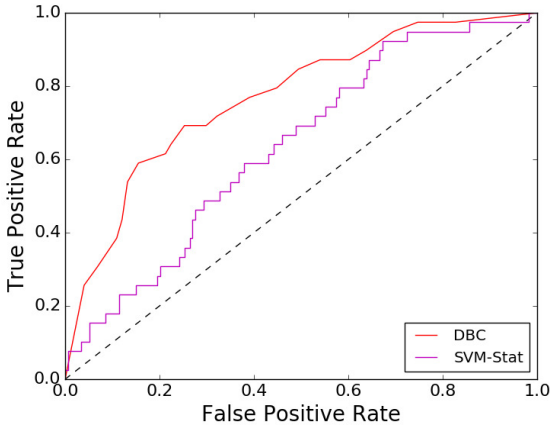


Fig. 7. ROC curves for predicting basal subtype (Positive = basal, Negative = other) on the test partition of TCGA-BRCA patients. SVM-Stat represents a standard approach to histopathological image summarization and classification, in which an image is summarized by statistics of the distribution of cellular features.

column in which they are shown, though since the codeword assignments are soft, some nuclei may be encoded as a mixture of several codewords. Patterns can be seen in the assignments, such as small, dark, lymphocytic cells and thin stromal cells, as well as the grouping of mostly hollow, highly textured nuclei. We found that codeword two never received the maximal mixture coefficient for any nuclei, which is why no examples are shown. With the knowledge of what types of cells are being mapped to which codewords, we can then investigate the BoW representations for WSIs to provide further insight. Fig. 6 shows BoW representations for the WSIs of several patients of basal and non-basal subtypes using the learned codeword assignment. The H&E images on the left of the two columns are randomly selected sample 500 $\times$ 500 pixel patches from the tiles of the WSIs of these patients which were used for feature extraction. The markings on the right indicate if the subtype of the patient was predicted correctly. A general trend for the basal subtype of higher counts of codeword five and lower counts of codewords one and six can be seen, though the true relationship lies in the weights of the learned neural network of the BoW classifier. We observe that most non-basal patients had a higher count of codeword one than codeword six. The upper left non-basal patient conversely had a higher count of codeword six and was predicted by the BoW classifier to be of the basal type. Similarly, the lower right basal patient had roughly equal counts of codewords one and six and a low count of codeword five and was predicted not to be of the basal type.

The output of our BoW classifier is a scalar value between $[0, 1]$, which can be interpreted as the probability of belonging to the basal subtype and which can be thresholded at varying levels to trade-off between the false and true positive rates. Varying this threshold, we generated the ROC curve of DBC on the separate partition of test patients shown in Fig. 7. From the ROC curve, we observe that our trained DBC is well-balanced between classifying basal and non-basal subtypes, achieving the reported class-balanced accuracy of 70% at a false positive rate of 0.3 and a true positive rate of 0.7. We compared the performance of DBC to a standard approach of summarizing histopathological images,[13] in which images are summa-

rized by statistics of the distribution of cellular features. For this comparison, we used the mean and variance, which became the feature vector for each patient to be fed into a classifier. We trained a SVM classifier with a RBF kernel using Scikit-learn[22] on this feature vector and called the overall method SVM-Stat. The ROC of this approach using the same training and testing sets is also shown in Fig. 7. DBC shows a significant improvement over this standard approach, which we attribute to the significant loss of information in summarizing all cellular features together.

4. Discussion

We have shown the potential for DBC as a screening tool of histopathological images for genomic markers. As a proof of principle, we showed the application of our approach for identifying the basal molecular subtype based on imaging features, though our method could easily be trained for other genomic markers, thereby learning new cell types relevant to that marker. The effectiveness of DBC stems from its flexibility in learning what types of cells are informative of the specific genomic marker and how best to jointly leverage the extracted cellular features to define these cell types. Our overall method requires only a few minutes per H&E image of 1000 $\times$1000 pixels to process, which is sufficiently fast for high-throughput image-genomic analysis if WSIs can first be efficiently pruned for representative tiles, which could be helped by the use of parallel GPU computing.

Furthermore, we believe several areas can be improved for our method. As a tumor progresses, the spatial relationship of cells changes, becoming increasingly disordered,[1] and so image features should consist not just of those of individual cells but also those describing their spatial relationship. Graphical features of cells have been shown to increase accuracy of automated non-small cell lung cancer subtype identification,[23] and these features could easily be appended to our BoW image representation to be used in the subsequent classification step. The DBC framework could also easily be extended hierarchically to learn words for local regions of nuclei in a spatial pyramid scheme[9] to capture cellular heterogeneity at various scales. Additionally, as more segmentation and imaging-genomic training data become available, the performance of DBC is expected to also increase. Nevertheless, we believe our approach has the potential to become a highly useful tool to connect imaging features with genomic signals to unravel the phenotypic impact of genomic alterations in cancer.

5. Acknowledgments

The authors would like to thank Jack P. Hou for helpful discussions about cancer genomics, Chang Hu for help in collecting nuclei training samples, and Sandhya Sarwate, M.D. for offering her professional pathologist expertise and consultation. We thank the TCGA Research Network for making the data publicly available.

References

1. D. Hanahan and R. A. Weinberg, *Cell* **144**, 646 (2011).
2. K. Polyak, *The Journal of Clinical Investigation* **121**, 3786 (2011).

3. C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. a. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. a. Akslen, O. Fluge, a. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, a. L. Børresen-Dale, P. O. Brown and D. Botstein, *Nature* **406**, 747 (2000).
4. TCGA Network, *Nature* **490**, 61 (2012).
5. H. Chang, G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, P. T. Spellman and B. Parvin, *BMC Bioinformatics* **12** (2011).
6. Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-f. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas and F. Markowetz, *Science Translational Medicine* **4**, 157ra143 (2012).
7. C. Wang, T. Pécot, D. L. Zynger, R. Machiraju, C. L. Shapiro and K. Huang, *Journal of the American Medical Informatics Association* **20**, 680 (2013).
8. L. A. D. Cooper, J. Kong, D. A. Gutman, W. D. Dunn, M. Nalisnik and D. J. Brat, *Laboratory Investigation* **95**, 366 (2015).
9. S. Lazebnik, C. Schmid and J. Ponce, Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
10. F. Moosmann, B. Triggs and F. Jurie, *Advances in Neural Information Processing Systems* , 985 (2007).
11. S. Lazebnik and M. Raginsky, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**, 1294 (2009).
12. A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland and D. M. Sabatini, *Genome Biology* **7**, p. R100 (2006).
13. D. L. Rubin, K.-h. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Re and M. Snyder, *Nature Communications* **7** (2016).
14. K. Sirinukunwattana, S. E. A. Raza, Y.-w. Tsang, D. R. J. Snead, I. A. Cree and N. M. Rajpoot, *IEEE Transactions on Medical Imaging* **35**, 1196 (2016).
15. J. Xu, L. Xiang, Q. Liu, S. Member, H. Gilmore, J. Wu, J. Tang and A. Madabhushi, *IEEE Transactions on Medical Imaging* **35**, 119 (2016).
16. F. Xing, Y. Xie and L. Yang, *IEEE Transactions on Medical Imaging* **35**, 550 (2016).
17. A. Janowczyk and A. Madabhushi, *Journal of Pathology Informatics* **7** (2016).
18. A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. C. Shih, J. Tomaszewski, F. A. González and A. Madabhushi, *Scientific Reports* **7**, p. 46450 (apr 2017).
19. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, TensorFlow: A System for Large-Scale Machine Learning, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, (USENIX Association, GA, 2016).
20. M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt and N. E. Thomas, A Method for Normalizing Histology Slides for Quanititative Analysis, in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009.
21. E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval and B. S. Manjunath, *BMC Bioinformatics* **10** (2009).
22. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
23. J. Yao, D. Ganti, X. Luo, G. Xiao, Y. Xie, S. Yan and J. Huang, Computer-Assisted Diagnosis of Lung Cancer Using Quantitative Topology Features, in *6th International Workshop on Machine Learning in Medical Imaging, MLMI'15*, 2015.

MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks

Lichy Han* and Maulik R. Kamdar*

Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA

E-mails: lhan2@stanford.edu, maulikrk@stanford.edu

Glioblastoma Multiforme (GBM), a malignant brain tumor, is among the most lethal of all cancers. Temozolomide is the primary chemotherapy treatment for patients diagnosed with GBM. The methylation status of the promoter or the enhancer regions of the O^6-methylguanine methyltransferase (MGMT) gene may impact the efficacy and sensitivity of temozolomide, and hence may affect overall patient survival. Microscopic genetic changes may manifest as macroscopic morphological changes in the brain tumors that can be detected using magnetic resonance imaging (MRI), which can serve as noninvasive biomarkers for determining methylation of MGMT regulatory regions. In this research, we use a compendium of brain MRI scans of GBM patients collected from The Cancer Imaging Archive (TCIA) combined with methylation data from The Cancer Genome Atlas (TCGA) to predict the methylation state of the MGMT regulatory regions in these patients. Our approach relies on a bi-directional convolutional recurrent neural network architecture (CRNN) that leverages the spatial aspects of these 3-dimensional MRI scans. Our CRNN obtains an accuracy of 67% on the validation data and 62% on the test data, with precision and recall both at 67%, suggesting the existence of MRI features that may complement existing markers for GBM patient stratification and prognosis. We have additionally presented our model via a novel neural network visualization platform, which we have developed to improve interpretability of deep learning MRI-based classification models.

Keywords: Deep learning; convolutional neural networks; MRI data; network visualization; Glioblastoma Multiforme

1. Introduction

Glioblastoma multiforme (GBM) is an aggressive brain cancer, with a median survival of only 15 months.[1] The efficacy of the first-line chemotherapy treatment, temozolomide, is in part dependent on the methylation status of the O^6-methylguanine methyltransferase (MGMT) regulatory regions (promoter and/or enhancer). MGMT removes alkyl groups from compounds and is one of the few known proteins in the DNA Direct Reversal Repair pathway.[2] Loss of the MGMT gene, or silencing of the gene through DNA methylation, may increase the carcinogenic risk after exposure to alkylating agents. Similarly, high levels of MGMT activity in cancer cells create a resistant phenotype by blunting the therapeutic effect of alkylating agents and may be an important determinant of treatment failure.[3] Thus, methylation of MGMT increases efficacy of alkylating agents such as temozolomide.[1]

As such, methylation status of MGMT regulatory regions has important prognostic im-

*These authors contributed equally to this work.

plications and can affect therapy selection in GBM. Currently, determining the methylation status is done using samples obtained from fine needle aspiration biopsies, which is an invasive procedure. However, several studies have demonstrated that some genetic changes can manifest as macroscopic changes, which can be detected using magnetic resonance imaging (MRI).[4,5] Previous approaches have constructed models to predict MGMT status from imaging and clinical data.[6,7] However, these models typically rely on hand curated features with classifiers such as SVM and random forests, and using neural networks may enable the discovery of novel biological features and increase the ease of implementation of such models.

Recently, convolutional neural networks (CNNs), a class of deep, feed-forward artificial neural networks, have emerged to be effective for autonomous feature extraction and have excelled at many image classification tasks.[8] A CNN consists of one or more convolutional layers, each layer composed of multiple filters. The architecture of a CNN captures different features (edges, shapes, texture, etc.) by leveraging the 2-dimensional spatial structure of an image using these filters. On the other hand, recurrent neural networks have shown a lot of promise to analyze ordered sequences of words or image frames, such as sentences or videos, for tasks such as machine translation, named entity recognition and classification.[9] Using fixed weight matrices (often termed, memory units) and vectorial representations for each sequence item (e.g. a word or a frame), an RNN can capture the temporal context in a dataset. While developing and implementing these neural network models may be inherently difficult, they can directly work on atomic features (e.g. pixels of an image, words in a sentence), and do not require exhaustive feature curation, as required in conventional machine learning methods.

Since MRI scans are 3-dimensional reconstruction of the human brain, they can be treated as volumetric objects or videos. Volumetric objects and sequences of image frames can be analyzed effectively by combining convolutional and recurrent neural networks.[10,11] However, few methods that combine CNNs with RNNs using end-to-end learning have been applied to radio-genomic analyses. Constructing an architecture that combines CNN and RNN for powerful image analysis while maintaining information transfer between image slices may reveal novel features that are associated with MGMT methylation.

In this work, **we present an approach using a bi-directional convolutional recurrent neural network (CRNN) architecture on brain MRI scans to predict the methylation status of MGMT**. We use a dataset of 5,235 brain MRI scans of 262 patients diagnosed with glioblastoma multiforme from The Cancer Imaging Archive (TCIA).[12,13] Genomics data corresponding to these patients is retrieved from The Cancer Genome Atlas (TCGA).[14] The CNN and RNN modules in the architecture are jointly trained in an end-to-end fashion. We evaluate our model using accuracy, precision, and recall. We also develop an interactive visualization platform to visualize the output of the convolutional layers in the trained CRNN network. The results of our study, as well as the visualizations of the MRI scans and the CRNN pipeline can be accessed at `http://onto-apps.stanford.edu/m3crnn/`.

1.1. *Deep learning methods over biomedical data*

Recently, several variations of deep learning architectures (neural networks, CNNs, RNNs, etc.) have been introduced for the analysis of imaging data, *-omics* data and biomedical lit-

erature.[15,16] A recent study by Akkus et al. has used CNNs to extract features from MRI images and predict chromosomal aberrations.[17] 2-dimensional and 3-dimensional CNN architectures have been used to determine the most discriminative clinical features and predict Alzheimer's disease using brain MRI scans.[10,18] Poudel et al. have developed a novel recurrent fully-connected CNN to learn image representations from cardiac MRI scans and leverage inter-slice spatial dependences through RNN memory units. The architecture combines anatomical detection and segmentation, and is trained end-to-end to reduce computational time.[11] For tumor segmentation, Stollenga et al. developed a novel architecture, PyramidLSTM, to parallelize multi-dimensional RNN memory units, and leverage the spatial-temporal context in brain MRI scans that is lost by conventional CNNs.[19] Chen et al. developed a transferred-RNN, which incorporates convolutional feature extractors and a temporal sequence learning model, to detect fetal standard plane from ultrasound videos. They implement end-to-end training and knowledge transfer between layers to deal with limited training data.[20] Kong et al. combined an RNN with a CNN, and designed a new loss function, to detect the end-diastole and end-systole frames in cardiac MRI scans.[21]

2. Methods

2.1. *Dataset and Features*

We used the brain MRI scans of glioblastoma multiforme (GBM) patients from The Cancer Imaging Archive (TCIA) and the methylation data, for those corresponding patients, from The Cancer Genome Atlas (TCGA).

2.1.1. *Preprocessing of Methylation Data*

We downloaded all methylation data files from GBM patients available via TCGA. The methylation consisted of 423 unique patients, with 16 patients having duplicate samples. We extracted methylation sites that are located in the minimal promoter and enhancer regions shown to have maximal methylation activity and affect MGMT expression.[22–24] Specifically, these methylation sites are *cg02941816*, *cg12434587*, and *cg12981137*. These are the same sites used in previous MGMT methylation studies that use TCGA data.[25] Similar to Alonso et al., we considered a methylation beta value of at least 0.2 to be a positive methylation site. As methylation of either the minimal promoter or the enhancer were shown to decrease transcription, we considered a patient to have a positive methylation status if any of the three sites were positive.

2.1.2. *Preprocessing of the MRI scans*

We downloaded 5,235 MRI scans for 262 patients diagnosed with GBM from TCIA. Each brain MRI scan can be envisioned as a 3-dimensional reconstruction of the brain (**Figure 1**). Each MRI scan consists of a set of image frames captured at a specific slice thickness and pixel spacing (based on the MRI machine specifications). The raw dataset contained a total of 458,951 image frames. From these, we selected 'labeled' *T1/T2/Flair axial* MRI scans for those patients for whom we had corresponding methylation data.

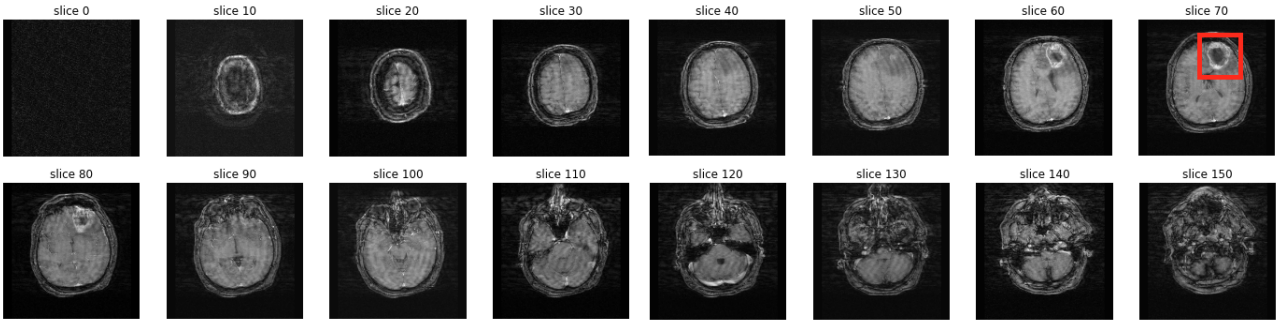


Fig. 1. **MRI scan.** A visualization of different MRI image frames in one MRI scan, with the GBM tumor highlighted in red on slice 70.

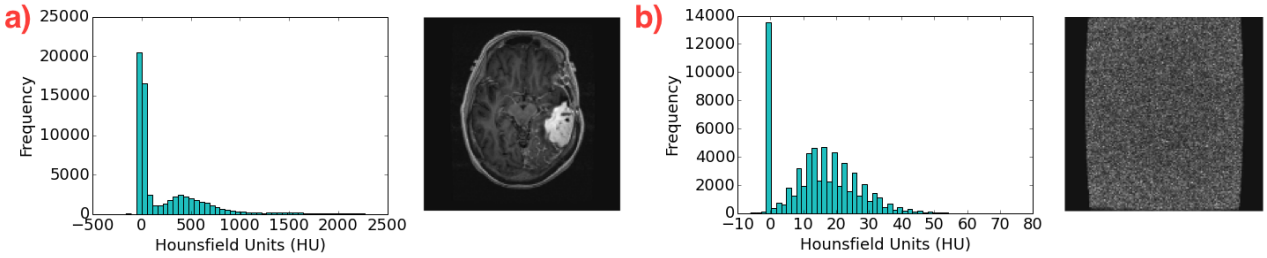


Fig. 2. **Removing noisy images.** We use the distributions of Hounsfield units (which vary drastically) to determine if an image is a valid MRI scan (**a**), or has only noisy pixels (**b**).

These image frames are made available in a DICOM format (Digital Imaging and Communications in Medicine), a non-proprietary data interchange protocol, digital image format, and file structure for biomedical images and image-related information.[26] The image frames are grayscale (1-channel) and the DICOM format allows storage of other patient-related metadata (sex, age, weight, etc.) as well as image-related metadata (slice thickness, pixel spacing etc.). As these image frames may be generated by different MRI machines with varying slice thickness (*range:* 1 to 10) and pixel spacing, we normalize these attributes across different MRI scans by resampling to a uniform slice thickness of 1.0 and pixel spacing of $[1, 1]$.

MRI image frames are grayscale, and instead of RGB channel values, each pixel is assigned a numerical value termed the Hounsfield Unit (HU), which is a measure of radiodensity. We filter out those image frames that are "noisy" by looking at the distribution of Hounsfield Units in the pixels. When removing noisy images, we used mean and standard deviation thresholds of 20 HU to determine image validity. An example of the distributions and the images are shown in **Figure 2**. We further limit our MRI scans to only those slices that contain the tumor to the nearest 10^{th} slice. This was achieved by annotating the MRI scans through our visualization platform[b]. Finally, we resize all images to 128×128 dimensions.

[b] `http://onto-apps.stanford.edu/m3crnn/`

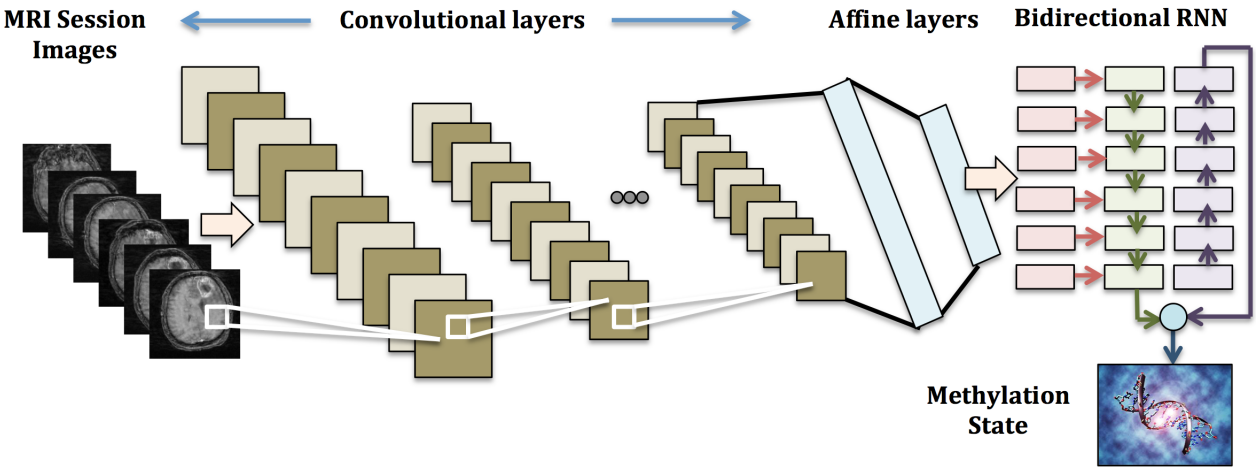


Fig. 3. **CRNN Architecture Overview.** Combining CNN and RNN to predict the methylation state from MRI scan images.

2.2. *Data Augmentation*

For our CRNN, we used data augmentation to increase the size of our dataset and to help combat overfitting. Specifically, we applied image rotation and MRI scan reversal, so that the methylation status and location of the tumor is preserved. Images were rotated every 4 degrees from -90 to +90 degrees, and were flipped such that in the RNN, the MRI scans were represented from superior to inferior and vice versa. This resulted in a 90 fold increase in the number of MRI scans.

2.3. *Training and Evaluation*

Given that our MRI scans are similar to video objects with a variable number of frames, we implemented a bi-directional convolutional recurrent neural network (CRNN) architecture (**Figure 3**). Each image frame of the MRI scan is first input into a CNN. Multiple convolutional layers extract essential features (e.g. shape, edges, etc.) from the image. The image is then processed through two fully connected neural network layers, so that the output from each image is a vector of length 512. All frames from one MRI scan are then represented by a series of vectors, which are input into a many-to-one bi-directional RNN. The bi-directional RNN is dynamic and can adjust for variable-length sequences, an advantage over using 3-dimensional CNN, which requires uniform volumes. Padding and bucketing of MRI scans of similar length was carried out for efficient computation. The RNN analyzes the sequence of MRI image frames and outputs a binary classification of methylation status per MRI scan. The entire architecture was developed using the Tensorflow Python library[c].

We split our the MRI scans into a 70% training set, 15% validation set, and 15% test set. As MRI scans of the same patient are highly correlated, we split our data such that all MRI scans pertaining to each patient are in the same set. We randomized the order of the training

[c]`https://www.tensorflow.org/`

Table 1. **Bi-directional CRNN Architecture.** Convolutional layers followed by fully connected layers and a many-to-one bi-directional RNN.

Layers	Hyperparmaeters
[5x5 Conv-ReLU-BatchNorm-Dropout-2x2 Max Pool] x 2 [5x5 Conv-ReLU-BatchNorm-Dropout] x 1 [5x5 Conv-ReLU-BatchNorm-Dropout-2x2 Max Pool] x 1	L2 Regularization: 0.05 Dropout Keep Probability: 0.9 Number of Filters: 8
FC-ReLU-BatchNorm-Dropout	Number of Neurons: 1024 L2 Regularization: 0.05 Dropout Keep Probability: 0.9
FC-ReLU-BatchNorm-Dropout	Number of Neurons: 512 L2 Regularization: 0.05 Dropout Keep Probability: 0.9
Bi-directional GRU with ReLU-Dropout	State Size: 256 L2 Regularization: 0.05 Dropout Keep Probability: 0.9
FC-ReLU	Number of Neurons: 256 L2 Regularization: 0.05 Dropout Keep Probability: 0.9
Softmax	

data based on the number of frames, bucketing MRI scans with similar frame numbers. We padded MRI scans within each bucket so all MRI scans in each batch had the same number of frames, while the number of frames differed across batches. We trained using softmax cross entropy as our loss function using the Adam optimizer with learning rates ranging from 5e-6 to 5e-1. We applied L2 regularization, with coefficients from 0.001 to 0.1 and dropout with keep probabilities ranging from 0.5 to 1. We varied the number of filters between 8 and 16, and trained our model until it converged, for ten epochs.

For comparison, we also implemented a random forest classifier, to evaluate how our CRNN performs in comparison to alternative, more conventional machine learning algorithms that do not capture spatial information. For our random forest classifier, each frame was considered one sample, where each pixel was one feature. Each MRI scan was treated as an ensemble of individual frames, where we averaged the prediction across all frames for each scan.

When assessing our results, we calculated the area under the receiver operator characteristic curve (AUC), accuracy, precision, and recall at the patient and MRI scan levels. We calculated methylation status probability as the proportion of positive individual MRIs. Out of these metrics, we used patient level accuracy in the validation set to tune our architecture and hyperparameters. The CRNN was then evaluated using the independent test set.

3. Results

3.1. *Data Statistics*

Our training dataset consisted of 344 positive MRI scans and 351 negative scans, which corresponded to 117 patients. Our validation dataset consisted of 21 patients, with 73 positive scans and 62 negative scans. Our test set also had 21 patients, with 62 positive and 62 negative scans. After data augmentation, this resulted in 62,550 examples in the training set, and

Table 2. **CRNN Performance Metrics** for test, validation, and training sets at the patient and MRI scan level.

Set	Level	AUC	Accuracy	Precision	Recall
Test	**Patient**	**0.61**	**0.62**	**0.67**	**0.67**
	MRI Scan	0.73	0.63	0.72	0.42
Validation	**Patient**	**0.66**	**0.67**	**0.67**	**0.73**
	MRI Scan	0.54	0.53	0.57	0.55

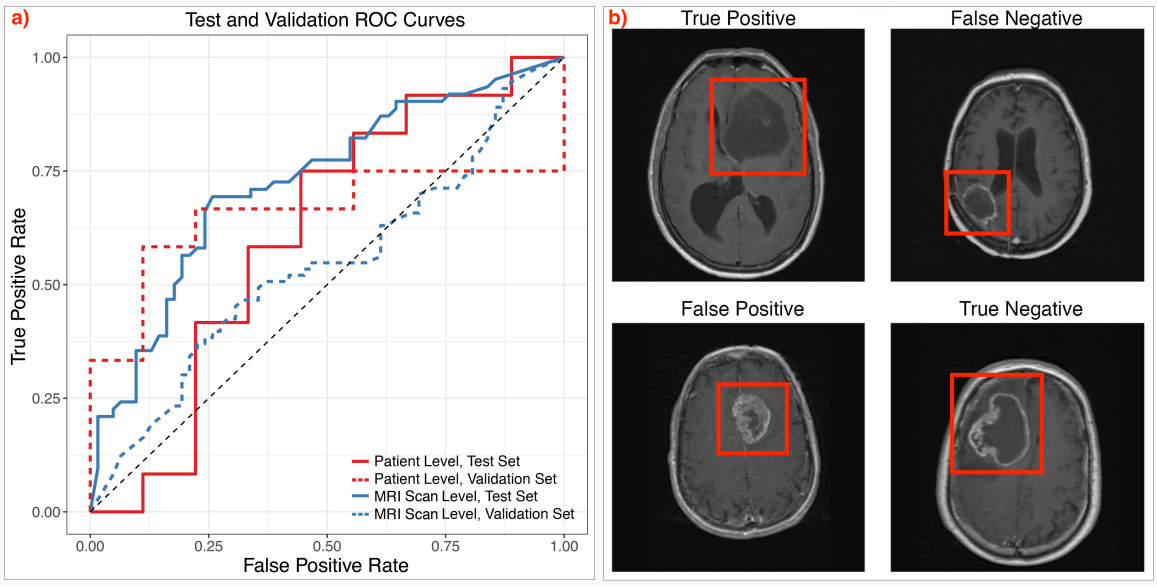


Fig. 4. **Evaluation of the CRNN method: a) ROC curves** depicting results at the patient and MRI scan levels in the validation and held-out test set, and **b) Classifier prediction examples.** True positive, true negative, and misclassified false positive and false negative examples from our test set. The tumors are highlighted in the red boxes.

12,150 in the validation set, and 11,160 in the test set. After preprocessing, we had an average of 45.9 frames per scan in the training set, 52.7 frames per scan in the validation set, and 43.2 frames per scan in the test set.

3.2. *Architecture and Hyperparameters*

The specific architecture of our CRNN is detailed in **Table 1**. Our architecture consisted mainly of alternating convolutional and pooling layers, using the rectified linear unit (ReLU) as our activation function. We used batch normalization, and we implemented L2 regularization and drop out layers to limit overfitting. We then followed these layers with fully connected (FC) layers to create the output for the RNN, which contained 512 neurons. We implemented a bi-directional RNN with gated recurrent units (GRU), with a state size of 256. We then followed the RNN with an additional FC layer before using the softmax classifier to predict methylation status. We used the Adam optimizer, with a learning rate of 1e-5.

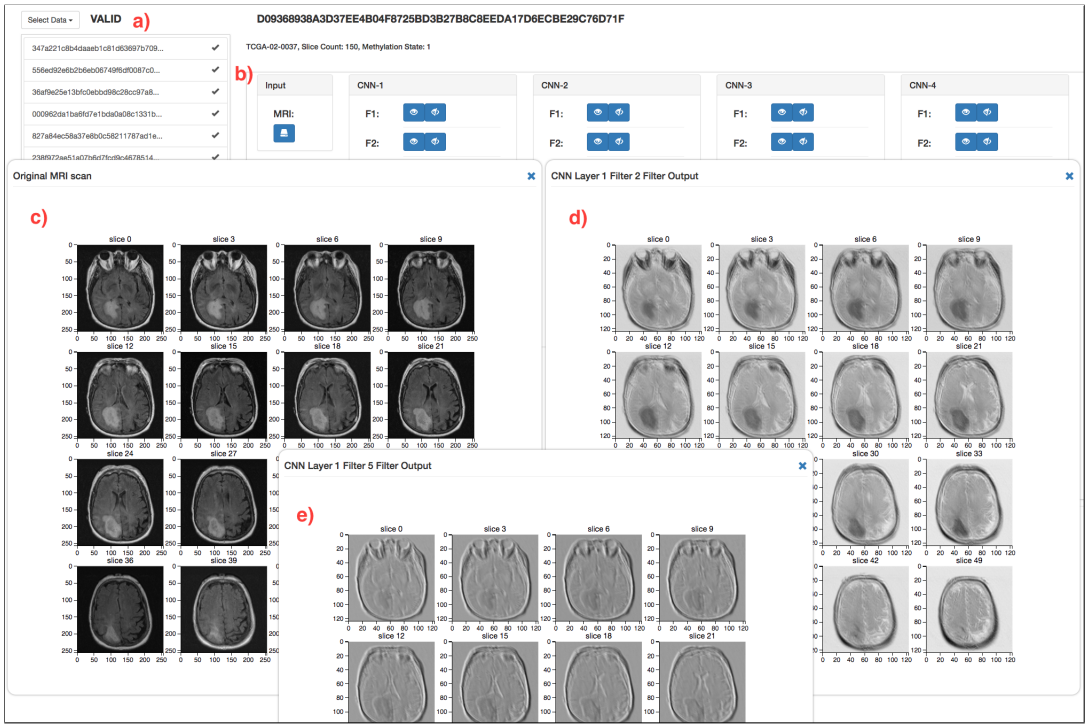


Fig. 5. **CRNN Visualization Interface. (a)** The domain user can select any MRI scan to load into the Tensorflow CRNN pipeline. **(b)** After the pipeline completes the computation to predict the MGMT methylation status, the user can visualize the original MRI Scan **(c)**, the output from any filter, in each convolutional layer **(d, e)**, as well as the output after ReLU activation.

3.3. *Evaluation*

For CRNN, our test set results are shown in **Table 2**. At the patient level, the test data yielded an accuracy of 0.61, with a precision of 0.67 and recall of 0.67. ROC curves are shown at the MRI scan and the patient level in **Figure 4a**. The training data obtained accuracies of 0.97 for MRI scans and at the patient level. Though we observe overfitting, increasing the dropout probability, increasing the L2 regularization coefficient, and decreasing model complexity did not result in significant gains in validation accuracy during model tuning. In comparison, our random forest classifier achieved an AUC of 0.56 on the validation set and 0.44 on the test set at the patient level.

We examined our classifier predictions in the test set, and show examples of true and false positives and negatives in **Figure 4b**. In particular, it appears that our classifier tends to classify lesions with ring enhancement as having a negative methylation status, and tumors with less clearly defined borders as positive. Predicted positive tumors also tended to have a more heterogeneous texture in appearance. Tumor location varied, and did not appear to be correlated with methylation status prediction.

3.4. *Visualization*

Deep learning methods, especially convolutional and recurrent neural networks, are thought to be less interpretable and clinically reliable, as compared to standard machine learning models.

To provide a more visual perspective on how our model perceives the input MRI scan, we have developed an interactive, online visualization interface deployed at `http://onto-apps.stanford.edu/m3crnn/`. The domain user (e.g. a radiologist or a biomedical researcher) can select an MRI scan from a list and load it through the pre-trained CRNN pipeline (**Figure 5a**). Once the pipeline completes the computation, the user can visualize the original MRI scan, click on each filter in each CRNN layer to see the output from each filter in each convolutional layer (**Figure 5b**). The user can also visualize the output after applying the ReLU activation function. Each visualization (either MRI scan, filter output or ReLU output) opens up in its own separate dialog window that can be dragged around the browser. Hence, multiple visualizations can be compared with each other (**Figure 5c-e**). Finally, the predicted output, the probability score as well as the actual methylation status, are also presented for the domain user to determine features and flaws of our model.

The output from two filters in the first convolutional layer are visualized (**Figure 5d,e**). As with many CNN architectures, the first layer places a heavy emphasis on edge detection, and we can clearly see the outline of the cranium and the tumor in each of these filters. Each filter also appears to show the brain slice at different contrasts. As specific tissues attenuate signal differently, in some sense these filters may be attempting to highlight different tissue types by varying the contrast. To the best of our knowledge, this is the first example of an online, interactive interface that can execute a deep learning pipeline over any selected MRI scan and can visualize intermediate layer outputs. It is very flexible, in the sense that the interface can easily be configured for variable number of convolutional layers and filters.

4. Discussion

In this work, we constructed a jointly trained, bi-directional convolutional recurrent neural network in order to predict the methylation status of MGMT from brain MRI scans. We explore macroscopic MRI features that may be correlated with MGMT methylation status to gain insight into GBM pathology. We use the publicly available data in TCGA and TCIA, where few studies, if any, have combined imaging data with *-omics* data using a deep learning framework. In addition, we present a generalizable platform for visualizing the different filters and layers of deep learning architectures for brain MRI scans to aid model interpretability for clinicians and biomedical researchers.

Our CRNN obtains modest patient level accuracies of 0.67 and 0.62 on the validation and test data, respectively, and on the test data, the precision and recall were both 0.67. Our data contained approximately equal proportions of positive and negative patients, indicating that our classifier is making predictions to balance precision and recall, and not relying on label distributions. Though the patient level performance does decrease from the validation to the test data set, the general similarity in performance indicates there are likely a subset of features that are correlated with MGMT methylation, as has been found in previous studies.[27,28] In comparison, the random forest model had an AUC of 0.57 in the validation set and 0.44 in the test set (versus CRNN with a validation AUC of 0.66 and test AUC of 0.61). This suggests that there is some useful information encoded in the individual pixels, but that reproducibility and performance are likely improved by using a method that can better capture spatial information.

We focused primarily on patient level results, leveraging multiple MRI scans per patient to obtain a prediction in an ensemble style. We secondarily assessed MRI scan results, as being able to predict methylation status from a single MRI scan would be highly relevant to clinicians and patients. The results at the MRI scan level were comparable to the patient level in the test set, but we see a decrease in performance in the validation set. This is likely due to our classifier being less confident at the MRI scan level, resulting in greater variability in results and prediction probabilities further from 0 or 1.

The difference in confidence between the patient level and MRI scan results suggests that combining information from multiple MRI scans is beneficial for MGMT methylation prediction. Deep learning models have been able to successfully learn multiple representations of the same object in other classification tasks.[8,9] However, we believe that combining different representations of the same tumor to reach a prediction per patient is more robust and clinically relevant. We accomplish this using majority voting. Incorporating additional layers into our model to combine MRI scans may also lead to further improvement in performance.

With a training set accuracy of nearly 1.0, our classifier is overfitted to the training set. To combat overfitting, we implemented L2 regularization, dropout layers, and data augmentation. Regularization had only a modest effect at curtailing overfitting and improving performance, and further increases in regularization resulted in decreasing validation set performance. Even though data augmentation was able to greatly decrease the speed of model overfitting, we still reach nearly perfect classification given enough training epochs. Data augmentation also substantially increased the number and variability of images for training, improving the robustness and performance of our model. However, due to the limited availability of publicly accessable patient data with both imaging and *-omics* measurements, our overall dataset of 159 patients can still be considered to be very small. The incorporation of additional patient data holds potential for further reduction of model variance and overfitting.

Currently, methylation status is not readily discernible by a human radiologist from MRI scans, even though multiple previous studies have attempted to correlate features to discover imaging-based biomarkers.[6,27–29] These studies typically require extensive manual feature curation, and may incorporate clinical data along with imaging features for classification. In comparison, our work is primarily focused on using raw MRI frames, which combines feature extraction and classification as one problem. Though we do manually annotate subsections of each MRI, we note that our method can work on full MRIs, and thus has the potential to be completely automated. While the validation accuracy of full MRI scans is simiar to the results in **Table 2**, training the CRNN on full scans requires additional computational time and resources. Additionally, though we have formulated our prediction task as binary classification, it is possible to use regression with CRNNs to predict methylation activity, which may be more informative. As we are interested in discovering MRI features independent of demographic or patient characteristics, we chose not to incorporate additional clinical data (e.g. age of onset or sex). However, these clinical data may provide additional signal from a classification standpoint.

When assessing our classifier predictions, our model had a tendency to assign positive methylation status to heterogeneous, larger tumors with poorly defined margins (**Figure 4b**).

Furthermore, many of our classification predictions are in concordance with previous results from Drabeyz et al.[30] and Eoli et al..[31] These studies discovered that ring-enhanced lesions were associated with negative MGMT promoter methylation. Hence, our model is able to autonomously determine some clinically relevant features correlated to MGMT methylation, without manual curation or predefined feature engineering as required in previous methods.

Deep learning methods have become powerful tools in image analysis and in the biomedical domain.[15,16] However, these methods typically are not easily interpretable, and it can be challenging for a clinician or researcher to understand the model's reasoning. Hence, these methods are often infamously termed *"black-box models"*. To address this challenge, we have developed a visualization platform that allows the domain user to select each MRI scan, load it through the CRNN computational pipeline, and interactively view and compare different filters and layers of our model. Our platform is generalizable, and can be easily extended for use with additional MRI prediction tasks and with different model architectures (e.g. variational number of filters and convolutional layers). For example, though we are primarily focused on the GBM tumor and its MGMT methylation status in this work, one may visualize whole brain MRI scans, in different orientations (e.g. saggital), for tasks such as risk stratification or lesion diagnosis. Moreover, any similar deep learning pipeline, that may use other type of MRI scans (e.g. cardiac) or other volumetric biomedical data (e.g. ultrasound), can be deployed with ease. Through the platform, we also visualize all the classifier predictions for our test set, and group them into four distinct sets — true and false positives and negatives. The domain users can browse and capture additional clinical features used by our model for prediction, or flaws in our model, that we may have not discussed here. We envision the visualization platform to be used in other relevant research and hence, we have released the source code.[d]

5. Conclusions

In this work, we implemented a convolutional recurrent neural network (CRNN) architecture to predict MGMT regulator methylation status using axial brain MRI scans from glioblastoma multiforme patients. Based on this model, we constructed a generalizable visualization platform for exploring the filtered outputs of different layers of our model architecture. Our CRNN achieved a test set accuracy of 0.62, with a precision of 0.67 and recall of 0.67. Using our predictions, we highlight macroscopic features of tumor morphology which may provide additional insight into the effects of MGMT methylation in glioblastoma multiforme. Though modest, our results support the existence of an association between MGMT methylation status and tumor characteristics, which merits further investigation using a larger cohort.

6. Acknowledgements and Funding

We would like to thank Fei-Fei Li, Justin Thompson, Serena Yeung and other staff members of the Stanford CS231N course (Convolutional Neural Networks for Visual Recognition) for their constructive feedback on this project. This work used the XStream computational resource, supported by the NSF Major Research Instrumentation program (ACI-1429830). LH is funded

[d] `https://github.com/maulikkamdar/M3CRNN`

by NIH F30 AI124553. The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. We dedicate this work to the memory of Rajendra N. Kamdar, father of Maulik R. Kamdar, who passed away during the course of this research.

References

1. J. P. Thakkar, T. A. Dolecek, C. Horbinski, Q. T. Ostrom, D. D. Lightner, J. S. Barnholtz-Sloan and J. L. Villano, *Cancer Epidemiology and Prevention Biomarkers* **23** (2014).
2. G. P. Margison, A. C. Povey, B. Kaina *et al.*, *Carcinogenesis* **24**, 625 (apr 2003).
3. M. E. Hegi, A.-C. Diserens *et al.*, *New England Journal of Medicine* **352**, 997 (mar 2005).
4. B. M. Ellingson, *Current Neurology and Neuroscience Reports* **15**, p. 506 (jan 2015).
5. S. Yamamoto, D. D. Maki *et al.*, *American Journal of Roentgenology* **199**, 654 (sep 2012).
6. P. Korfiatis, T. L. Kline, L. Coufalova, D. H. Lachance, I. F. Parney, R. E. Carter, J. C. Buckner and B. J. Erickson, *Medical Physics* **43**, 2835 (may 2016).
7. I. Levner, S. Drabycz, G. Roldan *et al.*, *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention* , 522 (2009).
8. A. Krizhevsky *et al.*, *Advances in neural information processing systems* , 1097 (2012).
9. J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, *Proceedings of the IEEE conference on computer vision and pattern recognition* , 4694 (2015).
10. A. Payan and G. Montana, *arXiv* (feb 2015).
11. R. P. K. Poudel, P. Lamata and G. Montana, *arXiv* (aug 2016).
12. L. Scarpace, T. Mikkelsen, S. Cha *et al.*, *The Cancer Imaging Archive* (2016).
13. K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox and F. Prior, *Journal of Digital Imaging* **26**, 1045 (dec 2013).
14. J. N. Weinstein, E. A. Collisson, G. B. Mills *et al.*, *Nature Publishing Group* **45** (2013).
15. G. Litjens, T. Kooi, B. E. Bejnordi *et al.*, *arXiv* (feb 2017).
16. S. Min, B. Lee and S. Yoon, *Briefings in Bioinformatics* , p. bbw068 (jul 2016).
17. Z. Akkus, I. Ali, J. Sedlar *et al.*, *arXiv* (nov 2016).
18. S. Sarraf and G. Tofighi, *arXiv* (mar 2016).
19. M. F. Stollenga, W. Byeon, M. Liwicki and J. Schmidhuber, *arXiv* (jun 2015).
20. H. Chen, Q. Dou, D. Ni, J.-Z. Cheng *et al.*, *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention* , 507 (2015).
21. B. Kong, Y. Zhan, M. Shin, T. Denny and S. Zhang, *Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention* , 264 (2016).
22. L. C. Harris, J. S. Remack and T. P. Brent, *Nucleic Acids Research* **22**, 4614 (1994).
23. L. C. Harris, P. M. Potter, K. Tano *et al.*, *Nucleic Acids Research* **19**, 6163 (1991).
24. T. Nakagawachi, H. Soejima, T. Urano *et al.*, *Oncogene* **22**, 8835 (2003).
25. S. Alonso, Y. Dai, K. Yamashita *et al.*, *Oncotarget* **6**, 3420 (2015).
26. P. Mildenberger, M. Eichelberg and E. Martin, *European Radiology* **12**, 920 (apr 2002).
27. W.-J. Moon, J. W. Choi, H. G. Roh, S. D. Lim and Y.-C. Koh, *Neuroradiology* **54**, 555 (2012).
28. A. Gupta, A. M. P. Omuro, A. D. Shah, J. J. Graber, W. Shi, Z. Zhang and R. J. Young, *Neuroradiology* **54**, 641 (jun 2012).
29. V. G. Kanas, E. I. Zacharaki, G. A. Thomas, P. O. Zinn, V. Megalooikonomou and R. R. Colen, *Computer Methods and Programs in Biomedicine* **140**, 249 (2017).
30. S. Drabycz, G. Roldán, P. de Robles, D. Adler, J. B. McIntyre, A. M. Magliocco, J. G. Cairncross and J. R. Mitchell, *NeuroImage* **49**, 1398 (2010).
31. M. Eoli, F. Menghi, M. G. Bruzzone, T. De Simone, L. Valletta, B. Pollo, L. Bissola, A. Silvani, D. Bianchessi, L. D'Incerti, G. Filippini *et al.*, *Clinical Cancer Research* **13**, 2606 (2007).

Deep Integrative Analysis for Survival Prediction

Chenglong Huang, Albert Zhang and Guanghua Xiao
Colleyville Heritage High School, Colleyville, TX, 76034, USA
Highland Park High School, Dallas, TX, 75205, USA
Department of Clinical Science, The University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

Survival prediction is very important in medical treatment. However, recent leading research is challenged by two factors: 1) the datasets usually come with multi-modality; and 2) sample sizes are relatively small. To solve the above challenges, we developed a deep survival learning model to predict patients' survival outcomes by integrating multi-view data. The proposed network contains two sub-networks, one view-specific and one common sub-network. We designated one CNN-based and one FCN-based sub-network to efficiently handle pathological images and molecular profiles, respectively. Our model first explicitly maximizes the correlation among the views and then transfers feature hierarchies from view commonality and specifically fine-tunes on the survival prediction task. We evaluate our method on real lung and brain tumor data sets to demonstrate the effectiveness of the proposed model using data with multiple modalities across different tumor types.

Keywords: Survival Prediction, Integrative Analysis, Deep Learning

1. Introduction

Survival analysis aims at modeling the time that will elapse from the present to the occurrence of a certain event of interest (e.g. biological death). The prognostic models generated by survival analysis can be used to explore interactions between prognostic factors in certain diseases, and also predict how a new patient will behave in the context of known data. In survival analysis, the Cox proportional hazards model[1] and parametric survival distributions[2] have long been used as important fundamental techniques. Clinicians and researchers usually apply these models to test for significant risk factors affecting survival. In order to handle the high-dimensional data, dimension reduction and penalized regression have been proposed in the Cox model.[3–7] However, the Cox model and its extensions are still built based on the assumption that a patient's risk is a linear combination of covariates. The parametric censored regression approaches[2,8] are highly dependent on the choice of the distribution. In fact, there are too many complex interactions that can affect the event (death) in various ways, and thus a more comprehensive survival model is needed to better fit data in real-world applications. To formulate the survival problem without any additional hypothesis, Li et al. modeled the prediction problem as standard multi-task learning using an additional indicator matrix.[9] However, the number of tasks corresponds to the maximum follow-up time of all the instances. In fact, recent cancer datasets are collecting patient electronic health records (EHR) with a very long follow-up time. Another limitation for existing survival models is that they mainly focus on one view and cannot efficiently handle multi-modalities data. Since more comprehensive multi-source data are available to health-care research, a powerful survival

analysis that can learn from those multi-view data is required.

One good way to learn highly complex survival functions is by using recent neural network techniques.[10–12] Katzman *et al.* proposed a deep fully-connected network (DeepSurv) to represent the nonlinear risk function.[10] They demonstrated that DeepSurv outperformed the standard linear Cox proportional hazard model. However, DeepSurv is still too simple to handle real cancer data. First, real datasets contain complex imaging and genomic data from different views. Although using multiple pieces of information can provide complementary characterizations of tumors at different levels, the view discrepancy and heterogeneity will bring challenges for survival prediction. Second, compared to computer vision applications, survival prediction problems only provide a very small training set due to the cost of multiple comprehensive data collections. To integrate multiple modalities and eliminate view variations, a good solution is to learn a joint embedding space in which different modalities can be compared directly. Such an embedding space will benefit the survival analysis since recent studies have suggested that common representation from different modalities provides important information for prognosis.[13,14] For example, molecular profiling data and pathological images actually share representations to describe the same event in tumor growth, which is very important for diagnosis. Stromal tissue has been verified to have a surprising role in predicting the overall survival of breast cancer patients.[13] The proportion of stromal cells correlated with the overexpression of genes, including FBLN1, FBLN2, COL6A2 and COL6A3, that encode extracellular matrix proteins .[14]

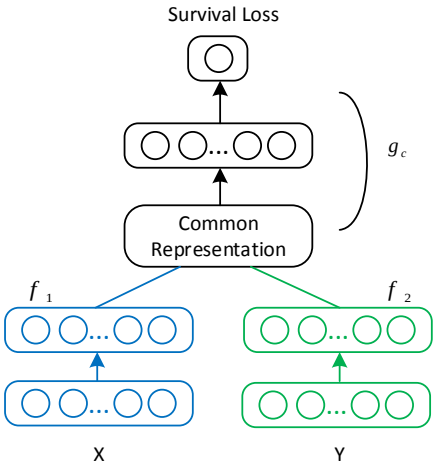


Fig. 1. An overview of the proposed model.

In order to take the advantage of both histopathological information and molecular profiles from imaging-genetics data, we developed an integrative pipeline as shown in Fig.1. It consists of two sub-networks, view-specific sub-network f_1, f_2 and common sub-network g_c. The view-specific sub-networks reduce the discrepancy between the view and the commonality of all views. The common sub-network is shared by all views and can extract a view-invariant representation for survival prediction. One advantage of the proposed architecture is that it

has good generality, since the network can handle any kind of data sources with well-designed view-specific sub-networks. Motivated by recent deep learning methods,[15,16] we proposed Convolutional Neural Networks (CNNs) and Fully Connected Networks (FCNs) to learn deep representations from pathological images and molecular profiling data, respectively.

To handle multi-modalities data, we integrate outputs of two networks into a common space where the two modalities have maximal correlation. The primary motivation of using such a model is to eliminate the view variations and find the maximum correlated representation from the views of pathological images and molecular data. Although the commonality of two views reduces the view-discrepancy, it still cannot guarantee that the common space is directly associated with survival outcomes. To address this issue, the proposed model transfers feature hierarchies from such common spaces and specifically fine-tunes on the survival regression task. This will not only help to avoid over-fitting, but also accelerates the model training. Moreover, it has the ability to discover important markers that cannot be found by previous deep correlational learning methods, which will benefit the survival prediction. The contribution of this paper can be summarized as: 1) We proposed a deep learning approach which can model very complex view distributions and learn good estimators for predicting patients' survival outcomes with insufficient training samples. 2) Our model used CNNs to represent much more abstract features from pathological images for survival prediction. Traditional survival models usually adopted hand-crafted imaging features. 3) Extensive experiments on TCGA-LUSC and GBM demonstrate that the proposed model can achieve better predictions across different tumor types.

2. Related Work

In this section, we give a brief survey on recent survival analysis methods with basic notations and then briefly review recent deep multi-modal embeddings.

2.1. *Survival Analysis*

Survival analysis aims to analyze the expected duration of time until events happen. It covers many topics as the event can be defined very broadly such as failure in mechanical systems and death in biological organisms. Survival analysis tries to find the answer of questions like: how does the proportion of a population survive past a certain time (e.g. 5 years)? what rate will they die or fail? Given a set of N patients, $\{x_i\}, i = 1 \dots N$, each patient has the label (t_i, δ_i) indicating the survival status where t_i is the observed time, δ_i is the indicator: 1 is for a uncensored instance (death event happens during the study), and 0 is for a censored instance (death not observed). If and only if $t_i = \min(O_i, C_i)$ can be observed during the study, the dataset is said to be right-censored.[17]

In Survival Analysis, the survival function $S(t|\mathbf{x}) = Pr(O \geq t|\mathbf{x})$ is used to identify the probability of being still alive at time t where $\mathbf{x} = (x_1, ...x_p)^\intercal$ is the covariates of dimension p. The hazard function is defined as

$$h(t|\mathbf{x}) = \lim_{\triangle t \to 0} \frac{Pr(t \leq O \leq t + \triangle t | O \geq t; \mathbf{x})}{\triangle t}, \tag{1}$$

which assesses the instantaneous rate of failure at time t. In the modeling methods, Cox proportional hazard model[1] is among the most popular one. The hazard function for the Cox proportional hazard model has the form

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\beta^\intercal \mathbf{x}) \tag{2}$$

where $\beta = (\beta_1, ..., \beta_p)^\intercal$ is a vector of regression parameters, and $h_0(t)$ is the baseline hazard. We can define $f(\mathbf{x}) = \beta^\intercal \mathbf{x}$ as a risk function. This gives the hazard rate at time t for the patient i with covariate vector $\mathbf{x}_i$.

A major challenge is that the number of features p is much larger than the number of patients n. To handle high-dimensional data, many feature selection methods have been adapted to the Cox regression setting for censored survival data.[3–7,18] Another type of hazard model is estimated by logistic regression such that the probability of surviving beyond t is $Pr(O \geq t|x) = (1 + \exp[x^\top \beta(t) + th])^{-1}$ with a threshold th.[19,20] Instead of defining the hazard function, one recent work transforms the original survival analysis problem into a multi-task learning problem by decomposing the regression component into related classification tasks; the new objective function can be solved by popular ADMM based optimization.[9] It is a good way to learn highly complex survival functions by using the advanced neural networks techniques.[10,12] We can get the risk score through neural networks and now denote the risk for the patient i as $\mathbf{o}_i$. Deepsurv[10] is the earlier attempt to learn a nonlinear risk function by replacing the linear part $\beta^T x$ in $f(x)$ with a nonlinear deep fully connected network.

One very simple way for data fusion is to create a concatenated feature vector comprising of all features selected individually from each modality.[21] However, a powerful feature selection is required to search for those important biomarkers from the original features, and each modality is processed individually without considering their inter-connections. The inherent challenge in combining data streams for survival analysis is that individual data sources are very heterogeneous due to the heterogeneity of tumors. However, recent studies have shown that different views actually share common representations to describe tumor morphology, which is very important for diagnosis.[14] A key challenge for survival analysis is how to eliminate view-discrepancies and learn such common representations.

2.2. *Deep multi-modal embeddings*

Recent deep multi-modal embeddings[22–26] provide a very good solution to the above challenge. They have been successfully applied in computer vision applications such as image-text matching[23,26] and image reconstruction utilizing multiple auto-encoders.[24,25]

In finding a correlated meta-space for data fusion, recent DNN-based multi-view methods provide very complex representation learning using deep neural networks (DNNs) that maximizes signals which are common to data from multiple modalities. They can learn much more comprehensive representation and more easily process large amounts of training data. However, these methods belong to unsupervised feature learning, which is incapable of survival analysis since it cannot guarantee that the integrated feature space is highly associated with patients' survival outcome. In addition, recent cancer datasets cannot provide multi-modalities data with sufficient patient samples, while deep multi-modal embeddings need large amounts of data.

3. Methodology

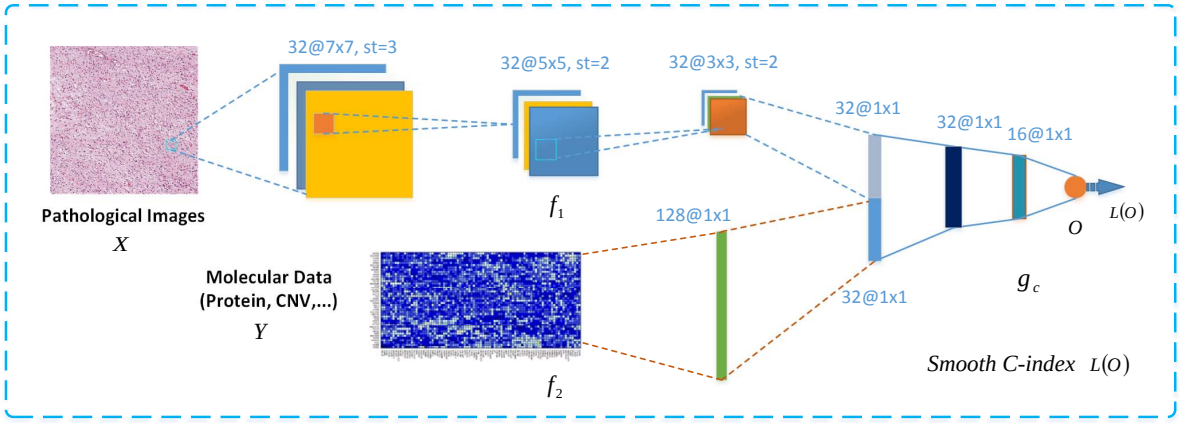


Fig. 2. The architecture of our framework. 'st' is short for 'stride'.

Figure 2 shows the pipeline of the proposed framework. f_1, f_2 is the view-specific sub-network and g_c is the common sub-network. The proposed model uses Convolutional Neural Networks (CNNs) as one image-view sub-network f_1 and Fully Connected Neural Networks (FCNs) as another view-specific sub-network f_2 to learn deep representations from pathological images and molecular profiling data, respectively. More details about the sub-network f_1 can be seen in 1. It consists of 3 convolutional layers, 1 max-pooling layer and 1 fully-connected layer. In each convolutional layer, we employ ReLU as the nonlinear activation function.

The sub-network f_2 has two fully connected layers equipped with ReLU activation function, with 128 and 32 neurons, respectively.

Table 1. The architecture of CNNs

Layer	Filter size, stride, number
Conv (ReLU)	7×7, 3, 32
Conv (ReLU)	5×5, 2, 32
Conv (ReLU) Max-pool	3×3, 2, 32 2×2
FC	32

3.1. *Deep Correlational Learning*

Denote $\mathbf{x}_i, \mathbf{y}_i$ from two views as i-th sample, its representation passing through the corresponding view sub-network is denoted as $f_1(\mathbf{x}_i; \mathbf{w_x})$ and $f_2(\mathbf{y}_i; \mathbf{w_y})$ respectively. $\mathbf{w_x}, \mathbf{w_y}$ represent all

parameters of the two sub-networks. The outputs of the two branches will be connected to a correlation layer to form the common representation.

In the correlational layer, deep correlational learning tries to find pairs of projections that maximize the correlation of two outputs from each network $f_1(\mathbf{x}_i; \mathbf{w_x}), f_2(\mathbf{y}_i; \mathbf{w_y})$. If $\mathbf{w_x}, \mathbf{w_y}$ represent all parameters of two networks, then the commonality is enforced by maximizing the correlation between two views as

$$L = corr(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{m}(f_1(\mathbf{x}_i) - \overline{f_1(\mathbf{X})})(f_2(\mathbf{y}_i) - \overline{f_2(\mathbf{Y})})}{\sqrt{\sum_{i=1}^{m}(f_1(\mathbf{x}_i) - \overline{f_1(\mathbf{X})})^2 \sum_{i=1}^{m}(f_2(\mathbf{y}_i) - \overline{f_2(\mathbf{Y})})^2}}, \tag{3}$$

where networks' parameters $\mathbf{w_x}, \mathbf{w_y}$ are omitted in the loss function (3). We can maximize the correlation loss function to generate the shared representation indicating the most correlated features from two modalities. Although different views of patients' data are very heterogeneous, there still share some common information for survival prediction. Correlational learning provides a very good way to find such common representation using the correlation function (3). However, it belongs to unsupervised learning and thus this procedure has a risk of losing the discriminant markers for predicting patients' survival outcomes.

3.2. *Survival prediction with smooth C-index loss function*

Denote $\mathbf{O} = [o_1, ..., o_N]^\top$ as the outputs of common sub-network $\mathbf{g}_c$, i.e., $o_i = \mathbf{g}_c(\mathbf{z}_i)$. The final model will be fine-tuned on the survival prediction task using the knowledge from the deep correlational learning. This will give the proposed model the ability to discover important markers that are ignored by the correlational model, and learn the best representation for survival prediction. Different from the use of negative log partial likelihood as survival loss in recent deep survival learning,[11] we propose to minimize the smoothed empirical risk function[27] which is from the concordance index (C-index) estimator and differentiable with respect to the predictor o_i.

During the past few decades, the C-index, a general discrimination measure for the evaluation of prediction models, has gained enormous popularity in biomedical research. The concordance index (C-index) quantifies the ranking quality of rankings and is calculated as

$$c = P(o_i > o_j | T_i < T_j) \tag{4}$$

where T_i, T_j and o_i, o_j are the event times and the predicted risk values. The C-index measures whether large values of o are associated with short survival times T and vice versa. Uno et al. proposed a modified C-index estimation as follows:[28]

$$C_{uno} = \frac{\sum_{i,k} \delta_i (G_m(T_i))^{-2} I(T_i < T_k) I(o_i > o_k)}{\sum_{i,k} \delta_i (G_m(T_i))^{-2} I(T_i < T_k)}. \tag{5}$$

where $G_m(t)$ denotes the Kaplan-Meier estimator of the unconditional survival function of Censored time (C_{cens}) estimated from the learning data. However, the Uno estimator is unfeasible because it is not differentiable to o_i. To solve this problem, the indicator function $I(o_i > o_k)$ is approximated by the sigmoid function:

$$L(\mathbf{o}) = \sum_{i,k} w_{i,k} \frac{1}{1 + \exp(\frac{o_k - o_i}{\sigma})}, \tag{6}$$

where o_i is the output of the i-th patient. We implement the smoothed C-index function (6) as the survival loss function in our method. The weights w_{ik} are defined as

$$w_{i,k} = \frac{\delta_i (G_m(T_i))^{-2} I(T_i < T_k)}{\sum_{i,k} \delta_i (G_m(T_i))^{-2} I(T_i < T_k)}. \tag{7}$$

where $I(T_i < T_k)$ is an indication function that indicates whether T_i is larger than T_k or not. It is easy to check the smoothed empirical risk is differentiable with respect to the predictor o_i. The derivative is given by

$$\frac{\partial L}{\partial o_i} = -\sum_{k} w_{i,k} \frac{\exp(\frac{o_k - o_i}{\sigma})}{\sigma(1 + \exp(\frac{o_k - o_i}{\sigma}))} \tag{8}$$

Compared with recent deep survival models,[10,29] which can only handle one specific view of data, our model can achieve more complex architecture for the integration of multi-modalities data, which can be used for practical applications on more challenging datasets.

4. Experiments

4.1. *Dataset Description*

TCGA (The Cancer Genome Atlas) data cohort[30] is a very large dataset which contains both high resolution whole slide pathological images and molecular profiling data. In TCGA-cohort, we focused on glioblastoma multiforme (GBM) and lung squamous cell carcinoma (LUSC). For each cancer type, we adopted a core sample set from UT MD Anderson Cancer Center[31] in which each sample has information for the overall survival time, pathological images, and molecular data related to gene expression. For model evaluation, 80% of patients were randomly selected for training and the remaining 20% were used for testing.

- **TCGA-LUSC**: Lung squamous cell carcinoma (LUSC) is one major type in Non-Small-Cell Lung Carcinoma (NSCLC). 106 patients with pathological images and protein expression (reverse-phase protein array, 174 proteins) are collected in our experiments.
- **TCGA-GBM**: Glioma is a type of brain cancer, and it is the most common malignant brain tumor. 126 patients are selected from the core set with images and CNV data (Copy number variation, 106 dimension).

4.2. *Comparison approaches*

We compare our model with four state-of-the-art survival approaches and three baseline deep survival models. The four survival methods include LASSO-Cox,[18] Parametric censored regression models with components with Weibull, Logistic distribution,[2] and Boosting concordance index (BoostCI).[27] Those above methods need hand-crafted features as inputs. To calculate imaging hand-crafted features, we used CellProfiler[32] to analyze pathological images in comparison survival models. CellProfiler is widely used as a state-of-the-art medical image feature

extracting and quantitative analysis tool. Motivated by the pipeline,[33] a total of 1,795 quantitative features were calculated from each image tile.

The three baseline deep survival models are as follows:

- **CNN-Surv**: Deep convolutional survival model;[29] we use the same architecture as the sub-network f_1 .
- **FCN-Surv**: FCN sub-network f_2 followed by negative log partial likelihood loss.[10]
- **DeepCorr+DeepSurv**: The shared representation learned by deep correlational learning is directly fed to another DeepSurv model.

To make fair comparisons, the architectures of different deep survival models are kept the same as the corresponding parts in the proposed method.

4.3. *Results and Discussion*

To evaluate the performances in survival prediction, we take the concordance index (CI) as our evaluation metric.

Table 2. Performance comparison of the proposed methods and other existing related methods

Data	Model	LUSC	GBM
Images	LASSO-Cox[18]	0.3411	0.5775
	BoostCI[27]	0.5088	0.5565
	Weibull[2]	0.4261	0.4787
	Logistic[2]	0.4217	0.4921
	CNN-Surv[29]	0.5797	0.5154
Protein/CNV	LASSO-Cox[18]	0.6231	0.4920
	BoostCI[27]	0.5714	0.4676
	Weibull[2]	0.4851	0.5659
	Logistic[2]	0.3915	0.4218
	FCN-Surv[10]	0.5462	0.5221
Integration	DeepCorr+DeepSurv	0.5622	0.5900
	Proposed	**0.6638**	**0.6045**

Results in Table 2 presents the C-index values by various survival methods on TCGA-LUSC and TCGA-GBM. It can be seen that the integration of both modalities in the proposed model achieves the best performance, for both lung and brain cancer. That is because the proposed method can remove view discrepancy as well as learn the survival-related common representations from both modalities. The difference in the DeepCorr+DeepSurv from ours is that those two models are trained separately. Performance shows that the common representation by maximizing the correlation in an unsupervised manner still has the risk of discarding markers that are highly associated with survival outcomes. In fact, the proposed model used a similar smoothed C-index as the survival loss function compared with BoostCI,[27] but the proposed method outperforms BoostCI in evaluation. This demonstrates that the proposed method can efficiently learn deep representation from two modalities and achieve better predictions.

From the results, we can see that it is not easy to find a general model that can successfully estimate patients' survival outcomes across different tumor types using only one specific view, either images or molecule data. The reason might be the heterogeneous of different tumor types and the original data in each view might contain variations or noise and thus affect the estimation of survival models. Because the proposed model can effectively integrate two views, it can achieve good prediction performance across different tumor types.

5. Conclusion

In this paper, we proposed a deep survival model to efficiently integrate multi-modalities from lung and brain tumor patients. Eliminating the view discrepancy between imaging data and molecular profiling data, deep correlational learning provides a good solution to maximize the correlation of two views and find the common embedding space. However, deep correlational learning belongs to an unsupervised learning which cannot ensure the common representation from correlational layer is suitable for survival prediction. To overcome this issue, the proposed model fine-tunes the whole network using smooth C-index loss after transferring knowledge from the embedding space. Experiments have demonstrated the proposed method can discover important markers that might be ignored by correlational learning. Our model can find non-linear relationships between factors and prognosis; it achieved quite promising performance with improvements. In the future, we will extend the proposed framework to directly process original whole slide images (WSIs).

References

1. D. R. Cox, *Journal of the Royal Statistical Society. Series B (Methodological)* , 187 (1972).
2. J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data* (John Wiley & Sons, 2011).
3. E. Bair and R. Tibshirani, *PLoS Biol* **2**, p. E108 (2004).
4. E. Bair, T. Hastie, D. Paul and R. Tibshirani, *Journal of the American Statistical Association* **101** (2006).
5. H. C. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer and L. F. Wessels, *Statistics in medicine* **25**, 3201 (2006).
6. M. Y. Park and T. Hastie, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 659 (2007).
7. H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi and O. C. Lingjærde, *Bioinformatics* **23**, 2080 (2007).
8. Y. Li, K. S. Xu and C. K. Reddy, Regularized parametric regression for high-dimensional survival analysis, in *In Proceedings of SIAM International Conference on Data Mining. SIAM*, 2016.
9. Y. Li, J. Wang, J. Ye and C. K. Reddy, A multi-task learning formulation for survival analysis, in *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
10. J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang and Y. Kluger, *arXiv preprint arXiv:1606.00931* (2016).
11. J. Yao, X. Zhu, F. Zhu and J. Huang, Deep correlational learning for survival prediction from multi-modality data, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.
12. X. Zhu, J. Yao, F. Zhu and J. Huang, Wsisa: Making survival prediction from whole slide

histopathological images, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
13. A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn and D. Koller, *Science translational medicine* **3**, 108ra113 (2011).
14. Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell *et al.*, *Science translational medicine* **4**, 157ra143 (2012).
15. A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*, 2012.
16. K. Chatfield, K. Simonyan, A. Vedaldi and A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in *British Machine Vision Conference*, 2014.
17. C. K. Reddy and Y. Li, A review of clinical prediction models, in *Healthcare Data Analytics*, (Chapman and Hall/CRC, 2015) pp. 343–378.
18. R. Tibshirani *et al.*, *Statistics in medicine* **16**, 385 (1997).
19. H.-c. Lin, V. Baracos, R. Greiner and J. Y. Chun-nam, Learning patient-specific cancer survival distributions as a sequence of dependent regressors, in *Advances in Neural Information Processing Systems*, 2011.
20. X. Song and C.-Y. Wang, *Statistics in medicine* **32** (2013).
21. X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar and J. Huang, Lung cancer survival prediction from pathological images and genetic data - an integration study, in *IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016.
22. G. Andrew, R. Arora, J. A. Bilmes and K. Livescu, Deep canonical correlation analysis., in *ICML*, 2013.
23. F. Yan and K. Mikolajczyk, Deep correlation for matching images and text, in *CVPR*, June 2015.
24. W. Wang, R. Arora, K. Livescu and J. Bilmes, On deep multi-view representation learning, in *Proc. of the 32st Int. Conf. Machine Learning (ICML 2015)*, 2015.
25. S. Chandar, M. M. Khapra, H. Larochelle and B. Ravindran, *Neural computation* (2016).
26. L. Wang, Y. Li and S. Lazebnik, *arXiv preprint arXiv:1511.06078* (2015).
27. A. Mayr and M. Schmid, Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations (1) (Public Library of Science, 2014) p. e84483.
28. H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino and L. Wei, *Statistics in medicine* **30**, 1105 (2011).
29. X. Zhu, J. Yao and J. Huang, Deep convolutional neural network for survival analysis with pathological images, in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016.
30. C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski *et al.*, *Nature* **502**, 333 (2013).
31. Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao *et al.*, *Nature biotechnology* **32**, 644 (2014).
32. A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat *et al.*, *Genome biology* **7**, p. R100 (2006).
33. J. Yao, S. Wang, X. Zhu and J. Huang, Imaging biomarker discovery for lung cancer survival prediction, in *MICCAI 2016, Part II*, eds. S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal and W. Wells, LNCS, Vol. 9901 (Springer, Heidelberg, 2016).

Genotype-phenotype association study via new multi-task learning model

Zhouyuan Huo
Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, United States
E-mail: zhouyuan.huo@pitt.edu

Dinggang Shen
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
E-mail: dinggang_shen@med.unc.edu

Heng Huang*
Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260, United States
E-mail: heng.huang@pitt.edu

Research on the associations between genetic variations and imaging phenotypes is developing with the advance in high-throughput genotype and brain image techniques. Regression analysis of single nucleotide polymorphisms (SNPs) and imaging measures as quantitative traits (QTs) has been proposed to identify the quantitative trait loci (QTL) via multi-task learning models. Recent studies consider the interlinked structures within SNPs and imaging QTs through group lasso, e.g. $\ell_{2,1}$-norm, leading to better predictive results and insights of SNPs. However, group sparsity is not enough for representing the correlation between multiple tasks and $\ell_{2,1}$-norm regularization is not robust either. In this paper, we propose a new multi-task learning model to analyze the associations between SNPs and QTs. We suppose that low-rank structure is also beneficial to uncover the correlation between genetic variations and imaging phenotypes. Finally, we conduct regression analysis of SNPs and QTs. Experimental results show that our model is more accurate in prediction than compared methods and presents new insights of SNPs.

Keywords: Quantitative Trait Loci; Single Nucleotide Polymorphisms (SNPs); Quantitative Traits (QTs); Multi-Task Learning.

1. Introduction

Research on the associations between genetic variations and imaging phenotypes is developing with the advance in high-throughput genotype and brain image techniques.[1–4] Alzheimers Disease Neuroimaging Initiative (ADNI) provides a suitable dataset for genotype-phenotype study, however it is still challenging to find out whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), genetic factors such as single nucleotide polymorphisms (SNPs) can be

*Corresponding Author. This work was partially supported by U.S. NIH R01 AG049371, NSF IIS 1302675, IIS 1344152, DBI 1356628, IIS 1619308, IIS 1633753

combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's Disease (AD). Given these data, researchers did the association study between genetic variation and imaging measures as quantitative traits (QTs), which was shown to have increased statistical power and decreased sample size requirements.[5] Through the analysis of strong associations between SNPs and imaging phenotypes, we can also identify candidate genes or loci which are relevant to the biological etiology of the disease.[2]

Traditional association studies use univariate or multivariate methods to discover the associations between single nucleotide polymorphisms (SNPs) and imaging measures as quantitative traits (QTs).[6,7] However, these methods treat each regression of imaging phenotype as an independent task, thus the correlations between SNPs and QTs are lost in this model. To solve this problem, regression analysis of SNPs and QTs has been proposed to identify the quantitative trait loci (QTL) via multi-task learning models.[4,8] In multi-task learning model, multiple tasks are handled jointly and dependently. For example, by imposing the interlinked structures within SNPs and imaging QTs through group lasso, e.g. $\ell_{2,1}$-norm,[9,10] it leads to better predictive results and more insights of the SNPs.[4] This assumption is suitable for the fact that only a small fraction of SNPs are responsible for the imaging manifestations of complex diseases. However, there are two limitations. Firstly, group sparsity is not enough for representing the intrinsic correlation between SNPs and imaging QTs. Apart from group sparsity, we can also benefit from the low-rank structure of the coefficient. Secondly, although $\ell_{2,1}$-norm regularization is common for the group sparsity, it is sensible to outliers.[11] For example, the value of $\ell_{2,1}$-norm of matrix $[[100], [0], [0]]$ is larger than $[[1], [1], [1]]$, however, the first matrix is more sparse rather than the second one.

In this paper, we propose a new multi-task learning model to analyze the associations between SNPs and QTs. We suppose that low-rank structure is also beneficial to uncover the correlation between genetic variations and imaging phenotypes. This assumption is reasonable because different SNPs may have similar effect on the imaging phenotypes. For example, both APOE SNPs rs429358 and rs7412 are the strongest known genetic risk factors for Alzheimer's Disease. In order to make the feature selection robust to outliers, we propose to use capped $\ell_{2,1}$-norm regularization in place of $\ell_{2,1}$-norm. We conduct regression analysis of SNPs and QTs from ADNI, and the experimental results show that our model is more accurate in prediction than compared methods and it presents new insights of SNPs as well.

2. Data Description

We use the dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). One goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. These data are obtained from 818 participants. Further information about ADNI can be found at see www.adni-info.org.

We use the genotype data[12] of all non-Hispanic Caucasian participants from the ADNI Phase 1 cohort. They were genotyped using the Human 610-Quad BeadChip. Only SNPs which belong to the top 40 AD candidate genes listed on the AlzGene database (www.alzgene.org) as of 4/18/2011[13] were selected after the standard quality control (QC) and imputation steps. The QC criteria for the

SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. After that, the quality-controlled SNPs were imputed using the MaCH software[14] to estimate the missing genotypes in the second pre-processing step. In this paper, we use 3123 SNPs in total. While most of them might be irrelevant to AD, only a small fraction of them are risk factors for the disease and associated with imaging phenotypes. For example, gene APOE and TOMM40 are known to be the contributors to AD.

Two widely employed automated MRI analysis techniques were used to process and extract imaging phenotypes from scans of ADNI participants as previously described.[3] First, Voxel-Based Morphometry (VBM)[15] is performed to define global gray matter (GM) density maps and extract local GM density values for target regions. Second, automated parcellation via FreeSurfer V4[16] is conducted to define volumetric and cortical thickness values for regions of interest (ROIs) and to extract total intracranial volume (ICV). All these measures were adjusted for the baseline ICV using the regression weights derived from the healthy control (HC) participants. Further details are available in.[3] In this paper, we use 36 ROIs from VBM and 24 ROIs from FreeSurfer which are known to be related to AD. VBM measures and FreeSurfer measures are treated as QTs for identifying QTLs independently.

3. Proposed Method

In this section, we propose a new multi-task learning model to study the intrinsic associations between SNPs and imaging phenotypes. Throughout our paper, we use $X \in \mathbb{R}^{d\times n}$ to denote the SNP data of all the ADNI participants, and $Y \in \mathbb{R}^{c\times n}$ to denote the selected imaging phenotypes, where n is the number of participants, d is the number of SNPs and c denotes the number of selected imaging phenotypes or QTs. It is a standard regression problem to predict continuous quantities Y using SNPs data X as follows:

$$\min_{W\in\mathbb{R}^{d\times c}} \|W^T X - Y\|_F^2 \tag{1}$$

The learned weight matrix W shows the importance of each SNP to predict imaging phenotypes, e.g. W_i^j denotes the importance of i-th SNP to predict j-th imaging phenotype. There are mainly three drawbacks of using model (1) as the objective function to learn the coefficient matrix W. Firstly, it is easy to overfit if there is no regularization, and the learned W is hard to generalize to new data. Secondly, the learned coefficient matrix W is not sparse. It is intuitive that only a small fraction of SNPs should be relevant to imaging quantitative traits (QTs), thus sparsity of W is a nontrivial property. The last but not the least, the associations within SNPs or imaging phenotypes are overlooked. Coefficient matrix W should come from a specific domain, we can impose a structured regularization on W to represent the intrinsic associations within SNPs or imaging phenotypes. We usually use l_2-norm regularization to avoid overfitting, however, the last two problems are still not solved yet. To handle these issues, we can treat the regression of each column of Y (each quantitative trait (QT)) as a task, then we can use multi-task learning model to learn multiple tasks jointly. The original problem (1) can be represented as a multi-task problem as follows:

$$\min_{W=[W^1,\ldots,W^T]\in\mathbb{R}^{d\times c}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 + \text{Reg}(W) \tag{2}$$

where $T = c$ (the number of tasks), $n_t = n, \forall t \in \{1, ..., T\}$ (the number of samples in task t). In task t, $x_{i,t} = X^i$, which is the column i of X; $y_{i,t} = Y_t^i$, which is the element of Y at the position of row t and column i; W^t denotes the column t of matrix W. Reg(W) is the regularization we impose on the multi-task learning problem, and it represents our assumption of the correlation between multiple tasks , *e.g.* low-rank or group sparsity.[17,18] In the following context, we propose to impose two new regularization terms in the multi-task problem to learn the associations between SNPs and imaging phenotypes, one for genetic association and the other one for quantitative trait loci (QTLs) identification.

3.1. *Capped Trace Norm Regularization for Genetic Association*

In multi-task learning, we assume that the regression tasks between SNPs and imaging phenotypes are correlated. Then we can benefit from learning multiple tasks jointly. Their correlation can be represented by imposing a structure on the coefficient matrix W. In this paper, we assume that matrix W has a low-rank subspace, which is widely used in many applications, such as recommendation system[19,20] and multi-task learning.[21,22] This assumption is also fit for the genome-phenotype associations, because multiple SNPs may have similar effects on the imaging phenotype. For example, both APOE SNPs rs429358 and rs7412 are the strongest known genetic risk factors for Alzheimer's Disease. The non-convex rank minimization regularization Reg$(W) =$ rank(W) is hard to optimize, for simplicity, trace norm is proposed as the best convex relaxation for the rank minimization regularization as follows[23] :

$$\text{Reg}(W) = \|W\|_* = \sum_{i=1}^{\min\{d,c\}} \sigma_i(W) \tag{3}$$

where σ_i is the singular value of matrix W. However, there is a big gap between rank minimization regularization and trace norm regularization. When some non-zero singular values of W changes, the value of trace norm also changes. In contrast, the rank of matrix W keeps constant. Besides, trace norm is also sensitive to outliers.

In this paper, we propose to use a tighter approximation of rank minimization than trace norm. Capped trace norm is more general than trace norm and it is represented as follows:

$$\text{Reg}(W) = \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i(W), \varepsilon_1\} \tag{4}$$

where ε_1 works as a threshold. If ε_1 is large enough, for any i, we have $\sigma_i(W) < \varepsilon_1$, then it is equal to trace norm regularization. When we reduce the value of ε_1, where $\varepsilon_1 \in \Big(min\{\sigma_i(W)\}, max\{\sigma_i(W)\}\Big)$, it's obvious that those singular values larger than ε_1 will be ignored in the optimization. So, instead of minimizing the sum of all singular values in the trace norm regularization, we focus on minimizing these singular values less than ε_1 and ignore large singular values. Therefore, capped trace norm regularization is more robust to outliers.

3.2. *Capped $\ell_{2,1}$-Norm Regularization for QTLs Identification*

There are 3123 SNPs in our dataset, and only a fraction of them is relevant to specific imaging quantitative traits (QTs). Therefore, W should be structured sparse, where each row of W is treated

as a unit. If SNP i is not important, $W_i = \mathbf{0} \in \mathbb{R}^{1\times c}$. $\ell_{2,0}$-norm regularization, $\text{Reg}(W) = \|\mathbf{w}\|_0$, minimizes the number of non-zero elements, where $\mathbf{w} \in \mathbb{R}^{d\times 1}$ and $\mathbf{w}_i = \|W_i\|_2$. However, it is a non-convex problem and hard to optimize. Alternatively, we usually use $\ell_{2,1}$-norm regularization enforce the structured sparsity on the learned coefficient matrix W:[4,9]

$$\text{Reg}(W) = \|W\|_{2,1} = \sum_{i=1}^{d} \|W_i\|_2 = \|\mathbf{w}\|_1 \tag{5}$$

where W_i denotes the i-th row of matrix W. Each row of W is treated as a unit, and if SNP i is negligible, $W_i = \mathbf{0} \in \mathbb{R}^{1\times c}$. Although $\ell_{2,1}$-norm regularization works fine, there is gap between $\ell_{2,0}$-norm regularization and $\ell_{2,1}$-norm regularization. Increasing the value of non-zero elements in $\mathbf{w}$ does not affect the number of its non-zero elements $\|\mathbf{w}\|_0$; on the contrary, $\|\mathbf{w}\|_1$ will increase. In this paper, we propose to use capped $\ell_{2,1}$-norm regularization as an alternative to $\ell_{2,0}$-norm as follows:

$$\text{Reg}(W) = \sum_{i=1}^{d} \min\{\|W_i\|_2, \varepsilon_2\} \tag{6}$$

Capped $\ell_{2,1}$-norm regularization is a better approximation of $\ell_{2,0}$-norm than $\ell_{2,1}$-norm. It treats $\|W_i\|_2$ equally if it is larger than ε_2, hence capped $\ell_{2,1}$-norm regularization is more robust to outliers. When ε_2 is large enough, we have $\min\{\|W_i\|_2, \varepsilon_2\} = \|W_i\|_2, \forall i$, thus capped $\ell_{2,1}$-norm is equal to $\ell_{2,1}$-norm.

To sum up, combining capped trace norm regularization and capped $\ell_{2,1}$-norm together makes our proposed objective function for multi-task learning (7) as follows:

$$\min_{W\in\mathbb{R}^{d\times c}} \sum_{t=1}^{T}\sum_{i=1}^{n_t} \min \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 + \gamma_1 \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i(W), \varepsilon_1\} + \gamma_2 \sum_{i=1}^{d} \min\{\|W_i\|_2, \varepsilon_2\} \tag{7}$$

where the notations are similar to problem (2). γ_1 and γ_2 are to balance the importance of two regularizations. In following sections, we will propose an efficient optimization algorithm for problem (7) and prove that it is sequence convergent.

4. Optimization Algorithm

In this section, we propose an efficient optimization algorithm to solve problem (7). Optimizing the non-smooth and non-convex problem (7) directly is very hard. Through re-weighted algorithm,[24] in each step, we can transform our objective function to a smooth and convex relaxed problem, so that we are able to compute the optimal solution to the new relaxed problem until convergence.

Firstly, we do Singular Value Decomposition (SVD) on the coefficient matrix W and we have $W = U\Sigma V^T$, where singular values $\sigma_i(W)$ of matrix W are in ascending order. Assuming there are k singular values smaller than ε_1, we define $D = \frac{1}{2}\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T$ where U^i is the i_{th} column of matrix U. Therefore, the second term in (7) can be represented as $\gamma_1 \text{Tr}(W^T D W)$. Secondly, we compute Z_{ii} for each row of matrix W:

$$Z_{ii} = \begin{cases} \frac{1}{2\|W_i\|_2} & \text{if } \|W_i\|_2 < \varepsilon_2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

All the non-diagonal elements of matrix Z are 0. Therefore, the third term in (7) can be represented by $\gamma_2 \text{Tr}(W^T ZW)$. When we fix the values of D and Z, the objective function (7) can be written as a smooth and convex problem as follows:

$$\min_{W=[W^1,\ldots,W^T]} \|W^T X - Y\|_F^2 + \gamma_1 \text{Tr}(W^T DW) + \gamma_2 \text{Tr}(W^T ZW) \tag{9}$$

where the loss term is from $\sum_{t=1}^{T} \sum_{i=1}^{n_t} \|(W^t)^T x_{i,t} - y_{i,t}\|_2^2 = \|W^T X - Y\|_F^2$ as per the definition of our variables. Finally, taking the derivative of (9) in terms of W and setting it to zero, we can get the optimal solution to the problem (9) as follows:

$$W = (XX^T + \gamma_1 D + \gamma_2 Z)^{-1} XY^T \tag{10}$$

To sum up, our proposed optimization algorithm is presented in Algorithm 1.

Algorithm 1 Algorithm to solve problem (7)

Input: Training data for multiple tasks $X \in \mathbb{R}^{d \times n}, Y \in \mathbb{R}^{c \times n}$
Output: $W \in \mathcal{R}^{d \times c}$.
Initialize W.
while not converge **do**
  Compute D and Z via (4) and (8).
  Fix D and Z, and compute matrix W via (10).
end while

5. Convergence Analysis

By optimizing our model with Algorithm 1, we can solve the non-smooth and non-convex objective function (7). In this section, we presents the convergence analysis of our proposed algorithm.

Theorem 1. *Through Algorithm 1, the values of objective function (7) are non-increasing monotonically, and it will converge to a local solution.*

In order to prove Theorem 1, we need the following Lemmas.

Lemma 1. *According to,[25] any two hermitian matrices $A, B \in R^{n \times n}$ satisfy the following inequality:*

$$\sum_{i=1}^{n} \sigma_i(A)\, \sigma_{n-i+1}(B) \leq \text{Tr}\left(A^T B\right) \leq \sum_{i=1}^{n} \sigma_i(A)\, \sigma_i(B) \tag{11}$$

where $\sigma_i(A)$, $\sigma_i(B)$ are singular values sorted in the same order.

Lemma 2. *Let $W = U\Sigma V^T$, Σ is a diagonal matrix and σ_i are singular values of W in ascending order. There are k singular values less than ε_1. $\hat{W}$ is coefficient matrix in next iteration by using Algorithm 1, and $\hat{W} = \hat{U}\hat{\Sigma}\hat{V}^T$, where $\hat{\sigma}_i$ are singular values of $\hat{W}$ in ascending order and U^i is the*

i-th column of U. There are $\hat{k}$ singular values less than ε_1. So it is true that:

$$\sum_{i=1}^{\min\{d,c\}} \min\{\hat{\sigma}_i, \varepsilon_1\} - \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T \hat{W}\hat{W}^T\right) \tag{12}$$

$$\leq \sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i, \varepsilon_1\} - \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T W W^T\right) \tag{13}$$

Proof: It's obvious that $\sigma_i - 2\hat{\sigma_i} + \sigma_i^{-1}\hat{\sigma_i}^2 = \frac{1}{\sigma_i}\left(\sigma_i^2 - 2\sigma_i\hat{\sigma_i} + \hat{\sigma_i}^2\right) \geq 0$. Thus we have:

$$\sum_{i=1}^{k}\left(\hat{\sigma}_i - \frac{1}{2}\sigma_i^{-1}\hat{\sigma}_i^2\right) \leq \frac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{14}$$

Because there are $\hat{k}$ singular values of $\hat{W}$ less than ε_1 and they are sorted in ascending order, so first $\hat{k}$ singular values $\hat{\sigma}_i$ are less than ε_1. Therefore, no matter $\hat{k} \geq k$ or $\hat{k} < k$, it holds that:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i - \hat{k}\varepsilon_1 \leq \sum_{i=1}^{k}\hat{\sigma}_i - k\varepsilon_1 \tag{15}$$

Combining (14) and (15), we get the following inequality:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i - \frac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 - \hat{k}\varepsilon_1 \leq \frac{1}{2}\sum_{i=1}^{k}\sigma_i - k\varepsilon_1 \tag{16}$$

Suppose there are $n = \min\{d, c\}$ singular values in total, adding $n\varepsilon_2$ on both sides, we are able to get the following inequality:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i + \left(n - \hat{k}\right)\varepsilon_1 - \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 \leq \sum_{i=1}^{k}\sigma_i + (n-k)\,\varepsilon_1 - \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{17}$$

According to the definition of matrix D in (4), the following equality holds that:

$$\frac{1}{2}\operatorname{Tr}(W^T D W) = \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T W W^T\right) = \frac{1}{2}\operatorname{Tr}\left(U\Lambda U^T U\Sigma^2 U^T\right) = \frac{1}{2}\sum_{i=1}^{k}\sigma_i \tag{18}$$

where Λ is the diagonal matrix where its first k elements are $\sigma_i^{-1}, i \in \{1, ..., k\}$ and other elements are 0. Via Lemma 1, we have:

$$\tfrac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\right) = \tfrac{1}{2}\operatorname{Tr}\left(U\Lambda U^T\hat{U}\hat{\Sigma}^2\hat{U}^T\right) \geq \tfrac{1}{2}\sum_{i=1}^{k}\sigma_i^{-1}\hat{\sigma}_i^2 \tag{19}$$

Substituting (18) and (19) in the inequality (17), it is satisfied that:

$$\sum_{i=1}^{\hat{k}}\hat{\sigma}_i + \left(n - \hat{k}\right)\varepsilon_1 - \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T\hat{W}\hat{W}^T\right)$$

$$\leq \sum_{i=1}^{k}\sigma_i + \left(n-k\right)\varepsilon_1 - \frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{k}\sigma_i^{-1}U^i(U^i)^T W W^T\right) \tag{20}$$

Finally, the following inequality holds that:

$$\begin{aligned} &\sum_{i=1}^{\min\{d,c\}} \min\{\hat{\sigma}_i, \varepsilon_1\} - \frac{1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T \hat{W}\hat{W}^T\Big) \\ \leq &\sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i, \varepsilon_1\} - \frac{1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T W W^T\Big) \end{aligned} \tag{21}$$

Lemma 3. *We define* $z = \begin{cases} \frac{1}{2|e|} & \textit{if } |e| < \varepsilon_2 \\ 0 & \textit{otherwise} \end{cases}$, *then the inequality holds that* $\min\{|\hat{e}|, \varepsilon_2\} - z\hat{e}^2 \leq \min\{|e|, \varepsilon_2\} - ze^2$.

Proof: If $|e| < \varepsilon_2$, we have $z = \frac{1}{2|e|}$. Via Lemma 2, let W and $\hat{W}$ be scalars $|e|$ and $|\hat{e}|$ respectively, thus $\sigma(|e|) = |e|$ and $\sigma(|\hat{e}|) = |\hat{e}|$. We substitute W, $\hat{W}$ and z in the inequality (21), it holds that:

$$\min\{|\hat{e}|, \varepsilon_2\} - z\hat{e}^2 \leq \min\{|e|, \varepsilon_2\} - ze^2 \tag{22}$$

On the other hand, if $|e| \geq \varepsilon_2$, we have $z = 0$. The following inequality always holds:

$$\min\{|\hat{e}|, \varepsilon_2\} \leq \min\{|e|, \varepsilon_2\} \tag{23}$$

Right now, we are able to prove Theorem 1 by using Lemma 2 and Lemma 3 above.

Proof: According to the step 2 in Algorithm 1, matrix W denotes the current values of our model, after we obtain the analysis solution $\hat{W}$ of function (9) through (10) . Therefore, it is guaranteed that:

$$\begin{aligned} &\|\hat{W}^T X - Y\|_F^2 + \gamma_1 \mathrm{Tr}(\hat{W}^T D \hat{W}) + \gamma_2 \mathrm{Tr}(\hat{W}^T Z \hat{W}) \\ \leq &\|W^T X - Y\|_F^2 + \gamma_1 \mathrm{Tr}(W^T D W) + \gamma_2 \mathrm{Tr}(W^T Z W) \end{aligned} \tag{24}$$

We define, $|e| = \|W_i\|_2$, $|\hat{e}| = \|\hat{W}_i\|_2$ and $z_i = Z_{ii}$. after substituting the value of $|e|$ in Lemma 3, we have:

$$\min\{\|\hat{W}_i\|_2, \varepsilon_2\} - Z_{ii}\|\hat{W}_i\|_2^2 \leq \min\{\|W_i\|, \varepsilon_2\} - Z_{ii}\|W_i\|_2^2 \tag{25}$$

By summing up from $i = 1$ to d, and multiplying both sides with γ_2, then the following inequality holds that:

$$\gamma_2 \sum_{i=1}^{d} \min\{\|\hat{W}_i\|_2, \varepsilon_2\} - \gamma_2 \mathrm{Tr}(\hat{W}^T Z \hat{W}) \leq \gamma_2 \sum_{i=1}^{d} \min\{\|W_i\|_2, \varepsilon_2\} - \gamma_2 \mathrm{Tr}(W^T Z W) \tag{26}$$

where $\sum_{i=1}^{d} Z_{ii}\|W_i\|_2^2 = \mathrm{Tr}(W^T Z W)$.

Via Lemma 2, we can easily know that:

$$\begin{aligned} &\gamma_1 \sum_{i=1}^{\min\{d,c\}} \min\{\hat{\sigma}_i, \varepsilon_1\} - \frac{\gamma_1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T \hat{W}\hat{W}^T\Big) \\ \leq \gamma_1 &\sum_{i=1}^{\min\{d,c\}} \min\{\sigma_i, \varepsilon_1\} - \frac{\gamma_1}{2}\,\mathrm{Tr}\Big(\sum_{i=1}^{k} \sigma_i^{-1} U^i (U^i)^T W W^T\Big) \end{aligned} \tag{27}$$

Finally, we combine inequalities (18), (24), (26) and (27), then we know that the objective value sequence is monotonically non-increasing:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{n_t}\|(\hat{W}^t)^T x_{i,t}-y_{i,t}\|_2^2+\gamma_1\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i(\hat{W}),\varepsilon_1\}+\gamma_2\sum_{i=1}^{d}\min\{\|\hat{W}_i\|_2,\varepsilon_2\}\\
\leq&\sum_{t=1}^{T}\sum_{i=1}^{n_t}\|(W^t)^T x_{i,t}-y_{i,t}\|_2^2+\gamma_1\sum_{i=1}^{\min\{d,c\}}\min\{\sigma_i(W),\varepsilon_1\}+\gamma_2\sum_{i=1}^{d}\min\{\|W_i\|_2,\varepsilon_2\} \qquad (28)
\end{aligned}$$

After several iterations, $\hat{W}\approx W$, the derivative of the objective function (9) is close to zero. So far, it is clear that the values of our proposed objective function will not increase by using our optimization algorithm, so we prove Theorem 1 that our optimization algorithm is non-increasing monotonically. We also know that the objective function (7) is lower bounded. We can conclude that our optimization algorithm is sequence convergent.

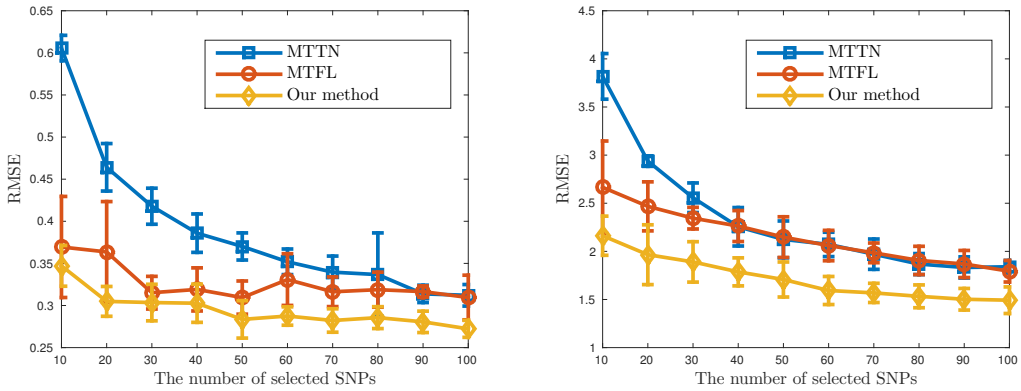
6. Experimental Results and Discussions

In this section, we evaluated our proposed model with other multi-task learning methods. The experimental dataset is from the ADNI cohort. Our goal is to select a subset of SNPs to predict the imaging phenotypes accurately. We conduct our experiments on two imaging phenotypes, FreeSurfer and VBM separately. There are two compared methods, multi-task learning with joint feature selection (MTFL)[9] and multi-task learning with trace norm regularization (MTTN),[26] both of them use least square loss to do regression. It is easy to observe that MTFL and MTTN can be represented by our proposed model. If $\gamma_2=0$ and $\varepsilon_1=\infty$, it is MTFL; if $\gamma_1=0$ and $\varepsilon_2=\infty$, it is MTTN.

We conduct 5-fold cross-validation, where 4 folds are training data and 1-fold is testing data. Then we perform internal 5-fold cross-validation on the training data, and tune parameters γ_1 and γ_2 in the range of $\{10^{-4},10^{-3},...,10^3,10^4\}$. Through the learned coefficient matrix W, we compute the weight of i_{th} SNP over all tasks by using $\sum_{j=1}^{c}|W_i^j|$. Then, we pick up the top $\{10,20,...,90,100\}$ SNPs to predict the regression responses of the testing data. For our method, although there are two other parameters ε_1 and ε_2 in the objective function (7), their values are set automatically during the optimization. In the first 5 iterations, ε_1 is set to be the 5_{th} largest singular value in $\sigma_i(W)$ and ε_2 is set to bet the 5_{th} largest value of SNP weight $\|W_i\|_2$. After that, we fix the values of ε_1 and ε_2 until convergence. In our experiments, we always stop our algorithm 1 after 20 iterations. The performance of compared method is evaluated by Root Mean Square Error (RMSE), which is a widely used measurement for regression analysis.

6.1. *Improved Phenotype Prediction*

The experimental results are presented in Figure 1. It shows the mean and standard deviation of the RMSEs obtained from 5 trails. In Figure 1, we observe that our proposed method consistently outperforms other two compared methods in both VBM phenotypes and FreeSurfer phenotypes. When we change the number of selected SNPs in our experiments, we can find out that models with joint feature selection regularization, $\ell_{2,1}$-norm or capped $\ell_{2,1}$-norm, are more stable. On the contrary, MTTN is very sensitive to the number of selected SNPs, and its performance is far worse when the number of SNPs is small. We can also observe that when the number of selected SNPs is larger than



(a) VBM (b) FreeSurfer

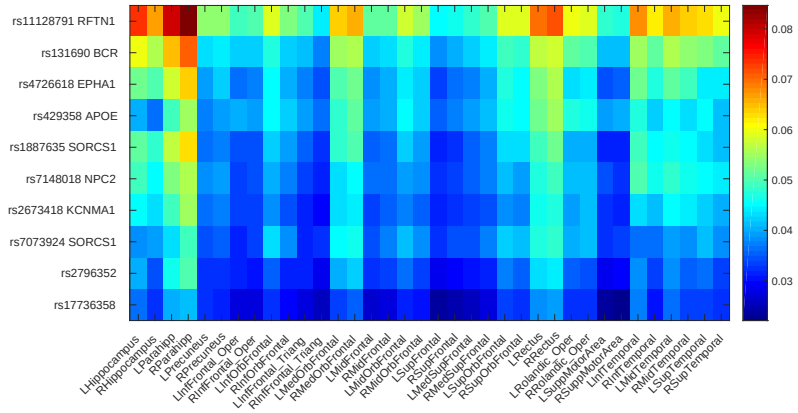
Fig. 1. Experimental results of three compared methods on two phenotypes. Average values are taken from five cross-validation and each error bar denotes $\pm$ standard deviation. Figure 1(a) shows the results of VBM phenotypes, Figure 1(b) shows the results of Freesurer phenotypes.

50, the improvement of prediction is small. Thus, we can draw a conclusion that our assumption of sparsity of coefficient matrix is correct. Although there are 3123 SNPs in our experiment, only a fraction of them is responsible for the imaging phenotypes.

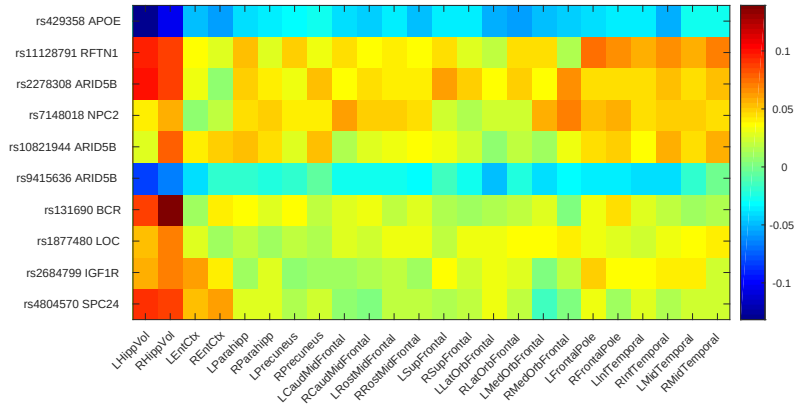
We also conduct ablation study of our method by setting $\gamma_1 = 0$ or $\gamma_2 = 0$ respectively. Table 1 presents the performance of compared methods when we select 20, 40 and 60 SNPs to predict imaging phenotypes. Firstly, we set $\gamma_2 = 0$, and our model becomes least square loss with capped trace norm regularization. We compare this model with MTTN, and experimental results demonstrate the effectiveness of capped trace norm. We also set $\gamma_1 = 0$, and our model is least square loss with capped $\ell_{2,1}$-norm regularization. We compare this model with MTFL, and it is clear that our method is more accurate in the prediction of imaging phenotypes. When we combine both of these two terms, $\gamma_1 \neq 0$ and $\gamma_2 \neq 0$, our model obtain the best results. We can draw a conclusion that although the performance of our method when $\gamma_2 = 0$ is much worse than the performance when $\gamma_1 = 0$, imposing low-rank structure on coefficient matrix is still beneficial to the regression analysis. Therefore, it is consistent with the fact that multiple SNPs may have similar effects on the imaging phenotypes.

Table 1. Ablation study of our method measured by RMSE. Value: RMSE, (comparison with corresponding method), e.g RMSE of capped $\ell_{2,1}$-norm (RMSE of capped $\ell_{2,1}$-norm $-$ RMSE of MTFL)

Phenotype	Method	20	40	60
VBM	capped trace norm ($\gamma_2 = 0$)	0.4566 (-0.0075)	0.3754 (-0.0105)	0.3398(-0.0120)
	capped $\ell_{2,1}$ ($\gamma_1 = 0$)	0.3381 (-0.0255)	0.3124 (-0.0067)	0.3066(-0.0242)
	Our Method	**0.3049**	**0.3027**	**0.2875**
FreeSurfer	capped trace norm ($\gamma_2 = 0$)	2.8623 (-0.0756)	2.2043(-0.0511)	1.9677 (-0.1047)
	capped $\ell_{2,1}$ ($\gamma_1 = 0$)	2.2030 (-0.2646)	1.8747 (-0.3883)	1.6389 (-0.4215)
	Our Method	**1.9653**	**1.7869**	**1.5934**



(a) VBM



(b) FreeSurfer

Fig. 2. Heat maps of regression coefficients learned genetic variations and quantitative traits (QTs). Top 10 selected SNPs of each matrix are visualized. Figure 2(a) shows the results from the regression of VBM measures, Figure 2(b) shows the results from the regression of FreeSurfer measures.

6.2. *Gene Selection*

Figure 2 visualizes the coefficient of top selected 10 SNPs. APOE is known to have relationship with the Alzheimer's disease (AD). Similar to previous research,[3,8] we find that APOE rs429358 shows the strongest associations with all imaging quantitative traits (QTs), especially in Figure 2(b). Clearly, our propose model is able to identify important quantitative trait loci (QTL) via joint regression analysis. Besides, we also observe that RFTN1 rs11128791 also takes important role in the imaging phenotypes, which is not identified in previous methods. These newly identified SNPs are highly correlated with the imaging phenotypes which are related to AD. They all have potential to serve as a useful generic risk factor for AD.

7. Conclusion

In this paper, we propose a new multi-task learning model with capped trace norm and capped $\ell_{2,1}$-norm regularizations. Capped trace norm helps to discover intrinsic structures within SNPs and imaging phenotypes; capped $\ell_{2,1}$-norm is more robust to select important SNPs. We propose efficient

algorithm to solve our model and provide convergence analysis. Finally, we conduct experiments on genotype-phenotype dataset from ADNI. Experimental results show that (1) our model works better in imaging phenotype prediction and (2) it helps to identify important quantitative trait loci (QTLs), which would be useful for the investigation of the generic risk factor for AD.

References

1. C. E. Bearden, T. G. Van Erp, R. A. Dutton, H. Tran, L. Zimmermann, D. Sun, J. A. Geaga, T. J. Simon, D. C. Glahn, T. D. Cannon *et al.*, *Cerebral Cortex* **17**, 1889 (2006).
2. S. G. Potkin, G. Guffanti, A. Lakatos, J. A. Turner, F. Kruggel, J. H. Fallon, A. J. Saykin, A. Orro, S. Lupoli, E. Salvi *et al.*, *PloS one* **4**, p. e6501 (2009).
3. L. Shen, S. Kim *et al.*, *Neuroimage* **53**, 1051 (2010).
4. H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin and L. Shen, *Bioinformatics* **28**, 229 (2012).
5. S. G. Potkin, J. A. Turner *et al.*, *Cogn Neuropsychiatry* **14**, 391 (2009).
6. D. H. Ballard, J. Cho and H. Zhao, *Genetic epidemiology* **34**, 201 (2010).
7. J. Bralten, A. Arias-Vásquez, R. Makkinje, J. A. Veltman, H. G. Brunner, G. Fernández, M. Rijpkema and B. Franke, *American Journal of Psychiatry* **168**, 1083 (2011).
8. H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen and A. D. N. Initiative, *Bioinformatics* **28**, i127 (2012).
9. A. Argyriou, T. Evgeniou and M. Pontil, *Machine Learning* **73**, 243 (2008).
10. G. Obozinski, B. Taskar and M. Jordan, *Statistics Department, UC Berkeley, Tech. Rep* **2** (2006).
11. H. Gao, F. Nie, W. Cai and H. Huang, Robust capped norm nonnegative matrix factorization: Capped norm nmf, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.
12. A. J. Saykin, L. Shen *et al.*, *Alzheimers Dement* **6**, 265 (2010).
13. L. Bertram, M. B. McQueen *et al.*, *Nat Genet* **39**, 17 (2007).
14. Y. Li, C. J. Willer *et al.*, *Genet Epidemiol* **34**, 816 (2010).
15. J. Ashburner and K. J. Friston, *Neuroimage* **11**, 805 (2000).
16. B. Fischl, D. H. Salat *et al.*, *Neuron* **33**, 341 (2002).
17. J. Zhou, J. Chen and J. Ye, Multi-task learning: Theory, algorithms, and applications, in *URL https://www. siam. org/meetings/sdm12/zhou_chen_ye. pdf*, 2012.
18. Z. Huo, D. Shen and H. Huang, New multi-task learning model to predict alzheimer's disease cognitive assessment, in *Medical Image Computing and Computer-Assisted Intervention*, 2016.
19. C.-J. Hsieh and P. Olsen, Nuclear norm minimization via active subspace selection, in *International Conference on Machine Learning*, 2014.
20. Z. Huo, J. Liu and H. Huang, Optimal discrete matrix completion, in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.
21. T. K. Pong, P. Tseng, S. Ji and J. Ye, *SIAM Journal on Optimization* **20**, 3465 (2010).
22. Z. Huo, F. Nie and H. Huang, Robust and effective metric learning using capped trace norm: Metric learning via capped trace norm, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
23. E. J. Candès and T. Tao, *IEEE Transactions on Information Theory* **56**, 2053 (2010).
24. F. Nie, J. Yuan and H. Huang, Optimal mean robust principal component analysis, in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014.
25. C. Theobald, An inequality for the trace of the product of two symmetric matrices, in *Mathematical Proceedings of the Cambridge Philosophical Society*, (02)1975.
26. S. Ji and J. Ye, An accelerated gradient method for trace norm minimization, in *Proceedings of the 26th annual international conference on machine learning*, 2009.

Codon bias among synonymous rare variants is associated with Alzheimer's disease imaging biomarker

Jason E. Miller[1], Manu K. Shivakumar[1], Shannon L. Risacher[2], Andrew J. Saykin[2], Seunggeun Lee[3], Kwangsik Nho[2]*, Dokyoon Kim[1, 4]*, for the Alzheimer's Disease Neuroimaging Initiative (ADNI)**

[1]*Biomedical and Translational Informatics Institute, Geisinger Health System, Danville, PA, USA*

[2]*Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA*

[3]*Department of Biostatistics,* University of Michigan*, Ann Arbor, MI, USA*

[4]*Huck Institute of the Life Sciences, Pennsylvania State University, University Park, PA, USA*

*Corresponding Author

Alzheimer's disease (AD) is a neurodegenerative disorder with few biomarkers even though it impacts a relatively large portion of the population and is predicted to affect significantly more individuals in the future. Neuroimaging has been used in concert with genetic information to improve our understanding in relation to how AD arises and how it can be potentially diagnosed. Additionally, evidence suggests synonymous variants can have a functional impact on gene regulatory mechanisms, including those related to AD. Some synonymous codons are preferred over others leading to a codon bias. The bias can arise with respect to codons that are more or less frequently used in the genome. A bias can also result from optimal and non-optimal codons, which have stronger and weaker codon anti-codon interactions, respectively. Although association tests have been utilized before to identify genes associated with AD, it remains unclear how codon bias plays a role and if it can improve rare variant analysis. In this work, rare variants from whole-genome sequencing from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort were binned into genes using BioBin. An association analysis of the genes with AD-related neuroimaging biomarker was performed using SKAT-O. While using all synonymous variants we did not identify any genome-wide significant associations, using only synonymous variants that affected codon frequency we identified several genes as significantly associated with the imaging phenotype. Additionally, significant associations were found using only rare variants that contains an optimal codon in among minor alleles and a non-optimal codon in the major allele. These results suggest that codon bias may play a role in AD and that it can be used to improve detection power in rare variant association analysis.

Keywords: Alzheimer's disease; neuroimaging; codon bias; synonymous variant; BioBin; SKAT-O; rare variant analysis.

1. Introduction

Rare and low-frequency variants have a significant influence on the heritability of disease. Rare variants are often spurious; thus, it can be difficult to run an association test on an individual locus because it will be underpowered [1]. In order to overcome this issue, rare variants can be grouped or "binned" together based on prior biological knowledge related to the genetic etiology of the disease [2-4]. For instance, rare variants can be binned into genes, pathways, intergenic, conserved regions, or any other defined region of the genome [3, 5]. This strategy has several strengths: first it increases the detection power by aggregating association signals in the variants in the bin; secondly, it reduces the multiple testing burden by not testing every variant, thus increasing the power to detect a significant association. In addition to binning by a specific region, filtering for a specific type of variant, such as non-synonymous changes, have important benefits in addition to reducing the testing burden by focusing the association on variants that are more likely to influence the phenotype and provide easier interpretation of the results [6].

Synonymous mutations represent a change in the coding sequencing at the nucleotide level without changing the amino acid sequence. Since multiple codons code for the same amino acid, the genetic code is called "degenerate". It is likely that these characteristics of the genome is partially responsible for leading investigators to the assumption that synonymous mutations and variants have little to no impact on the protein, and are thus often dubbed "silent" without further investigation. However, it has been shown that different organisms prefer some codons over others and codon usage can also vary between genes in the same organism, suggesting there has been evolutionary pressure to optimize synonymous codons [7, 8]. Further investigation has demonstrated the many gene regulatory mechanisms by which codon bias can impart its affects such as splicing, RNA secondary structure, and translation [9, 10]. Moreover, synonymous variants have been implicated in a number of diseases including neurological, immune, cancer, blood-related, heart, and others [9]. The synonymous variants associated with these diseases are attributed to multiple mechanisms, therefore it will be important to study multiple forms of codon bias.

There are a number of ways in which codon usage can be biased and thus measured (Figure 1). For instance, the relative synonymous codon usage (RSCU) score represents the frequency for which the codon is used relative to other synonymous codons, thus providing a metric for determining whether a mutation replaces a more common codon with a rarer codon or vice versa [9, 11]. Substituting rare and common synonymous codons can affect translation and protein activity *in vitro* [12]. Both single cellular and multicellular eukaryotic organisms utilize codons that use rare and common tRNAs at the beginning and end of the gene, respectively, to impart control over translation rates [13]. Another means by which codon bias has been observed is through codon optimality. Some codons are more optimal than others by having stronger interactions with their cognate tRNA, or having more tRNAs available resulting in translation proceeding with less pausing and with higher fidelity, and in some cases affecting the stability of the mRNA [14].

In this study, we identified synonymous rare variants that have a functional impact on gene regulatory mechanisms in whole genome sequencing data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and then performed an association analysis of the functional synonymous variants with AD-related neuroimaging biomarker. AD is a progressive neurodegenerative disorder. Currently AD has no cure or preventive therapy. Genetic risk clearly

plays an important role in AD and neuroimaging has been used in concert with genetic information to improve our understanding in relation to how AD arises and how it can be potentially diagnosed.

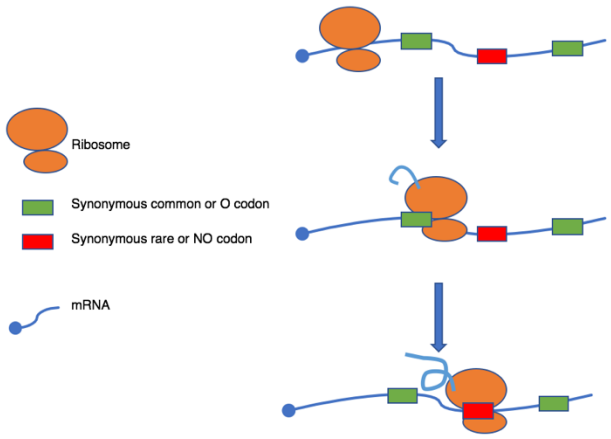


Fig. 1. Codon bias effects translation of mRNA. When the ribosome translates the mRNA, it will come into contact with both common, rare, optimal (O), and/or non-optimal codons (NO). After the ribosome starts translating (top), it may come into contact with a common or optimal codon. These codons are likely to lead to the ribosome continuing on the mRNA. Rare codons or non-optimal codons (bottom figure) may lead to the ribosome pausing or slowing down, possibly to allow for the protein to fold correctly.

2. Methods

2.1. *Study sample*

Data (whole genome sequencing and MRI imaging) used in this study were obtained from the ADNI database (http://adni.loni.usc.edu/). Samples were collected as described previously [6]. There was a total of 750 non-Hispanic Caucasian participants (425 were male and 325 female). The average age and years of education was 73.1 +/- 7.0 and 16.1 +/- 2.8 years, respectively.

2.2. *Neuroimaging analysis*

Pre-processed baseline 1.5T and 3T MRI scans were downloaded from the ADNI and T1-weighted brain MRI scans were processed using previously described automated MRI analysis technique, FreeSurfer software, which was used to extract mean bilateral entorhinal cortical thickness and total intracranial volume (ICV) [15]. Mean entorhinal cortical thickness, AD-related neuroimaging biomarker, was used as endophenotype for the association analysis.

2.3. *Variant annotation*

750 ADNI non-Hispanic Caucasian participants with baseline MRI scans and whole-genome sequencing (WGS) were used in this study. The VCFs containing the genomic information for

these 750 individuals were annotated using the variant effect predictor (VEP) software package. Using VEP, "synonymous" variants were selected using the filter function. Codons were then annotated as either optimal or non-optimal based on previous studies that characterized the codon anti-codon affinities [16-18]. The transition from optimal (O) to non-optimal (NO) was defined as the most common allele was O and the alternate allele was NO, and the reciprocal is true for the NO to O variants. Additionally, the relative synonymous codon usage (RSCU) score was calculated as:

$$RSCU = SN_c/N_a$$

N_c refers to the frequency of a specific codon
N_a is the frequency of the amino acid N_c codes for
S represents the number of synonymous codons for N_a

The codon frequencies for Homo sapiens were acquired from the Codon Usage Database (http://www.kazusa.or.jp/codon/). RSCU increasing was defined as the most common allele was in a codon with a lower RSCU score than the synonymous codon that the alternate allele produced. Whereas, for a decreasing RSCU score the most common allele was in a codon with a higher RSCU score than the synonymous codon that the alternate allele produced. Since only synonymous codons for the same amino acid were compared, the RSCU comparisons were effectively just comparing codon frequency in this work.

2.4. *BioBin analysis and association test*

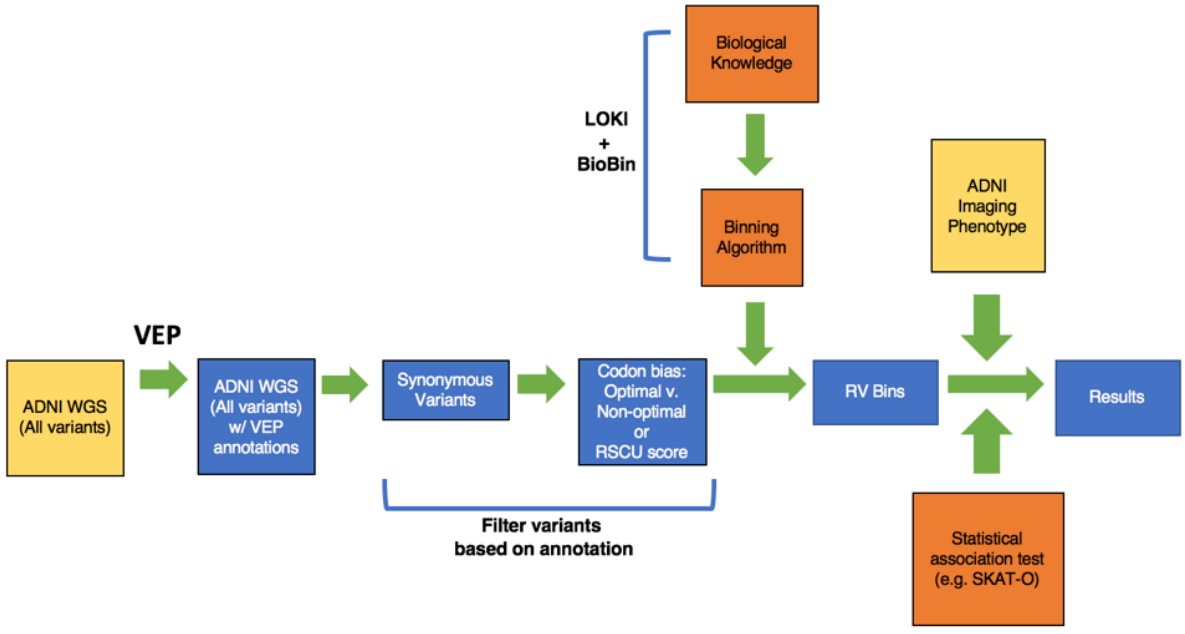


Fig. 2. Variants were from ADNI were annotated using VEP. Synonymous variants were then annotated using optimality and RSCU score. BioBin utilized the annotation from LOKI to bin rare variants into genes. SKAT-O was used to test genes for an association with the ADNI imaging phenotype.

BioBin was employed to group variants together by genic region [2-5]. BioBin uses gene annotations from LOKI (the library of knowledge integration), which contains information from several databases including but not limited to NCBI Entrez, UCSC Genome Browser, and Kyoto Encyclopedia of Genes and Genomes (KEGG). Only rare variants with minor allele frequency (MAF) less than 0.01 were binned and Madsen & Browning weighting was applied as previously described [5]. Association tests were performed using SKAT-O [19], adjusting for age, gender, years of education, intracranial volume (ICV) and MRI field strength as covariates. Although neuroimaging MRI scans from all ADNI participants included in the analysis were obtained from multiple sites, all sites followed the same ADNI MRI protocol and each raw scan was processed using a FreeSurfer pipeline at the Indiana University. Thus, the site was not included as a covariate since there is likely to be little site effects if any. The advantage of using SKAT-O is that it can utilize both dispersion or burden tests in order to detect a significant association [19]. P-values were adjusted for multiple tests using the p.adjust function in R, using the "FDR" method [20].

3. Results

Variants that represent a synonymous alteration were identified using VEP. The rare (MAF < 0.01) synonymous variants were then binned based on the genes they were located in, using BioBin (Figure 2). Each gene was required to have at least five variants across the cohort to be included in the analysis. Setting a minimum bin size establishes a more stringent threshold for finding an association. Additionally, by having fewer bins, there will be fewer tests performed, thus increasing the power to detect a significant association. previous studies have utilized a threshold when attempting to identify significant associations between genes and phenotypes of interest [21]. An association test was then performed between the genes and the imaging phenotype (entorhinal cortical thickness) using SKAT-O and corrected for multiple testing. When using all synonymous variants, there were no genes that reached genome-wide significance (FDR < 5%) nor were there any suggestive of being significant (FDR < 10%) (Table. 1).

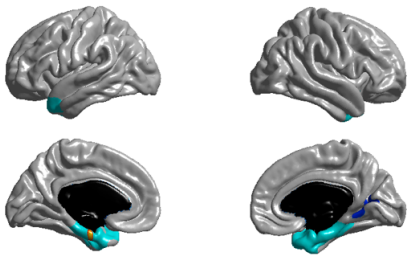
Table 1: Top 5 associations for all synonymous variants (11,236 genes total)

Gene	# of Loci	*p*-value	Corrected *p*-value
RHOB	4	1.14E-05	0.121
TMEM201	9	2.15E-05	0.121
MLST8	8	6.02E-05	0.212
MOB3B	4	7.56E-05	0.212
DTL	13	1.24E-04	0.278

However, using only synonymous variants with decreasing RSCU scores or increasing RSCU scores, we identified two (*MLST8* and *RHOB*) and six genes (*FLG2*, *CHD6*, *CD244*, *FLG-AS1*, *SERPINB5*, and *GTF3C1*) as significantly associated with entorhinal cortical thickness after multiple testing adjustment, respectively (Table 2 and Table 3). There are also two genes that were suggestive of being significant (Table 3). In addition, we performed a detailed unbiased whole-

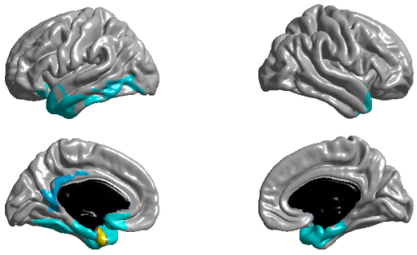
brain surface-based analysis using multivariate regression models to assess the effects of synonymous rare variants in *MLST8* and *RHOB* on whole-brain cortical thickness. First, we calculated a single polygenic risk score by collapsing all rare variants and counting minor alleles with a dominant genetic model. Figure 3 displays the results of the main effect of synonymous rare variants with decreasing RSCU scores in a surface-based whole-brain analysis. We identified highly significant clusters as associated with the risk scores in the entorhinal cortex after multiple comparison adjustment.

(a) MLST8



Cluster ID	# of vertices	Region	Corrected *p*-value
1	5479	Right entorhinal cortex	1.33E-05
2	4206	Left entorhinal cortex	0.00041

(b) RHOB



Cluster ID	# of vertices	Region	Corrected *p*-value
1	12360	Left entorhinal cortex	2.24E-06
2	4392	Right entorhinal cortex	0.000283
3	2452	Left temporal lobe	0.003054

Figure 3. **Surface-based whole-brain analysis results.** A whole-brain analysis of cortical thickness was performed to visualize the topography of genetic association (a) MLST8 and (b) RHOB in an unbiased manner. A threshold for statistical maps was set using a random field theory adjustment to a corrected significance level of p=0.05.

Table 2: Top 5 associations for synonymous variants with decreasing RSCU scores (8,066 genes total)

Gene	# of Loci	*p*-value	Corrected p-value
MLST8	6	4.10E-06	0.033
RHOB	4	1.14E-05	0.046
TRMT44	3	5.12E-05	0.122
RCC2	4	6.79E-05	0.122
MOB3B	4	7.56E-05	0.122

Table 3: Top associations for synonymous variants with increasing RSCU scores (4774 genes total)

Gene	# of Loci	*p*-value	Corrected *p*-value
FLG2	9	2.52E-05	0.039
CHD6	6	2.63E-05	0.039
CD244	2	3.48E-05	0.039
FLG-AS1	10	3.95E-05	0.039
SERPINB5	3	4.10E-05	0.039
GTF3C1	3	5.60E-05	0.045
GABRG1	2	1.07E-04	0.073
SRP72	4	1.39E-04	0.083

In addition, synonymous variants were separated into variants that introduce a non-optimal codon (O to NO) and those which introduce an optimal codon (NO to O). The results from each association analysis are represented in tables 4 and 5. Although no genes met genome-wide significance using the O to NO variants, the NO to O rare variants in five genes (*DTL*, *FLG2*, *SERPINB5*, *FLG-AS1*, and *ZNF599*) were significantly associated with entorhinal cortical thickness after multiple comparison adjustment (FDR < 5%).

Table 4: Top 5 associations for synonymous variants in non-optimal codons (O to NO; 8,401 genes total)

Gene	# of Loci	*p*-value	Corrected *p*-value
RHOB	4	1.14E-05	0.096
MLST8	6	6.78E-05	0.246
MSH2	13	1.54E-04	0.246
AKAP3	3	2.08E-04	0.246
RASGRF2	4	2.10E-04	0.246

Table 5: Top associations for synonymous variants in optimal codons (NO to O; 2,625 genes total)

Gene	# of Loci	*p*-value	Corrected *p*-value
DTL	4	4.45E-06	0.012
FLG2	9	2.68E-05	0.028
SERPINB5	3	4.10E-05	0.028
FLG-AS1	10	4.30E-05	0.028
ZNF599	2	9.42E-05	0.049
SRP72	4	1.39E-04	0.056
DLL4	3	1.50E-04	0.056

4. Discussion

Here we have performed an association analysis of synonymous rare variants from WGS with a functional impact on gene regulatory mechanisms with AD-related neuroimaging biomarker. Variants that represented synonymous changes between the codon of the major and minor alleles were first selected. BioBin was then used to count the number of variants per gene. No significant associations were identified using all synonymous variants. However, by focusing on specific groups, like those which affect frequency or optimality, significant associations were identified. In other words, by focusing on variants that are more likely to impact gene expression and possibly protein function, associations with genes that were previously undetected using all synonymous variants with AD neuroimaging biomarker were identified. Using all synonymous variants may be less likely to identify significant associations because it increases the likelihood of including synonymous variants that are in fact benign, thus drowning out the signal from synonymous variants that are more likely to be functional. Furthermore, by selecting only variants of a certain type, the number of tests performed was also reduced when compared to using all synonymous variants (compare table 1 to tables 2 through 5). With fewer association tests to run, the power to detect an association will also increase. The significant associations may provide useful insights into the biology of AD.

Several genes were associated with the imaging phenotype through variants that had a synonymous change which caused a change in relative codon usage. *MLST8* is a subunit of the *TOR* complex which is a key regulator of the cellular growth and survival in response to environmental cues [22-24]. Furthermore, it was found that a SNP near *MLST8* has a cis-regulatory effect on its expression in the brain in an age dependent manner [25]. Interestingly, gene expression also overlapped with genes that had epigenetic signatures that implicated it in Alzheimer's [25]. *RHOB* is a member of the Rho GTPase family of proteins responsible for modulating the actin cytoskeleton and gene expression [26]. *RHOB* is induced during neuro-trauma which is a known risk factor for AD [27-29]. *CHD6* is a chromatin remodeler that is a member of the *SNF2/RAD54* helicase protein family with no recognized link to AD [30]. Although expressed in most tissues, not much is known about the anti-sense *FLG-ASH1* transcript (genecards.org). Thus, multiple genes that have been previously connected to AD were identified here along with genes that have not previously been found to be associated with AD, or have little known about them at all. However, even though some genes had been previously associated with AD, this work presents a novel mechanism by which those associations may have arose.

Finding associations through the unique types of codon bias sheds light on possible mechanisms that may be at play. Generally, more simple eukaryotes like yeast often have a positive correlation between codon frequency and tRNA abundance, making codon bias easier to dissect, however human codon bias is more complicated [31]. Thus, it was surprising that significant associations were identified simply by using variants that either increased or decreased in frequency. It has been shown that changing codons from rare to common can impact translation and protein activity [12]. So while there may not be as easily an explainable relationship between rare and common codons in humans, they may still impact the expression of some genes. *RHOB* was significantly associated with the phenotype using frequency and almost significant using optimality, thus another possibility is that frequency could be a surrogate for other types of codon bias. Codon optimality offers a more refined characterization in terms of why the association may exist between these genes and the phenotype. In this study, significant associations were found among the variants that went from non-optimal to optimal. More optimal codons are expected to reduce pausing of the ribosome on the transcript [12]. It has been suggested that ribosome pausing may be important for allowing the protein to properly fold before the translation continues [12, 32]. Codon optimality can also affect mRNA stability [14]. Thus, the variants in the genes with an increase in optimal codons may have altered protein activity and/or expression levels which may eventually reach its way to impacting the AD related phenotype.

Although it was possible to detect significant associations between the imaging phenotype and binned rare variants, there are a number of ways the methodology can be improved for future work, and are thus limitations of the methodology as it currently stands. For instance, it has been illustrated that codon bias can be observed when comparing the codon usages among highly expressed lowly expressed genes [8, 33, 34]. Currently, the method employed here is not be able to address such complex mechanisms. Thus, future analysis could divide codon bias among highly or lowly expressed genes in cell types such as brain tissue. Another limitation of our study is the way in which we calculated codon bias, as there are other ways of measuring bias in terms of frequency

and optimality [31, 35], so these calculations should also be tested for their ability to improve signal strength in a rare variant association test for future work. It will be important for follow-up association tests to replicate these findings to illustrate that the results and conclusions are robust. Of course, another limitation is that without experimental follow up studies it cannot be suggested that these variants are causal. Thus, functional validation would be incredibly valuable to measure empirically how codon bias mediates the relationship between these genes and the imaging phenotype, AD, and/or neurological diseases in general. Codon bias has also been investigated with respect to cancer, where non-optimal codons mutations were enriched among multiple types of cancers [16]. Furthermore, synonymous variants have been associated with a variety of disease including, but not limited to, blood-related, bone, immune and other neurological disorders [9], suggesting the methods utilized in this work could contribute to our understand of a wide range of diseases. In summary, this work has illustrated variants that contribute to codon bias can be used to increase detection power. Moreover, codon bias is associated with an AD-related neuroimaging biomarker, suggesting synonymous variants can be used to explain the etiology of AD.

Acknowledgments

**Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Samples from the National Cell Repository for AD (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (AIG), were used in this study. Funding for the WGS was provided by the Alzheimer's Association and the Brin Wojcicki Foundation.

Additional support for data analysis was provided by NLM R01 LM012535, NIA R03 AG054936, NIA R01 AG19771, NIA P30 AG10133, NLM R01 LM011360, NSF IIS-1117335, DOD W81XWH-14-2-0151,

NCAA 14132004, NIGMS P50GM115318, NCATS UL1 TR001108, NIA K01 AG049050, the Alzheimer's Association, the Indiana Clinical and Translational Science Institute, and the IU Health-IU School of Medicine Strategic Neuroscience Research Initiative. This project was also funded, in part, under a grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

References

1. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* Am J Hum Genet, 2014. **95**(1): p. 5-23.
2. Moore, C.B., et al., *BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge.* BMC Med Genomics, 2013. **6 Suppl 2**: p. S6.
3. Moore, C.B., et al., *Using BioBin to explore rare variant population stratification.* Pac Symp Biocomput, 2013: p. 332-43.
4. Basile, A.O., et al., *Knowledge Driven Binning and Phewas Analysis in Marshfield Personalized Medicine Research Project Using Biobin.* Pac Symp Biocomput, 2016. **21**: p. 249-60.
5. Moore, C.C., et al., *A biologically informed method for detecting rare variant associations.* BioData Min, 2016. **9**(1): p. 27.
6. Kim, D., et al., *Knowledge-driven binning approach for rare variant association analysis: application to neuroimaging biomarkers in Alzheimer's disease.* BMC Med Inform Decis Mak, 2017. **17**(Suppl 1): p. 61.
7. Grantham, R., et al., *Codon catalog usage and the genome hypothesis.* Nucleic Acids Res, 1980. **8**(1): p. r49-r62.
8. Ikemura, T., *Codon usage and tRNA content in unicellular and multicellular organisms.* Mol Biol Evol, 1985. **2**(1): p. 13-34.
9. Sauna, Z.E. and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human disease.* Nat Rev Genet, 2011. **12**(10): p. 683-91.
10. Hunt, R.C., et al., *Exposing synonymous mutations.* Trends Genet, 2014. **30**(7): p. 308-21.
11. Sharp, P.M., T.M. Tuohy, and K.R. Mosurski, *Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.* Nucleic Acids Res, 1986. **14**(13): p. 5125-43.
12. Komar, A.A., T. Lesnik, and C. Reiss, *Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation.* FEBS Lett, 1999. **462**(3): p. 387-91.
13. Tuller, T., et al., *An evolutionarily conserved mechanism for controlling the efficiency of protein translation.* Cell, 2010. **141**(2): p. 344-54.
14. Presnyak, V., et al., *Codon optimality is a major determinant of mRNA stability.* Cell, 2015. **160**(6): p. 1111-24.
15. Saykin, A.J., et al., *Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans.* Alzheimers Dement, 2015. **11**(7): p. 792-814.
16. Wu, X. and G. Li, *Prevalent Accumulation of Non-Optimal Codons through Somatic Mutations in Human Cancers.* PLoS One, 2016. **11**(8): p. e0160463.
17. Watkins, N.E., Jr. and J. SantaLucia, Jr., *Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes.* Nucleic Acids Res, 2005. **33**(19): p. 6258-67.

18. Frenkel-Morgenstern, M., et al., *Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels.* Mol Syst Biol, 2012. **8**: p. 572.
19. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* Am J Hum Genet, 2012. **91**(2): p. 224-37.
20. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
21. Christophersen, I.E., et al., *Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation.* Nat Genet, 2017. **49**(6): p. 946-952.
22. Loewith, R., et al., *Two TOR complexes, only one of which is rapamycin sensitive, have distinct roles in cell growth control.* Mol Cell, 2002. **10**(3): p. 457-68.
23. Hara, K., et al., *Raptor, a binding partner of target of rapamycin (TOR), mediates TOR action.* Cell, 2002. **110**(2): p. 177-89.
24. Sarbassov, D.D., S.M. Ali, and D.M. Sabatini, *Growing roles for the mTOR pathway.* Curr Opin Cell Biol, 2005. **17**(6): p. 596-603.
25. Lu, A.T., et al., *Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum.* Nat Commun, 2016. **7**: p. 10561.
26. Hall, A., *Rho GTPases and the actin cytoskeleton.* Science, 1998. **279**(5350): p. 509-14.
27. Brabeck, C., et al., *Lesional expression of RhoA and RhoB following traumatic brain injury in humans.* J Neurotrauma, 2004. **21**(6): p. 697-706.
28. Conrad, S., et al., *Prolonged lesional expression of RhoA and RhoB following spinal cord injury.* J Comp Neurol, 2005. **487**(2): p. 166-75.
29. Magnoni, S. and D.L. Brody, *New perspectives on amyloid-beta dynamics after acute brain injury: moving between experimental approaches and studies in the human brain.* Arch Neurol, 2010. **67**(9): p. 1068-73.
30. Hall, J.A. and P.T. Georgel, *CHD proteins: a diverse family with strong ties.* Biochem Cell Biol, 2007. **85**(4): p. 463-76.
31. Quax, T.E., et al., *Codon Bias as a Means to Fine-Tune Gene Expression.* Mol Cell, 2015. **59**(2): p. 149-61.
32. Bali, V. and Z. Bebok, *Decoding mechanisms by which silent codon changes influence protein biogenesis and function.* Int J Biochem Cell Biol, 2015. **64**: p. 58-74.
33. Ikemura, T., *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system.* J Mol Biol, 1981. **151**(3): p. 389-409.
34. Ikemura, T. and H. Ozeki, *Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents.* Cold Spring Harb Symp Quant Biol, 1983. **47 Pt 2**: p. 1087-97.
35. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.* Nucleic Acids Res, 1987. **15**(3): p. 1281-95.

Building trans-omics evidence: using imaging and 'omics' to characterize cancer profiles

Arunima Srivastava
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: srivatava.1@osu.edu

Chaitanya Kulkarni
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: kulkarni.132@osu.edu

Parag Mallick*
Canary Center for Cancer Early Detection, Stanford University, 3155 Porter Dr., Palo Alto, CA, 94305
Email: paragm@stanford.edu

Kun Huang*
Department of Medicine, Indiana University School of Medicine, 340 W 10th St #6200, Indianapolis, IN 46202
Email: kunhuang@iu.edu

Raghu Machiraju*
Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210
Email: Machiraju.1@osu.edu

* *Corresponding authors*

Utilization of single modality data to build predictive models in cancer results in a rather narrow view of most patient profiles. Some clinical facets relate strongly to histology image features, e.g. tumor stages, whereas others are associated with genomic and proteomic variations (e.g. cancer subtypes and disease aggression biomarkers). We hypothesize that there are coherent "trans-omics" features that characterize varied clinical cohorts across multiple sources of data leading to more descriptive and robust disease characterization. In this work, for 105 breast cancer patients from the TCGA (The Cancer Genome Atlas), we consider four clinical attributes (AJCC Stage, Tumor Stage, ER-Status and PAM50 mRNA Subtypes), and build predictive models using three different modalities of data (histopathological images, transcriptomics and proteomics). Following which, we identify critical multi-level features that drive successful classification of patients for the various different cohorts. To build predictors for each data type, we employ widely used "best practice" techniques including CNN-based (convolutional neural network) classifiers for histopathological images and regression models for proteogenomic data. While, as expected, histology images outperformed molecular features while predicting cancer stages, and transcriptomics held superior discriminatory power for ER-Status and PAM50 subtypes, there exist a few cases where all data modalities exhibited comparable performance. Further, we also identified sets of key genes and proteins whose expression and abundance correlate across each clinical cohort including (i) tumor severity and progression (incl. GABARAP), (ii) ER-status (incl.

ESR1) and (iii) disease subtypes (incl. FOXC1). Thus, we quantitatively assess the efficacy of different data types to predict critical breast cancer patient attributes and improve disease characterization.

1. Introduction

Recent advances in whole slide imaging digitization and compilation in the form of the TCGA compendium[1], alongside matching high throughput profiling data has opened up many avenues for modeling different facets of an experiment. Histopathological images have been very successful in predicting clinical outcomes in the context of various TCGA cancers[2]. Similarly, transcriptomics and proteomics profiling has showcased distinct discriminatory power when modeling cancer subtypes for the purpose of targeted drug therapies and biomarker excavation[3]. The purpose of this work is to comprehensively compare the characterizing abilities of these three modalities of data across varying types of clinical cohorts, when traditionally, each are analyzed in the context of specific attributes only (**Figure 1**). The overarching goal is to (a) qualitatively compare the predictive power of each type of data modality while modeling different patient attributes and (b) find coherent biological signatures and features (e.g. driver genes and subtype-specific protein biomarkers) that provide a framework for evidence based depiction of each attribute cohort.

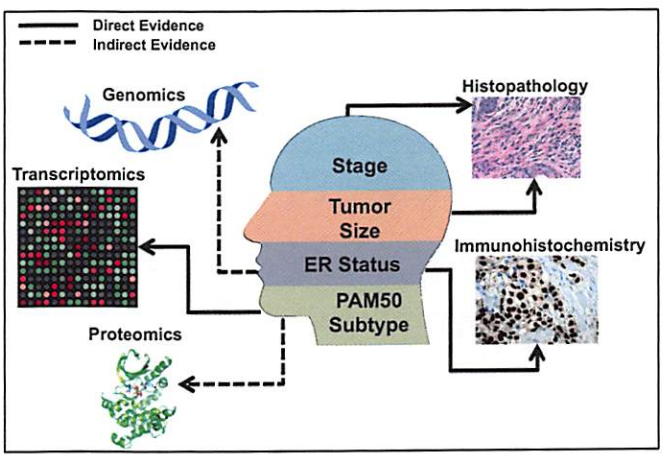


Figure 1. Various patient attributes and the data modalities that present validation or evidence for them.

To achieve the above goals, we utilize 105 patients from the TCGA-BRCA (Breast Cancer) dataset, which contain a clinically diverse set of patients and multiple levels of data. Namely, histopathological (H-hematoxylin and E-eosin stained) tissue images of the tumor region, transcriptomics (RNA-Seq) data and proteomics (Isobaric tag for relative and absolute quantitation –iTRAQ) data are available for all the patients categorized in this study. The data modalities were used to model prediction for the four (4) clinical attributes, namely, AJCC (American Joint Committee on Cancer) staging, tumor staging, ER-status and lastly PAM50 panel breast cancer subtypes.

Data modalities and biological models

Histology Images - There has been a recent upsurge in publications describing the utilization of H/E whole slide images (WSIs) to predict clinical outcomes. Many of these present the use of a deep learning approach on tiles of histology images, as opposed to semantic image features, to build classification models[2] (**supplementary material**). There are many commonalities between the inferences that can be drawn from morphological features identified by CNNs (atypical shape of cells, disintegration of tissue architecture and visible stromal invasiveness) and the hallmarks of cancer (cells resisting apoptosis, metastatic tendencies and limitless replicative potential). We

chose GoogLeNet[4] inspired CNNs with histology data to classify the dataset into clinical cohorts as listed above.

Transcriptomics - From the four (4) patient stratifications that we aimed to assess, the molecular subtypes (ER-Status and PAM50 subtypes) are natively associated with transcript level variations across patients. Histology images remain the preferred and proven method for largely predicting clinical status, while a few analyses using transcriptomics, modeling staging and similar attributes have been performed with some success[5]. Transcript level expression has been primarily utilized to subtype patients, extract biomarkers and understand signaling and regulation using data from microarrays and RNA-Seq[6] using traditional methods of analysis that are well understood and widely implemented[7].

Proteomics - High throughput proteomics experiments have recently gained popularity as they enable the study of regulatory mechanisms in cancer. Combining proteomics and transcriptomics, models disease more effectively than using these datasets independently and thus proteogenomics analysis has been utilized for gauging better subtypes of disease, associating genomic variations with signaling and isolating disease driver genes and proteins[3,8].

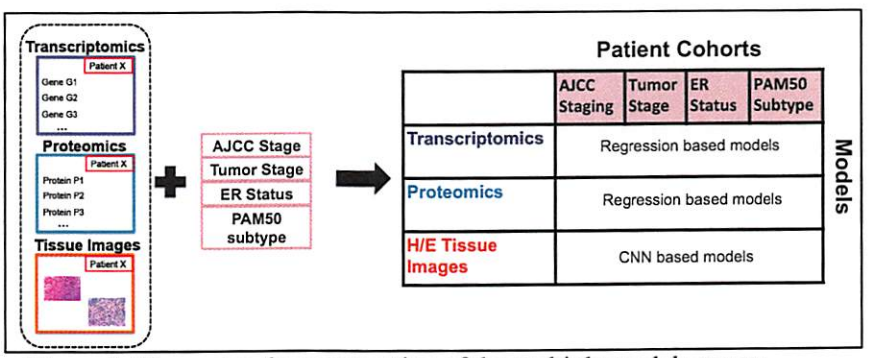


Figure 2. Showcases the construction of the multiple models across different data modalities. Comparing the performance of these models quantifies each data level's utility for identification of characteristic features for each of the four clinical attributes.

Our aim was to build models with "best practice" methods for each data modality, and for the same set of patients, train, test and predict the different clinical attributes. The goal was to compare utility of data types, using traditional, widely used methods for each data type. For the best performing models for each data modality, we compare the corresponding results for all four clinical cohorts (**Figure 2**). As expected, histology based models are ideal for predicting clinical outcomes and genomic data outperforms all other data types when predicting molecular subtypes. However, we also observe comparable performance metrics amongst non-traditional data models in some cases, which allow us to reinforce the characterization of cohorts across the scales. For instance, our results show that there exist salient genes and proteins that are able to distinguish between AJCC stages and tumor stages relatively successfully. These included genes that are verifiably associated with disease severity and tumor progression – e.g. GABARAP, CKS2 and CDH1. A more comprehensive list of molecular markers is included in subsequent sections. While image-based classifiers outperform them, they can still help build trans-omic evidence and reinforce disease profiles. In summary, this work explores the likelihood of finding evidence in data types that include trans-omic indicators. Additionally it forms a critical

foundational layer for future integrative imaging genomics models by highlighting utility of data types and extracting meaningful features.

2. Methods and Materials

Below we describe, a workflow for (a) building data driven predictive models (Section 2.2), (b) evaluating performance of all the models for comparison (Section 2.3) and (c) extracting key driving features from successful models (Section 2.4). Scripts used to build and evaluate models as well as find key features are included in the **supplementary material**.

2.1. *TCGA Breast Cancer multi-level dataset*

The TCGA breast cancer study[9] is a well-characterized and thoroughly comprehensive experimental study of breast invasive carcinoma[5,10]. A total of 105 patients, for whom all three (3) types of data were available were utilized in this work. The three data modalities include RNA-Seq mRNA expression, selected parts of H/E stained whole slide tissue images and proteomic abundance from mass spectrometry (LC-MS/MS –Liquid Chromatography-Mass Spectrometry) analysis. Further, each of these 105 patients have associated clinical profiles that detail their AJCC Stage (Stage I, II, III and IV), Tumor Stage (T1, T2, T3 and T4), ER status (+/-) and finally PAM50 mRNA subtypes of the cancer (Luminal A, Luminal B, HER2 enriched and Basal like).

2.2. *Different data modalities and subsequent model construction*

2.2.1. *Histology images and CNN based classification models*

As mentioned previously, we utilized the GoogLeNet (2014) Convolutional Neural Network to construct a classifier using the tiles of histology images that were acquired from the Genomic Data Commons (GDC)[11]. As is typical with neural networks, multiple parameters (structure of input, number of network layers, pre-training data, etc.) are instrumental to the eventual performance. To this end, we constructed classifiers using five (5) different versions of the standard GoogLeNet (with and without pre-training data, larger input tile sizes, spatially invariant input tiles and only epithelial regions as training data) and evaluated the performance for each across all clinical cohorts. See **supplementary material** for more details including key transformative features of the different CNN models.

2.2.2. *Transcriptomics data and modeling*

Transcript expression RNA-Seq data (percentile-normalized version) was accessed from the UCSC Xena (http://xena.ucsc.edu/)[12]

Project. Superfluous gene names or transcript measurements (tagged as NA) were removed and Z-transform normalization of the resulting data was performed. The final data matrix contained the normalized transcript measures for 20501 unique genes across 105 patients, and the corresponding four (4) dimensional clinical attribute vectors. To build predictive models for multi-class labels, we employed multinomial log-linear regression (using function "*mutinom*" from R-package "*nnet*"[13]) on different subsets of the transcriptomics data. To build these regression models from the transcriptomics dataset, we find subsets of the genome wide dataset that potentially represented a variety of cohorts (histology and molecular profiling based). **Table 1** presents the gene lists[11,14–17] which helped extract those data subsets and a summation of what patient profiles they help characterize (**Figure 3**).

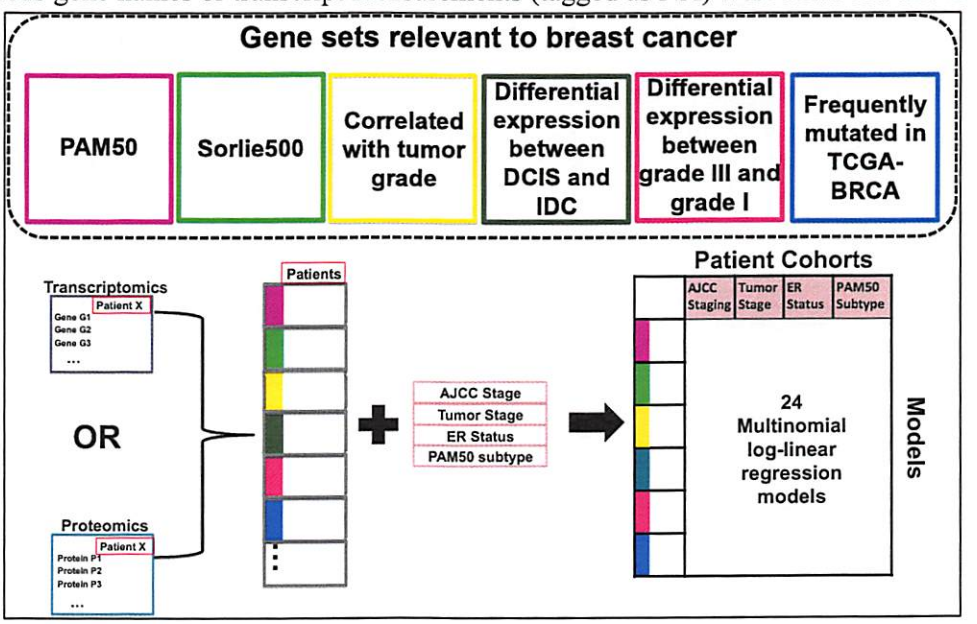


Figure 3. Workflow detailing the construction of regression models from transcriptomics and proteomics data. 6 different models are constructed for each clinical cohort (AJCC Staging, Tumor Staging, ER Status and PAM50 Subtypes). These 6 different models are constructed by extracting subsets of the proteogenomic data for gene sets relevant to various different attributes of breast cancer. Using these data subsets, and clinical cohort labels, we construct multinomial log-linear regression predictive models.

Table 1. Gene lists proven to be relevant to different breast cancer patient attributes

Gene Lists	Patient Attribute Characterization
PAM50	50 genes, defining expression based intrinsic subtypes of breast cancer
Sorlie500	500 intrinsic unique genes, defining refined subtypes of breast cancer
Top 200 genes related to tumor grade	200 genes, highly and significantly correlated with tumor grade
Most frequently mutated in TCGA-BRCA	Top 20 most frequently mutated genes in the TCGA-BRCA project patients, as reported by data analysis performed by the GDC portal
Differentially expressed between DCIS-S and IDC-S	305 genes classifying tumor invasion, presenting significant differential expression between stroma of ductal carcinoma *in situ* and invasive ductal carcinoma
Differentially expressed between grade III and grade I	620 genes, found differentially expressed between stroma of tumor grade III and tumor grade I patients, classifying the extent of abnormality of the tumor

2.2.3. *Proteomics data and modeling*

Normalized proteomics abundance (iTRAQ - Isobaric tag for relative and absolute quantitation) for the 105 patients considered for this work were extracted from supplementary data provided in the work of Mertins Et al.[3] (Supplementary Table 03 – Global Proteome G1). The details of data normalization and pre-processing are described in the supplementary information of the publication cited above. Redundant profiles or samples were removed from the normalized dataset mimicking the technique utilized previously for transcriptomics data. The final data matrix of normalized protein abundances contained 6386 unique proteins across the relevant 105 patients. Similar to the analysis technique engaged for the transcriptomics dataset (regression modeling with biologically driven subsets of data), we built models (**Figure 3**) with subsets of the proteomic dataset in the context of the four (4) selected clinical attributes.

2.3. *Metrics of performance for each model*

Standard performance metrics of precision, recall and F-score for all versions of the models constructed from each data type are used. As described above, multiple variations of models from each data modality were constructed, and the ones that performed best, within a single data level, were chosen for performance comparison across data levels. For proteogenomics datasets, we calculated the performance measures using cross-validation and partitioning 70% of the data for training, using function "*prediction*" from the R-package "*ROCR*"[18]. For the CNN based models, suitable programmatic metrics[19] were utilized for the same. Specifics of model evaluation, including the details regarding division of data to training and testing sets, are further detailed in the **supplementary material**.

2.4. *Isolating the key biological signature features characterizing each cohort*

We aim to identify key features from all data modalities, which project the highest discriminatory power for each clinical attribute.

Transcriptomics and Proteomics - We select the model that predominantly outperforms all other proteogenomic models and extract genes and proteins that contain high predictive power for each of the patient attributes. We utilize the RFE (Recursive Feature Elimination)[20,21] workflow (see **supplementary material**) to find subsets of features for each model that are critical for guiding the classification (using function "*rfe*" provided by the R-package "*caret*"[20]). Due to selection bias, we execute the algorithm multiple times, and extract the features that are consistently deemed important for the model.

Histology Images –While CNNs produce weighted feature maps used to drive classification, it is challenging to map them to verifiable histology image features. For the purposes of this work, we assess genomic features only, but propose to expand this study in the future to utilize CNN saliency maps in an effort to extract relevant feature images from CNNs.

3. Results and Discussion

In this section we report performance metrics for models generated from all data types, and the key genes and proteins that drive successful classification. All other relevant information (gene lists, modeling parameters, R data frames) is detailed in the **supplementary material**.

3.1. *Classification of patient attributes using histopathological tissue images using CNNs*

We compared the performance metrics of precision, recall and F-score across all the different versions of the image classification CNNs. We observed that the best predictive model for classifying patients to AJCC stages was borne from the "*Inception (v3 2015) with pre-training version of the CNN*" (0.52 F-score), whereas for tumor stages the "*Inception (v3 2015) with pre-training and boosted data*" (0.50 F-score) outperformed the other models. For clinical cohorts (ER-Status and PAM50 subtypes) traditionally related to genomic data, the best predictive models compared to all other CNN based models were again the "*Inception (v3 2015) with pre-training and boosted data*" (0.64 F-score) and "*Inception (v3 2015) with no pre-training*" (0.35 F-score) respectively. To summarize, it is not surprising that tissue images predict the AJCC and tumor staging as well as ER-status drastically better than PAM50 subtypes. Additionally, the varying parameters and boosting data techniques had little to no discernible effect to the quality of the models produced.

3.2. *Modeling and classification using proteogenomics*

We now describe the results and the attained efficacy of the multinomial log-linear regression models built with transcriptomics and proteomics measures. Examples of the genes and proteins identified as critical to the predictive model (using RFE analysis) are listed in **Figure**

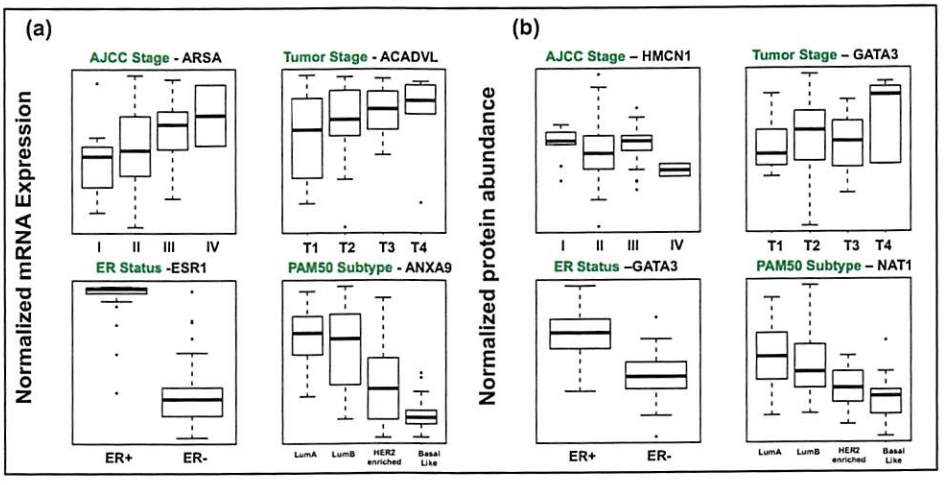


Figure 4. (a) Boxplots presenting gene expression (percentile normalized) of example top ranking critical features (genes) derived from "best" models for each clinical cohort using transcriptomics data. (b) Boxplots presenting protein abundance (mixture model-based normalization) of example top ranking critical features (proteins) derived from "best" models for each clinical cohort using proteomics data.

4, and examples with corresponding literature evidence are listed below. Performance metrics for all the models and literature evidence for key genes and proteins as identified by RFE analysis is listed in **supplementary material**.

3.2.1. *Transcriptomics model results*

AJCC and Tumor Stage-The performance metric indicates that the transcriptomics model generated with *Top 200 genes related to tumor grade* outperform all other models when predicting both AJCC Stage and Tumor Stage (0.43 and 0.4 F-score respectively). The RFE approach isolates 11 genes that are critical to the model predicting AJCC Stage and 10 genes for the tumor stage model. All 21 of the extracted key genes from the above two models are part of a verified breast cancer grading signature[22]. A large number of these genes are shown to be significantly associated with clinical outcome, a result that can be easily extended to stage and tumor size association (analyzed using the PRECOG database[23] (https://precog.stanford.edu), which associates gene expression to clinical outcome, tabulated in **supplementary materials**). These genes include CKS2, overexpression of which is known to be involved in tumor development by blocking cell cycle S-phase signaling which causes cells to abnormally proliferate in stress conditions[24] and ARPC1A, associated with poor prognosis in many cancers[25].

ER-Status-The classification between patients with differing ER-Status is performed perfectly by the *PAM50* set of intrinsic genes (0.925 F-score). Gene expression of ESR1 and other co-expressed genes, included within the *PAM50* set are widely known as having high discrimination for ER-positive status[26]. The RFE guided list of driver genes for this model lists 7 key genes including ESR1. Investigating these critical genes using Gene Set Enrichment Analysis (GSEA)[27], we observe that 6 out of the 7 identified genes are known to be up regulated specifically in ESR1 positive breast cancer tumors (false discovery rate of $2.2\ e^{-11}$)[28]. This gene set includes driver genes such as FOXA1, which is involved with ESR1 for regulation[29], NAT1, known to be commonly overexpressed in ER-alpha (+) tumors[30], and MDM2, which regulates ER-alpha and estrogen responsiveness in breast cancer cells[31].

PAM50 mRNA Subtype-While one would expect the data subset from *PAM50* gene set to perform ideally when predicting the PAM50 subtype status, *Sorlie500* transcriptomics subset in fact presents a superior classifier when performing classification to PAM50 mRNA subtypes (0.65 F-score versus 0.59 for the *PAM50* transcriptomics subset). A total of 13 genes from the *PAM50* gene set are included in the *Sorlie500* gene set, along with other intrinsic subtyping genes. Acquisition, normalization, pre-processing and composition of data all are known to fluctuate the results of PAM50 mRNA based subtyping[32]. We hypothesize that while highly relevant (known biomarkers) genes like ESR1, ERBB2, FOXA1, FOXC1[33] etc. (included in both *PAM50* and *Sorlie500* gene sets, and identified as highly important for this model) drive the subtype classification, additional intrinsic genes encompassed in the *Sorlie500* set further help stratification. The hypothesis is confirmed by the RFE analysis, which selected 23 highly important genes to the model, 7 of which also belonged to the *PAM50* gene set and included known discriminants such as ESR1 and ERBB2[34].

3.2.2. *Proteomics model results*

AJCC, Tumor Stage and ER-Status-All three of these cohorts are best predicted by proteomics data subset of the *Most frequently mutated genes*. The proteomics data subsets presented F-score measures of 0.43, 0.45 and 0.80 for the AJCC stage, tumor stage and ER-status classification respectively, compared to 0.43, 0.39 and 0.92 F-score observed for the best models from the transcriptomics dataset. The RFE analysis further isolates a list of two (2) (e.g. DST, found to be breast cancer tumor progression suppressor[35]), three (3) (e.g. disease severity biomarker CDH1[36]) and four (4) (e.g. GATA3, strongly associated with ER-Status[37]) proteins respectively that are the main drivers of these models.

PAM50 mRNA Subtype - The subset of *PAM50* proteins and their corresponding normalized abundances outperform all other models when attempting to classify the PAM50 mRNA subtype of the patients using the proteomics data. All performance measures showcase that this model is still lacking in predictive power as compared to the best *Sorlie500* model derived from the transcriptomics data (0.65 F-score for the transcriptomics *Sorlie500* model versus 0.54 for proteomics *PAM50* model). This may be caused, potentially due to a disconnect between various transcripts and corresponding proteins due to post-transcriptional regulation. The RFE analysis outlines 10 proteins that drive the classification for this model and they include well known breast cancer relevant proteins such as ESR1, ERBB2 and RRM2 (associated with basal proliferative tumors[38]). Four (4) of the key genes from the corresponding list derived from the transcriptomics model (ESR1, FOXA1, ERBB2 and NAT1) are included in the key proteins list as well.

3.3. *Comparison of all data modalities*

As previously mentioned, the best performing models, for each data modality, were ultimately compared across all three data levels for each cohort (**Figure 5**). Histology image based models outdid the transcriptomics and proteomics models while predicting the AJCC and tumor stage. Both ER-Status and PAM50 subtypes were ideally characterized by transcriptomics data, with proteomics data models performing second best. While this was in accordance with our expectations, there were a few points of interest within these comparative results. For instance, while imaging models were the most precise in classifying

Precision	AJCC Stage	Tumor Stage	ER Status	PAM50 Subtype
Best Imaging Model	**0.513149**	**0.497086**	0.639785	0.365476
Best Transcriptomics Model	0.472848164	0.467049688	**0.92923747**	**0.704195609**
Best Proteomics Model	0.463979274	0.462117632	0.81571361	0.582639235
Recall	**AJCC Stage**	**Tumor Stage**	**ER Status**	**PAM50 Subtype**
Best Imaging Model	**0.596441**	**0.555039**	0.645003	0.356984
Best Transcriptomics Model	0.4175	0.381875	**0.9246875**	**0.6553125**
Best Proteomics Model	0.518125	0.49	0.804375	0.5409375
F-score	**AJCC Stage**	**Tumor Stage**	**ER Status**	**PAM50 Subtype**
Best Imaging Model	**0.524801**	**0.501531**	0.642337	0.357809
Best Transcriptomics Model	0.431051241	0.399065605	**0.92426716**	**0.653987855**
Best Proteomics Model	0.438919713	0.453932236	0.80132172	0.541250002

Figure 5. Precision, recall and F-score across all clinical cohorts for the best performing models for each data modality respectively. Standard deviations for measures calculated using cross-validation are included in supplementary materials.

patient staging, it is important to note that both transcriptomics and proteomics models were not drastically less precise (0.51 vs ~0.47). This indicates that there exist features within genomic variations, which have the discriminatory power to effectively characterize histology-based staging. These comparisons not only quantified the utility of each data type in modeling various clinical subtypes, they also identified previously unexplored associations between data types and patient profiles.

4. Conclusions and Future Work

In this study we showcased the different data modalities and how they are utilized for modeling different facets of cancer. We employed "best practice" modeling techniques for three different data modalities to predict four (4) varied attributes of breast cancer patients in the TCGA compendium. We quantified the predictive power of data modalities for different aspects of patient profiles. Finally, key genomic features critical to all different clinical attributes were identified and validated using existing literature. We wish to expand this work by exploring the various pathways (e.g. FOXA1/ESR1/GATA3 interacting pathway) that include the identified key genes and proteins, in the context of various different cohorts. Additionally, we wish to perform analysis to explain what causes the differences between predictive powers across data modalities (e.g. transcriptomics and proteomics models). Further, we wish to utilize more sophisticated methods, including more robust regression to account for the structure of data, for each data modality to perform better stratification of patient subtypes and construct robust patient similarity frameworks using these trans-omic evidences.

Supplementary Materials - https://github.com/arunima2/PSB_2018

References

1. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol.* 2015;1A:A68-A77. doi:10.5114/wo.2014.47136.
2. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using Convolutional Neural Networks. Sapino A, ed. *PLoS One.* 2017;12(6):e0177544. doi:10.1371/journal.pone.0177544.
3. Mertins P, Mani DR, Ruggles K V., et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature.* 2016;534(7605):55-62. doi:10.1038/nature18003\rhttp://www.nature.com/nature/journal/v534/n7605/abs/nature18003.html#supplementary-information.
4. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol 07-12-June. ; 2015:1-9. doi:10.1109/CVPR.2015.7298594.
5. Yao F, Zhang C, Du W, Liu C, Xu Y. Identification of gene-expression signatures and protein markers for breast cancer grading and staging. *PLoS One.* 2015;10(9). doi:10.1371/journal.pone.0138213.
6. Bertucci F, Finetti P, Rougemont J, et al. Gene expression profiling identifies molecular

subtypes of inflammatory breast cancer. *Cancer Res.* 2005;65(6):2170-2178. doi:10.1158/0008-5472.CAN-04-4115.
7. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:13. doi:10.1186/s13059-016-0881-8.
8. Beretov J, Wasinger VC, Millar EKA, Schwartz P, Graham PH, Li Y. Proteomic Analysis of Urine to Identify Breast Cancer Biomarker Candidates Using a Label-Free LC-MS/MS Approach. Aboussekhra A, ed. *PLoS One.* 2015;10(11):e0141876. doi:10.1371/journal.pone.0141876.
9. Koboldt DC, Fulton RS, McLellan MD, et al. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61-70. doi:10.1038/nature11412.
10. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. *J Biomed Inform.* 2015;56:220-228. doi:10.1016/j.jbi.2015.05.019.
11. Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med.* 2016;375:1109-1112. doi:10.1056/NEJMp1002530.
12. Goldman M, Craft B, Swatloski T, et al. The UCSC cancer genomics browser: Update 2015. *Nucleic Acids Res.* 2015;43(D1):D812-D817. doi:10.1093/nar/gku1073.
13. Venables WN, Ripley BD. Modern Applied Statistics with S. *Issues of Accuracy and Scale.* 2002;(March):868. doi:10.1198/tech.2003.s33.
14. Bernard PS, Parker JS, Mullins M, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370.
15. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A.* 2003;100(14):8418-8423. doi:10.1073/pnas.0932692100.
16. Ma X-J, Dahiya S, Richardson E, Erlander M, Sgroi DC. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.* 2009;11(1):R7. doi:10.1186/bcr2222.
17. Ma X-J, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A.* 2003;100(10):5974-5979. doi:10.1073/pnas.0931261100.
18. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics.* 2005;21(20):7881. http://rocr.bioinf.mpi-sb.mpg.de.
19. Bowles M. Machine learning in Python. *Igarss 2014.* 2014;(1):1-5. doi:10.1007/s13398-014-0173-7.2.
20. from Jed Wing MKC, Weston S, Williams A, et al. caret: Classification and Regression Training. 2015. http://cran.r-project.org/package=caret.
21. Zeng X, Chen Y-W, Tao C, Alphen D Van. Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition. *2009 Fifth Int Conf Intell Inf Hiding Multimed Signal Process.* 2009:1205-1208. doi:10.1109/IIH-MSP.2009.145.
22. Ma XJ, Sgroi DC, Erlander MG. Grading of breast cancer. 2017. https://www.google.com/patents/US20170073767.
23. Gentles AJ, Newman AM, Liu CL, et al. The prognostic landscape of genes and infiltrating

immune cells across human cancers. *Nat Med.* 2015;21(8):938-945. doi:10.1038/nm.3909.
24. Liberal V, Martinsson-Ahlzén H-S, Liberal J, et al. Cyclin-dependent kinase subunit (Cks) 1 or Cks2 overexpression overrides the DNA damage response barrier triggered by activated oncoproteins. *Proc Natl Acad Sci U S A.* 2012;109(8):2754-2759. doi:10.1073/pnas.1102434108.
25. Lomakina ME, Lallemand F, Vacher S, et al. Arpin downregulation in breast cancer is associated with poor prognosis. *Br J Cancer.* 2016;114(5):545-553. doi:10.1038/bjc.2016.18.
26. Iwamoto T, Booser D, Valero V, et al. Estrogen Receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. *J Clin Oncol.* 2012;30(7):729-734. doi:10.1200/JCO.2011.36.2574.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.
28. Doane a S, Danso M, Lal P, et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene.* 2006;25(28):3994-4008. doi:10.1038/sj.onc.1209415.
29. Chaudhary S, Krishna B, Mishra S. A novel FOXA1/ESR1 interacting pathway: A study of Oncomine™ breast cancer microarrays. *Oncol Lett.* June 2017. doi:10.3892/ol.2017.6329.
30. Abba MC, Hu Y, Sun H, et al. Gene expression signature of estrogen receptor alpha status in breast cancer. *BMC Genomics.* 2005;6(1):37. doi:10.1186/1471-2164-6-37.
31. Kim K, Burghardt R, Barhoumi R, Lee S ook, Liu X, Safe S. MDM2 regulates estrogen receptor α and estrogen responsiveness in breast cancer cells. *J Mol Endocrinol.* 2011;46(2):67-79. doi:10.1677/JME-10-0110.
32. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst.* 2015;107(1):357. doi:10.1093/jnci/dju357.
33. Han B, Bhowmick N, Qu Y, Chung S, Giuliano AE, Cui X. FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene.* 2017;(February):1-7. doi:10.1038/onc.2017.48.
34. Kao J, Salari K, Bocanegra M, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One.* 2009;4(7). doi:10.1371/journal.pone.0006146.
35. Lee S, Stewart S, Nagtegaal I, et al. Differentially expressed genes regulating the progression of ductal carcinoma in situ to invasive breast cancer. *Cancer Res.* 2012;72(17):4574-4586. doi:10.1158/0008-5472.CAN-12-0636.
36. Memni H, Macherki Y, Klayech Z, et al. E-cadherin genetic variants predict survival outcome in breast cancer patients. *J Transl Med.* 2016;14(1). doi:10.1186/s12967-016-1077-4.
37. Voduc D, Cheang M, Nielsen T. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer Epidemiol Biomarkers Prev.* 2008;17(February):365-373. doi:10.1158/1055-9965.EPI-06-1090.
38. Bertucci F, Finetti P, Cervera N, et al. How different are luminal A and basal breast cancers? *Int J Cancer.* 2009;124(6):1338-1348. doi:10.1002/ijc.24055.

PRECISION MEDICINE: FROM DIPLOTYPES TO DISPARITIES TOWARDS IMPROVED HEALTH AND THERAPIES

DANA C. CRAWFORD

Population and Quantitative Health Sciences, Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, 44106 USA
Email: dana.crawford@case.edu

ALEXANDER A. MORGAN

Khosla Ventures, Menlo Park, CA, 94025 USA
Email: alex@khoslaventures.com

JOSHUA C. DENNY

Vanderbilt University Medical Center, Nashville, TN 37203 USA
Email: josh.denny@vanderbilt.edu

BRUCE J. ARONOW

Center for Computational Medicine, Cincinnati Children's Hospital Medical Center and the University of Cincinnati, Cincinnati, OH 45229 USA
Email: bruce.Aronow@cchmc.org

STEVEN E. BRENNER

University of California, Berkeley, CA 94720-3012 USA
Email: brenner@compbio.berkeley.edu

Precision medicine research efforts both in basic science discovery and clinical implementation are well underway and promise to provide individualized preventions and treatments, improving overall health care delivery. To achieve these goals, advances in data capture and analysis are needed spanning different types of 'omic and clinical data. The efforts to enhance precise treatments for all may accentuate healthcare disparities unless specific challenges are identified and addressed. This session of the 2018 Pacific Symposium on Biocomputing presents the latest developments in this transdisciplinary research space of genomics, medicine, and population health.

1. Introduction

Precision medicine is often described as providing a patient the optimal tailored treatment the first time as opposed to standard treatment or trial and error. Precision medicine has arguably been practiced since the emergence of modern medicine[1]. The definition of precision medicine has recently evolved from the 20th century's addition of genetic variants that impact drug response[2-4] to the 21st century's recognition that lifestyle, social, and environmental factors interact with the patient's genome[5], impacting a range of health consequences including conferring risk to disease and differential response to treatmen[6].

Although the concept of precision medicine is not new, the implementation of precision medicine

is relatively nascent. Recent key advances in genomics[7] and the emergence of electronic health records (EHRs) in part through the HITECH Act[8, 9] make it feasible to put precision medicine to practice[10]. Despite these advances, the adoption of genomic data in routine clinical practice has been slow likely due to a variety of reasons including costs or inconsistent reimbursement policies (such as changing and limited coverage by Medicare) and access to genomic testing[11].

Another major driver behind the lack of implementation is the lack of data. The lack of data is distributed across active research areas that fuel precision medicine including basic discovery and functional or biological characterization[12]. Inherent in these broad areas of research are needed expansions of informatics approaches to extract health-informative data from various big data sources coupled with advances in novel statistical and computational methodology to integrate and interpret disparate data types for predictive modeling or further discovery and functional characterization[13, 14]. Beyond the traditional boundaries of basic research are data needs related to precision medicine implementation ranging in topics from cost-effectiveness and clinical utility to clinical decision support[15, 16], among other topics[17].

Precision medicine research and implementation programs will undoubtedly evolve rapidly with the recent infusion of support from the previous administration[18]. The initial structure and recruitment sites for the Precision Medicine Initiative Cohort Program, an ambitious effort to ascertain one million participants in the United States for precision medicine research, have been established as of 2016, and recruitment and collection of these data are anticipated this year. Of note is the Program's emphasis on racial/ethnic, geographic, and economic diversity, variables that continue to be underrepresented in the basic discovery studies[19-21] despite their known influence on human health[22, 23]. The absence of these data may lead to misdiagnoses and missed opportunities for many patients not represented in discovery studies[24]. The new Cohort endeavor, now known as "All of Us," is expected to both generate new data but also to inspire new regulations and guidelines based on safety and bioethics[25-27]. Given innovative biocomputing is the engine of precision medicine's implementation vehicle, PSB is the optimal forum for the presentation, discussion, and debate of these diverse topics that eventually fuel true individualized health for all.

The investigators and research featured in this session each represent a facet of precision medicine research highlighting needs and gaps that must be addressed to achieve the goals of translational research. Topics covered in this session include the problem of finding sufficient numbers of patients or participants with similar characteristics required to achieve adequate power to identify important biomarkers that distinguish subtypes including genetic, metabolomic or other phenotypic features that have a molecular or mechanistic relationship. Without careful attention, this can lead to health disparities, as less information may be available from vulnerable groups, and thus leading to less effective diagnosis and treatment. Such challenges need to be identified and actively addressed.

This session also addresses development of incorporating social network information to recruit patients in context where they are and better understand the variables that are operative for those individuals in disadvantaged coverage and difficult complex environments, all of which opens up

for new possibilities in research and care.

2. Podium presentations

A group of four manuscripts describe novel and emerging approaches to tackle 'omic data generated by metabolomics and transcriptomics. **A Orlenko** and colleagues[28] present a case study of Tree-based Pipeline Optimization Tool (TPOT)[29], an Automated Machine Learning (AutoML) tool, applied to the clinical metabolic profiling of patients exposed to metformin. In their assessment of TPOT in 546 samples and 42 metabolites, the investigators identified a putatively novel association between increased homocysteine and long-term exposure to metformin. The investigators also developed several considerations for future studies including suggestions for adjustments for confounding features and recommendations for the ideal study design or dataset characteristics that minimize bias and improve AutoML performance.

J Westra and colleagues[30] take a different approach in the analysis of metabolomic data with the observation that these data are better represented by Gaussian mixture distributions rather than the linear models that assume normality commonly used to test for SNP-trait associations. The investigators present an adaptation of Kim et al where tests of association between copy number variants and traits were performed with a likelihood ratio test[31]. The adapted test presented here is a likelihood ratio test that can be constrained based on *a priori* biological data describing the genotype-phenotype relationship, if available. This adapted tested was applied to simulated data to evaluate performance of the test versus liner models as well as to natural data from 20,315 SNPs on chromosome 11 for 5,936 Framingham Heart Study participants with gas-chromatography-measured red blood cell fatty acid levels[32]. A total of 28 SNPs from five different regions of chromosome 11 were associated with the metabolic trait, including 19 SNPs containing the FADS gene complex known to enzymatically desaturate arachidonic acid to dihoo-gamma-linoleneic acid. While further work is needed to extend the model, the present study suggests that powerful analysis of metabolomics data may require different models compared with the traditional linear models so popular in GWAS.

Both J Berghout and S Rachid Zaim offer methods applicable to single-subject transcriptomics. **J Berghout** and colleagues[33] describe a mixture mole for transcript fold-change clustering from isogenically paired samples known as the "N-of-1 pathways MixEnrich."[34] The Gene Ontology Biological Processes (GO-BP)[35, 36] is applied to both reduce dimensionality as well as to identify functional attributes. The method was validated using a microarray dataset of inbred mouse strains exposed to different diets. The MixEnrich results for the paired mouse liver transcriptomes are compared with results from two other methods, Linear Models for Microarray (*limma*)[37] and Gene Set Enrichment Analysis[38], both of which require a minimum of three pairs as opposed to a single sample pair required by MixEnrich. Results suggest that MixEnrich reproduces GO-BP signals in similar priority order compared with the other approaches, thus offering an efficient alternative to cost prohibitive cohort-derived gold standards used for validation.

S Rachid Zaim and colleagues[39] use simulations to address limitations in using transcriptomics as a clinical biomarker. Here, the investigators assume that a transcriptional signal or association can

be detected at the pathway level regardless of patient-level heterogeneity in gene expression. The simulations are performed to assess 54 different scenarios using single-subject and cohort-based techniques describing variability at various levels including pathway gene set size, fraction of expressed gene responsive within the gene set, fraction of up and down-regulated gene expression response, and sample size. Results of these simulations suggest that single-subject pathway detection methods that include patient-level variability can detect transcriptional dysregulation at the pathway level, scenarios likely relevant to heterogeneous clinical populations expected in precision medicine.

Of all the 'omics, genomics, represented by genotyping and sequencing, is the most mature and readily available in the clinic. Whole genome sequencing has been particularly groundbreaking in diagnosing previously undiagnosed rare diseases[40-42], but these data remain an analysis challenge for more common, complex diseases. **A Gupta** and colleagues[43] note, as have others[44], that the genetic architecture of some complex diseases such as autism spectrum disorder (ASD) remains unknown despite evidence of a strong genetic component. While it is now recognized that complex diseases are often a result of many variants across multiple genes with independent and interacting effects, and sequencing methods are available to capture genome-wide variability, few statistical methods have emerged to detect these complex genetic architectures. The investigators posit that Coalition Game Theory (CGT) can be applied to genomic data to identify individual genes ("players") who improve the performance of the coalition or, in this case, the relationship to ASD. The investigators apply CGT to 2,710 whole-genome sequences from ASD multiplex families and identify eight genes with significantly elevated "player scores." All eight genes are in biological pathways known to be affected by ASD and directly interact with genes previously associated with ASD. Although further follow-up is needed, these results suggest CGT is a promising method for large-scale genome data generated for complex diseases.

Two manuscripts centered on the use of automated clinical systems in the delivery or practice of precision medicine. In the first, **C-L Chi** and colleagues[45] apply a treatment simulation and optimization approach to develop decision support for warfarin dosing. Warfarin is a commonly prescribed anti-coagulant, and variability in initial dosing has a strong and known genetic component[46, 47]. Genetically-guiding warfarin dosing was an early poster-child for pharmacogenomics in precision medicine, but lackluster clinical trials[48-50] save for one[51] among other issues have dampened enthusiasm. Logistically, the development of algorithm-based dosing delivered via EHR-automated decision support has been a challenge for this and other drugs impacted by genetics, particularly for diverse populations[52]. The investigators present results of simulations that employed the property of minimal entropy to minimize overall risks for the largest patient groups. The investigators further discuss these results through the lens of ease of implementation, a factor highly relevant for this[53] and other potential precision medicine clinical applications.

S Poole and colleagues[54] address alarm fatigue, an unfortunate consequence of easy and constant automated vital sign monitoring. The investigators aim to improve default vital sign alarm thresholds to decrease the number of unnecessary alarms. The investigators develop personalized vital sign thresholds based on a large heart rate database used to identify the 1st and 99th

percentiles of a patient's heart rate on the patient's first day of monitoring. Results suggest that these new thresholds would decrease low and heart rate alarms while preserving sensitivity and boosting specificity. Overall, the investigators suggest these thresholds will reduce alarm fatigue thereby improving both patient care and hospital costs, major goals underlying the original HITECH Act[55].

3. Posters with published papers

This year's poster session with papers published in the proceedings will feature two research groups. In the first, **A Fish** and colleagues[56] present evidence that genetic associations are modified by local ancestry transitions inferred from genome-wide association study (GWAS) fine-mapping Metabochip data available on ~10,000 African Americans with de-identified EHRs[57]. African Americans are considered admixed with an average 75-93% their genomes originating from West African ancestral populations and the remaining from European ancestral populations[58-61]. Local ancestry, as opposed to inferred genome-wide global ancestry, is the inference of ancestral state at the locus-level. Local ancestry estimates have been instrumental in admixture mapping efforts[62] as well as the estimation of recombination rates and the identification of recombination hotspots[63]. The investigators leveraged this admixed population with clinical data to identify SNP x transition interactions where the transition represent a switch in ancestral backgrounds between nearby variants. Five Bonferroni-corrected significant interactions were identified, and subsequent statistical follow-up suggested that a European to African transition modifies the association rs16890640 between mean corpuscular hemoglobin and mean corpuscular volume. Bioinformatic data coupled with model organism data suggest that alterations in the region chromatin conformation are the biological basis for the modifying effect of the ancestral transition. More broadly, this study offers an example of epistasis where the interaction, and the not variant itself, is associated with the phenotype[64]. Furthermore, this study highlights yet another example of the importance of diverse populations in the search for all genetic variants and their modifiers important in human health and health disparities.

In the second, **B Li** and colleagues[65] test two functions of PrediXcan: 1) its ability to predict gene expression and 2) its ability to prioritize GWAS results. PrediXcan is a gene-based association method rooted in the observation that phenotypic variability can be explained by regulatory variants that modulate the expression levels of genes[66]. PrediXcan uses reference transcriptome data to infer gene expression in GWAS data by estimating the genetically determined component of gene expression. This gene-based approach reduces multiple testing and proffers biological insights or mechanisms compared with standard single SNP tests of association common in GWAS. This study evaluates PrediXcan using genotypic and transcriptomic datasets available from the 1000 Genomes Project (Yoruba; YRI) and GWAS data from the AIDS Clinical Trials Group (ACTG) protocol A5202[67, 68]. To characterize the accuracy in predicting gene expression levels, the investigators compared the PrediXcan-inferred YRI data based on whole blood models and transcriptomic data from the Genotype-Tissue Expression (GTEx) Project[69] and Depression, Genes and Networks (DGN)[70] to the actual YRI expression data and found that the slopes of correlation between predicted and actual were negative for almost one-half of the genes tested. Despite these differences, PrediXcan identified 19 genes in the A5202 cohort dataset associated

with triglyceride change from baseline to 24 or 48 weeks that were not previously identified using phenome-wide association testing, attesting to PrediXcan's potential ability to prioritize GWAS findings. Of note is the poor transcriptome prediction in YRI despite the fact that the GTEx cohort includes African Americans. These data suggest that testing the limits of PrediXcan in gene-trait associations will require more, larger, and diverse populations with both GWAS and transcriptome-level data.

4. Acknowledgements

We would like to thank the members of the program committee for reviewing all submissions and providing expert critiques used in evaluating manuscripts for inclusion into the session and the PSB proceedings. We would also like to thank the PSB 2018 chairs and Tiffany Murray of Stanford University for their efforts in organizing the meeting. This work is supported in part by NIH grants including U41 HG007346 to S.E.B and by a collaborative research agreement with Tata Consultancy Services.

5. References

1. Murray JF. Personalized Medicine: Been There, Done That, Always Needs Work! *American Journal of Respiratory and Critical Care Medicine*. 2012;185:1251-1252.
2. Gurwitz D and Motulsky AG. Drug reactions, enzymes, and biochemical genetics: 50 years later. *Pharmacogenomics*. 2007;8:1479.
3. Motulsky AG. Drug reactions, enzymes, and biochemical reactions. *JAMA*. 1957;165:835-837.
4. Motulsky AG and King MC. The great adventure of an American human geneticist. *Annu Rev Genomics Hum Genet*. 2016;17:1-15.
5. Hall MA, Moore JH and Ritchie MD. Embracing Complex Associations in Common Traits: Critical Considerations for Precision Medicine. *Trends in Genetics*. 2016;32:470-484.
6. Carlsten C, Brauer M, Brinkman F, Brook J, Daley D, McNagny K, Pui M, Royce D, Takaro T and Denburg J. Genes, the environment and personalized medicine. *We need to harness both environmental and genetic data to maximize personal and population health*. 2014;15:736-739.
7. van Nimwegen KJM, van Soest RA, Veltman JA, Nelen MR, van der Wilt GJ, Vissers LELM and Grutters JPC. Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. *Clinical Chemistry*. 2016;62:1458-1464.
8. Feero W, Guttmacher AE and Collins FS. The genome gets personal--almost. *JAMA*. 2008;299:1351-1352.
9. Blumenthal D. Launching HITECH. *New England Journal of Medicine*. 2010;362:382-385.
10. Duffy DJ. Problems, challenges and promises: perspectives on precision medicine. *Briefings in Bioinformatics*. 2016;17:494-504.
11. Lynch JA, Berse B, Dotson WD, Khoury MJ, Coomer N and Kautter J. Utilization of genetic tests: analysis of gene-specific billing in Medicare claims data. *Genet Med*. 2017.
12. Green ED and Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature*. 2011;470:204-213.

13. Iyengar R, Altman RB, Troyanskya O and FitzGerald GA. Personalization in practice. *Science*. 2015;350:282-283.
14. Ritchie MD, Holzinger ER, Li R, Pendergrass SA and Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16:85-97.
15. Schully SD and Khoury MJ. What is translational genomics? An expanded research agenda for improving individual and population health. *Applied & Translational Genomics*. 2014;3:82-83.
16. Herr T, Bielinski S, Bottinger E, Brautbar A, Brilliant M, Chute C, Cobb B, Denny J, Hakonarson H, Hartzler A, Hripcsak G, Kannry J, Kohane I, Kullo I, Lin S, Manzi S, Marsolo K, Overby C, Pathak J, Peissig P, Pulley J, Ralston J, Rasmussen L, Roden D, Tromp G, Uphoff T, Weng C, Wolf W, Williams M and Starren J. Practical considerations in genomic decision support: The eMERGE experience. *Journal of Pathology Informatics*. 2015;6:50-50.
17. Manolio TA. Implementing genomics and pharmacogenomics in the clinic: The National Human Genome Research Institute's genomic medicine portfolio. *Atherosclerosis*. 2016;253:225-236.
18. Collins FS and Varmus H. A New Initiative on Precision Medicine. *New England Journal of Medicine*. 2015;372:793-795.
19. Bustamante CD, De La Vega FM and Burchard EG. Genomics for the world. *Nature*. 2011;475:163-165.
20. Popejoy AB and Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538:161-164.
21. Adler NE and Stead WW. Patients in Context — EHR Capture of Social and Behavioral Determinants of Health. *New England Journal of Medicine*. 2015;372:698-701.
22. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D and Risch N. The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice. *New England Journal of Medicine*. 2003;348:1170-1175.
23. Braveman P, Egerter S and Williams DR. The social determinants of health: coming of age. *Annu Rev Public Health*. 2011;32:381-398.
24. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J and Kohane IS. Genetic Misdiagnoses and the Potential for Health Disparities. *New England Journal of Medicine*. 2016;375:655-665.
25. Kohane IS. Ten things we have to do to achieve precision medicine. *Science*. 2015;349:37-38.
26. Sankar PL and Parker LS. The Precision Medicine Initiative/'s All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet Med*. 2016.
27. Thompson B and Boiani J. The Legal Environment for Precision Medicine. *Clinical Pharmacology & Therapeutics*. 2016;99:167-169.
28. Orlenko A, Moore JH, Orzechowski P, Olson RS, Cairns J, Caraballo PJ, Weinshilboum RM, Wang L and Breitenstein MK. Considerations for automated machine learning in clinical metabolic profiling: altered homocysteine plasma concentration associated wtih metformin exposure. *Pac Symp Biocomput*. 2017;23.
29. Olson RS, Bartley N, Urbanowicz RJ and Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. *Proc Genet Evol Comput Conf*. 2016;GECCO '16:485-492.

30. Westra J, Hartman N, Lake B and Tintle N. Analyzing metabolomics data for association with genotypes using two-component Gaussian mixture distributions. *Pac Symp Biocomput*. 2017;23.
31. Kim W, Gordon D, Sebat J, Ye KQ and Finch SJ. Computing Power and Sample Size for Case-Control Association Studies with Copy Number Polymorphism: Application of Mixture-Based Likelihood Ratio Test. *PLOS ONE*. 2008;3:e3475.
32. Harris WS, Pottala JV, Lacey SM, Vasan RS, Larson MG and Robins SJ. Clinical correlates and heritability of erythrocyte eicosapentaenoic and docosahexaenoic acid content in the Framingham Heart Study. *Atherosclerosis*. 225:425-431.
33. Berghout J, Li Q, Pouladi N, Li J and Lussier YA. Single subject transcriptome analysis reproduces signed gene set functional activation singals from cohort analysis of murine response to high fat diet. *Pac Symp Biocomput*. 2017;23.
34. Li Q, Schissler AG, Gardeux V, Achour I, Kenost C, Berghout J, Li H, Zhang HH and Lussier YA. N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes. *BMC Medical Genomics*. 2017;10:27.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H and Cherry JM. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25.
36. Consortium TGO. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res*. 2017;45:D331-D338.
37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. limma powers differential expression analyses for RNA-sequencing and mciroarray studies. *Nucleic Acids Res*. 2015;43:e47.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005;102:15545-15550.
39. Rachid Zaim S, Li Q, Schissler AG and Lussier YA. Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous Genome-by-Environment transcriptomic responses. *Pac Symp Biocomput*. 2017;23.
40. Chong Jessica X, Buckingham Kati J, Jhangiani Shalini N, Boehm C, Sobreira N, Smith Joshua D, Harrell Tanya M, McMillin Margaret J, Wiszniewski W, Gambin T, Coban Akdemir Zeynep H, Doheny K, Scott Alan F, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, Watkins L, Patterson Karynne E, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor Holly K, Leal Suzanne M, Gunel M, Mane S, Gibbs Richard A, Boerwinkle E, Hamosh A, Shendure J, Lupski James R, Lifton Richard P, Valle D, Nickerson Deborah A and Bamshad Michael J. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*. 2015;97:199-215.
41. Biesecker LG and Green RC. Diagnostic Clinical Genome and Exome Sequencing. *New England Journal of Medicine*. 2014;370:2418-2425.
42. Willig LK, Petrikin JE, Smith LD, Saunders CJ, Thiffault I, Miller NA, Soden SE, Cakici JA, Herd SM, Twist G, Noll A, Creed M, Alba PM, Carpenter SL, Clements MA, Fischer RT, Hays JA, Kilbride H, McDonough RJ, Rosterman JL, Tsai SL, Zellmer L, Farrow EG and Kingsmore SF. Whole-genome sequencing for identification of Mendelian disorders in critically

ill infants: a retrospective analysis of diagnostic and clinical findings. *The Lancet Respiratory Medicine*. 2015;3:377-387.
43. Gupta A, Sun MW, Paskov KM, Stockham NT and Wall DP. Coalitional game theory as a promising approach to identify candidate autism genes. *Pac Symp Biocomput*. 2017;23.
44. Merikangas KR and Risch N. Genomic Priorities and Public Health. *Science*. 2003;302:599-601.
45. Chi C-L, He L, Ravvaz K, Weissert J and Tonellato PJ. Optimized decision support rules of precision warfarin treatment. *Pac Symp Biocomput*. 2017;23.
46. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, Blough DK, Thummel KE, Veenstra DL and Rettie AE. Effect of VKORC1 Haplotypes on Transcriptional Regulation and Warfarin Dose. *New England Journal of Medicine*. 2005;352:2285-2293.
47. Perera MA, Cavallari LH, Limdi NA, Gamazon ER, Konkashbaev A, Daneshjou R, Pluzhnikov A, Crawford DC, Wang J, Liu N, Tatonetti NJ, Bourgeois S, Takahashi H, Bradford Y, Burkley BM, Desnick RJ, Halperin JL, Khlifa SI, Langaee TY, Lubitz SA, Nutescu EA, Oetjens M, Shahin MH, Shitalben RP, Tector M, Rieder MJ, Scott SA, Wu AHB, Burmester JK, Deloukis P, Wagner MJ, Mushiroda T, Kubo M, Roden DM, Cox NJ, Altman RB, Klein TE, Nakamura Y and Johnson JA. Genetic variants associated wtih warfarin dose in African-American individuals: a genome-wide association study. *Lancet*. 2013;382:790-796.
48. Pirmohamed M, Burnside G, Eriksson N, Jorgensen AL, Toh CH, Nicholson T, Kesteven P, Christersson C, Wahlstr+¦m B, Stafberg C, Zhang JE, Leathart JB, Kohnke H, Maitland-van der Zee A, Williamson PR, Daly AK, Avery P, Kamali F and Wadelius M. A Randomized Trial of Genotype-Guided Dosing of Warfarin. *New England Journal of Medicine*. 2013;369:2294-2303.
49. Kimmel SE, French B, Kasner SE, Johnson JA, Anderson JL, Gage BF, Rosenberg YD, Eby CS, Madigan RA, McBane RB, Abdel-Rahman SZ, Stevens SM, Yale S, Mohler ER, Fang MC, Shah V, Horenstein RB, Limdi NA, Muldowney JAS, Gujral J, Delafontaine P, Desnick RJ, Ortel TL, Billett HH, Pendleton RC, Geller NL, Halperin JL, Goldhaber SZ, Caldwell MD, Califf RM and Ellenberg JH. A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *New England Journal of Medicine*. 2013;369:2283-2293.
50. Verhoef TI, Ragia G, de Boer A, Barallon R, Kolovou G, Kolovou V, Konstantinides S, Le Cessie S, Maltezos E, van der Meer FJM, Redekop WK, Remkes M, Rosendaal FR, van Schie RMF, Tavridou A, Tziakas D, Wadelius M, Manolopoulos VG and Maitland-van der Zee AH. A Randomized Trial of Genotype-Guided Dosing of Acenocoumarol and Phenprocoumon. *New England Journal of Medicine*. 2013;369:2304-2312.
51. Gage BF, Bass AR, Lin H, Woller SC, Stevens SM, Al-Hammadi N, Li J, Rodriguez T, Miller JP, McMillin GA, Pendleton RC, Jaffer AK, King CR, Whipple BD, Porche-Sorbet R, Napoli L, Merritt K, Thompson AM, Hyun G, Anderson JL, Hollomon W, Barrack RL, Nunley RM, Moskowitz G, Davila-Roman V and Eby CS. Effect of genotype-guided warfarin dosing on clinical events and anticoagulation control among patients undergoing hip or knee arthoplasty. The GIFT Randomized Clinical Trial. *JAMA*. 2017;318:1115-1124.
52. Johnson JA, Caudle KE, Gong L, Whirl-Carrillo M, Stein CM, Scott SA, Lee MT, Gage BF, Kimmel SE, Perera MA, Anderson JL, Pirmohamed M, Klein TE, Limdi NA, Cavallari LH and Wadelius M. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. *Clinical Pharmacology & Therapeutics*. 2017;102:397-404.

53. Emery JD. Pharmacogenomic testing and warfarin. What evidence has the GIFT Trial provided? *JAMA*. 2017;318:1110-1112.
54. Poole S and Shah N. Addressing vital sign alarm fatigue using personalized alarm thresholds. *Pac Symp Biocomput*. 2017;23.
55. Adler-Milstein J and Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Affairs*. 2017;36:1416-1422.
56. Fish AE, Crawford DC, Captra JA and Bush WS. Local ancestry transitions modify SNP-trait associations. *Pac Symp Biocomput*. 2017;23.
57. Crawford DC, Goodloe R, Farber-Eger E, Boston J, Pendergrass SA, Haines JL, Ritchie MD and Bush WS. Leveraging epidemiologic and clinical collections for genomic studies of complex traits. *Human Heredity*. 2015;79:137-146.
58. Bryc K, Durand E-á, Macpherson J-á, Reich D and Mountain J-á. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*. 2015;96:37-53.
59. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE and Shriver MD. Estimating African American Admixture Proportions by Use of Population-Specific Alleles. *The American Journal of Human Genetics*. 1998;63:1839-1851.
60. Reiner AP, Ziv E, Lind DL, Nievergelt CM, Schork NJ, Cummings SR, Phong A, Burchard EGí, Harris TB, Psaty BM and Kwok PY. Population Structure, Admixture, and Aging-Related Phenotypes in African American Adults: The Cardiovascular Health Study. *The American Journal of Human Genetics*. 2005;76:463-477.
61. Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, Bustamante CD, Kenny EE, Williams SM, Aldrich MC and Gravel S. The Great Migration and African-American Genomic Diversity. *PLoS Genet*. 2016;12:e1006059.
62. Shriner D. Overview of admixture mapping. *Curr Protoc Hum Genet*. 2017;94:1.23.1-1.23.8.
63. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, Bock CH, Boerwinkle E, Cai Q, Caporaso N, Casey G, Adrienne Cupples L, Deming SL, Ryan Diver W, Divers J, Fornage M, Gillanders EM, Glessner J, Harris CC, Hu JJ, Ingles SA, Isaacs W, John EM, Linda Kao WH, Keating B, Kittles RA, Kolonel LN, Larkin E, Le Marchand L, McNeill LH, Millikan RC, Murphy, Musani S, Neslund-Dudas C, Nyante S, Papanicolaou GJ, Press MF, Psaty BM, Reiner AP, Rich SS, Rodriguez-Gil JL, Rotter JI, Rybicki BA, Schwartz AG, Signorello LB, Spitz M, Strom SS, Thun MJ, Tucker MA, Wang Z, Wiencke JK, Witte JS, Wrensch M, Wu X, Yamamura Y, Zanetti KA, Zheng W, Ziegler RG, Zhu X, Redline S, Hirschhorn JN, Henderson BE, Taylor Jr HA, Price AL, Hakonarson H, Chanock SJ, Haiman CA, Wilson JG, Reich D and Myers SR. The landscape of recombination in African Americans. *Nature*. 2011;476:170-175.
64. Ritchie MD. Finding the Epistasis Needles in the Genome-Wide Haystack. In: J. H. Moore and S. M. Williams, eds. *Epistasis: Methods and Protocols* New York, NY: Springer New York; 2015: 19-33.
65. Li B, Verma SS, Veturi YC, Verma A, Bradford Y, Haas DW and Ritchie MD. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pac Symp Biocomput*. 2017;23.

66. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL, Cox NJ and Im HK. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47:1091-1098.
67. Verma SS, Frase AT, Verma A, Pendergrass SA, Mahony SA, Haas DW and Ritchie MD. Phenome-wide interaction study (PheWIS) in AIDS Clinical Trials Group Data (ACTG). *Pac Symp Biocomput*. 2016:5768-68.
68. Verma A, Bradford Y, Verma SS, Pendergrass SA, Daar ES, Venuto C, Morse GD, Ritchie MD and Haas DW. Multiphenotype associaton study of patients randomized to initiate antiretroviral regimens in AIDS Clinical Trials Group protocl A5202. *Pharmacogenet Genomics*. 2017;27:101-111.
69. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J and Moore HF. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580-585.
70. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF and Koller D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*. 2014;24:14-24.

Single subject transcriptome analysis to identify functionally signed gene set or pathway activity [a]

Joanne Berghout[*†]*
Center for Biomedical Informatics and Biostatistics (CB2) & The Center for Applied Genetics and Genomic Medicine, Department of Medicine, University of Arizona, Tucson, AZ 85721, USA
Email: jberghout@email.arizona.edu

Qike Li[†]
CB2, Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721, USA
Email: qikeli@email.arizona.edu

Nima Pouladi and Jianrong Li
CB2, University of Arizona, Tucson, AZ 85721, USA;
Email: nimapouladi@email.arizona.edu; jianrong@email.arizona.edu

Yves A. Lussier*
CB2, BIO5 Institute, University of Arizona Cancer Center, Department of Medicine, University of Arizona, Tucson, AZ 85721, USA; Email: yves@email.arizona.edu

Analysis of single-subject transcriptome response data is an unmet need of precision medicine, made challenging by the high dimension, dynamic nature and difficulty in extracting meaningful signals from biological or stochastic noise. We have proposed a method for single subject analysis that uses a mixture model for transcript fold-change clustering from isogenically paired samples, followed by integration of these distributions with Gene Ontology Biological Processes (GO-BP) to reduce dimension and identify functional attributes. We then extended these methods to develop functional signing metrics for gene set process regulation by incorporating biological repressor relationships encoded in GO-BP as *negatively_regulates* edges. Results revealed reproducible and biologically meaningful signals from analysis of a single subject's response, opening the door to future transcriptomic studies where subject and resource availability are currently limiting. We used inbred mouse strains fed different diets to provide isogenic biological replicates, permitting rigorous validation of our method. We compared significant genotype-specific GO-BP term results for overlap and rank order across three replicate pairs per genotype, and cross-methods to reference standards (*limma*+FET, SAM+FET, and GSEA). All single-subject analytics findings were robust and highly reproducible (median area under the ROC curve=0.96, n=24 genotypes x 3 replicates), providing confidence and validation of this approach for analyses in single subjects. R code is available online at http://www.lussiergroup.org/publications/PathwayActivity

Keywords: N-of-1, reproducibility, transcriptome, atherosclerosis, high fat, ontology.

[†] Contributed equally

[a] This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)

* Corresponding authors

1. Introduction

While precision medicine moves towards understanding disease in individuals, transcriptome expression analysis remains largely in the realm of cohort-level understanding. In large part, this is due to the high dimension, broad range of expression values, and dynamic nature of the transcriptome. However, these qualities also mean that the transcriptome has major potential to reveal important processes during the dynamic course of a disease including onset, progression, and response to therapy[1,2]. Extracting signal from these data requires combating biological noise as well as platform and analysis pipeline factors, with variability in transcript expression levels influenced by an individual's genome sequence, environment, and stochastic processes[3-5].

To counter these challenges, computationally focused investigations using whole transcriptome data to draw inferences about single subjects have typically required either a large set of reference samples for comparison, or used paired samples drawn from the same subject (e.g. tumor-normal) to control for a large component of individual variation. Then, they have (1) sought to find outlier expression patterns correlated with the phenotype of interest[6-9], or (2) used clustering algorithms to bin individuals into *a priori* interpretable classifiers (ex. disease subtypes)[10,11]. Other more biologically focused approaches have limited the dimensionality of the whole transcriptome by constructing curated gene panels with known functional relevance for targeted analysis[9,12]. Combination approaches have integrated expression pattern discovery with biological knowledgebases (ex. Gene Ontology; **GO**)[13,14] to identify functional gene set level signals that require the coordinated activation of many genes, thus reducing the dimension and impact of false positive gene-level prioritization. However, establishing how well a computational analysis method of transcript or pathway prioritization represents the ground truth of a disease is elusive and reproducibility between experiments or across analysis methods has been an issue that must be solved before transcriptomics can be applied to clinical data for non-research use, whether used for diagnosis, clinical decision making, or used to evaluate a patient's response to therapy.

We have previously described the "N-of-1-*pathways* **MixEnrich**" combination method for identification of physiologically responsive biological processes (gene sets) in a single subject using paired RNA-Seq data[15]. That study established that the MixEnrich method could recapture an average signal when applied to single subjects, though there were pathways identified by MixEnrich that did not appear as significant in the cohort analysis. In the absence of available replicate data or access to biological samples, we could not confidently establish whether these individually identified pathways were, in fact, biologically true for those patients, indicating heterogeneous characteristics relevant for personalized medicine. The goal of this study is to quantify reproducibility of single-subject results in a controlled setting with real biological data, as well to extend our bioinformatics methods development to incorporate functional logic that reveals activation, suppression or other regulatory alterations to a biological process.

Some of the most powerful genetic tools for modeling human biology are inbred laboratory mouse strains that have been maintained via brother-sister mating for over a hundred generations, becoming genetically fixed and identical (isogenic) in the process[16]. As such, within a given strain they are genotype replicates, while across strains they maintain differences and can be considered as modeling distinct individuals, exhibiting a broad range of phenotypic variation[17]. We used this

principle to select a microarray dataset of inbred mice placed on an atherosclerotic high fat diet, originally published by Shockley and colleagues[18]. Biological/genotype replicates in this set allowed us to explore the data as a cohort (n=3/group) using paired *limma*[19]+FET, SAM+FET and GSEA analyses comparing high fat diet cohorts to normal diet cohorts within a genotype, or as three independent isogenic pairs by MixEnrich (3 replicate pairs of high fat:normal/strain).

2. Methods

2.1. *Details of microarray and annotation datasets*

We downloaded 144 raw microarray samples from GEO (GSE10493), where Shockley et al.[18] used transcript profiling to study the effect of an atherosclerotic high fat diet on 12 inbred mouse strains. Microarray labeling errors were corrected as indicated on the Center for Genome Dynamics website (http://cgd.jax.org/datasets/expression/10strain.shtml). Briefly, male and female mice of each strain were fed a high-fat diet (30% Kcal from dairy fat) or a normal (6% fat) chow diet for 4 weeks. Then, livers were dissected and total mRNA expression profiling was done on n=3 mice per group using Affymetrix Mouse 430 v2 arrays.

Probe sets were converted to 18, 017 gene identifiers using *mouse4302mmentrezgprobe* v21, downloaded from Brainarray at the University of Michigan (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/21.0.0/entrezg.asp). GO Biological Process (**GO-BP**) terms with gene annotations were downloaded from Mouse Genome Informatics[20,21] (http://www.informatics.jax.org/downloads/reports/index.html#go; gene_association.mgi.gz) on 8 March 2017, and ontology terms were filtered to those 4,682 GO IDs with annotated gene set size between 15-500 (with subsumption) for increased biological resolution. The resulting GO-BP file was used for all enrichment analyses.

2.2. *Data normalization and pre-processing*

All data files were individually normalized using R/bioconductor package *SCAN.UPC* for Single-Channel Array Normalization (SCAN)[22,23]. Batch effects were removed using *ComBat*[24] implemented via the *SVA* package[25].

2.3. *Creation of diet-responsive "individuals" through pairing of isogenic mice*

Although we were using these data to model individual responses to diet, we could not use paired microarrays drawn from a single biological individual as collection of liver RNA is a terminal procedure, and being assigned to one diet precludes the other.

So, we created pairs of microarrays matched according to strain and sex (hereafter, "genotype"), but comprised of one chow-fed mouse and one high fat-fed mouse to create three replicates for quasi-single subject analysis. For example, if we consider the male C57BL/6J (B6) mice fed regular chow as #1-3 while those on high fat diet are #4-6, we would pair #1&4 (rep. 1), #2&5 (rep. 2), and #3&6 (rep. 3). As mice within a strain are isogenic and the experiment was conducted under rigorously controlled conditions, differences in gene expression for these pairs should represent the differences attributable to variation in the mouse's diet, plus stochastic noise.

Within a strain, 0.8%[26]-11%[27] (FDR=0.1) of transcripts expressed in mouse liver have been estimated to vary between individuals at baseline due to developmental, hormonal, circadian, and other factors. This may introduce some "false positive" results in our single-subject analyses that are more indicative of inter-individual variation than diet effect. Still, overall gene expression is highly consistent between mice of the same strain[18,28] (**Suppl. Fig. 1**) and we consider this an acceptable caveat, as similar levels of variation can be observed in human blood samples collected from healthy individual volunteers over a time series[29,30]. An advantage of our model in contrast to simulations, is that by using biological replicates, we do not require manually set parameters to model noise and can capture a more realistic biological scenario – encompassing a range of expression variation, distributing that variation non-randomly across genes, and including gene-gene interactions and gene-gene expression correlation where these apply in natural systems.

2.4. *Calculation of reference standards from cohorts (***limma+*FET, SAM+FET and GSEA)***

To establish reference results for diet-responsive pathways in each genotype, we used three algorithmically distinct methods. For differentially expressed genes (**DEG**) followed by GO term enrichment (DEG+Enrichment) references, we used both the Linear Models for Microarray (*limma*)[19,31] package for single channel microarray experiments and Significance Analysis of Microarrays (SAM)[32] using *siggenes*[33] to identify DEGs using the pairing matrix described in **2.3** and matching that used for MixEnrich analyses. For *limma*, DEGs were identified for each genotype (n=3 high fat versus n=3 normal diet) at $p_{adj} \leq 0.05$ following Benjamini-Hochberg (B-H) multiple test correction. SAM was conducted as a moderated paired t-test with unequal variance with DEGs determined as significantly responsive according to within-genotype delta values based on 8 permutations and q-values calculated by the software. Deltas ranged from 0.8 (DBA/2J female) to 7.4 (CAST/EiJ female) with a median of 1.95, each selected to balance false positive and true positive rates. DEGs were used as input in a Fisher's Exact Test (**FET**) to identify overrepresented GO-BP terms at $FDR_{B-H} \leq 5\%$ (**Fig 3A**).

We also created a non-parametric reference of diet-responsive GO-BP terms for each genotype using Gene Set Enrichment Analysis (GSEA)[8] software downloaded from the Broad Institute (http://software.broadinstitute.org/cancer/software/gsea/) and implemented in Java. GSEA GO-BP terms were called significant at FDR $p_{adj} \leq 0.20$, following the package developers recommended default parameters[8]. FDR was calculated via permutation after shuffling transcript labels 5000 times without replacement. Note that GSEA does not allow paired sample design.

2.5. *Identification of individually responsive GO-BP terms using MixEnrich*

We used the MixEnrich method published recently in Li *et al* (2017)[15] to identify GO-BP terms for each sample pair, representing diet-responsive pathways in single subjects (72 files: 3 isogenic replicate pairs x 12 strains x 2 sexes). Briefly, MixEnrich uses paired data from the same subject to control for genotype effects, and models the absolute value of the log-transformed fold change ($|\log_2 FC|$) across conditions by a probabilistic Gaussian mixture. MixEnrich assumes that these $\log_2 FCs$ follow two distributions where one corresponds to transcripts whose expression is biologically altered between the two conditions (DEGs), and the other distribution corresponds to

the transcripts whose expression remains unaltered. Transcripts assigned to the "altered" distribution with a posterior probability >0.5 become inputs to an FET for identification of overrepresented GO-BP terms, with a Benjamini-Yekutieli multiple hypothesis testing correction applied (**Fig. 3B**). See full details and equations in Li (2017)[15]. R code for MixEnrich is available online at http://www.lussiergroup.org/publications/PathwayActivity

2.6. *Assigning gene set functional direction in significant GO-BP terms*

We mined the ontology structure of GO to identify parent-child relationships with the edge annotation of *negatively_regulates*, representing child GO-BP terms whose gene product annotations have been curated as functional repressors of the activity described in the parent (**Suppl File 1**)[34]. For these transcripts, an increase in expression indicates functional repression or a decrease of GO-BP activity, while a decrease in expression indicates activation of GO-BP activity, or, more accurately – removal of suppression (**Fig. 3C**). When transcripts with these annotations were identified as DEGs, we reversed the sign of the $\log_2$FC in the parent term to mirror the functional impact of that gene's change in expression on the parent term's function.

To sign a GO-BP term as activated, we identified DEGs with increased expression and no regulatory annotation (or exclusively *regulates* and/or *positively_regulates* edges), plus those genes with decreased expression and *negatively_regulates* edges in a direct child term as described. This 'upregulated and upregulatory' set was used in a new contingency table (**Fig. 3D**), and an FET for functional enrichment of BP activation was conducted. In parallel, a separate FET calculation was done to determine evidence that each GO-BP term was functionally suppressed, based on DEGs with decreased expression across the paired condition plus those with increased expression and a *negative_regulates* edge. These differ from the original FET for GO-BP term over-representation *per se* by using only the subset of signal from the (reciprocally) concordantly responsive transcripts. Final functionally signed pathway outputs were generated using the FET p-value and OR from the complete set of altered transcripts (**Fig 3A**), together with a categorical direction determined from the FET described in this section and **Fig 3C**. Functional direction was assigned as (1) activated: FDR<5% FET in up-regulated/up-regulatory and non-significant FET in down-regulated/down-regulatory, (2) reduced activity: non-significant FET in up-regulated/up-regulatory and FDR<5% in down-regulated/down-regulatory, (3) bi-directionally altered: FDR<5% in both FETs, or (4) ambiguous: non-significant in both FETs. GO-BP edge annotation files and R code are available online at http://www.lussiergroup.org/publications/PathwayActivity

3. Results and Discussion

3.1. *GO-BP pathway discovery by MixEnrich, limma*+FET, SAM+FET, *and GSEA*

Across all of the analysis methods we observed that each genotype responded differently to the high fat diet (**Table 1**), which was as expected based on a large body of related work including original analyses using these data[18,35-37]. GSEA tended to identify fewer pathways than limma+FET or ME, even with a more relaxed FDR threshold at 20%. This could be due to the approach's requirement that the underlying data contributing to the enrichment signal must be

directionally concordant[8]. SAM+FET results were also generally low with variable numbers of significant GO-BP term results relative to the other methods for certain strains (male A/J,, B6, I/Ln, female NZB). This could be the result of the increased solution space due to use of two parameters (delta, FC), and likely represents increased stringency applied by the SAM algorithm at the chosen values. On average, 55% of MixEnrich pathways identified in a given genotype replicate pair was common to all three ME replicate analyses of the same genotype, with an average of 42% overlap between all three ME replicates and *limma*+FET, and 41% overlap with SAM+FET analysis. These values are comparable to the overlap observed between SAM+FET and *limma*+FET (41% of *limma*+FET). Exact GO-BP term overlap across all six analyses was modest, averaging 40% of the smallest input value (range: 9%-75%). Based on its ubiquitous use in transcriptome analysis, high power and robust performance with smaller sample sizes[38], we chose *limma*+FET as our reference standard when conducting more in-depth comparisons.

Table 1. Count of GO-BP terms identified for each genotype by *limma*+FET (n=3 paired subjects/genotype; FDR 5%), SAM+FET (n=3 paired subjects/genotype; FDR 5%), GSEA (n=3/diet/genotype; FDR 20%), or as 3 replicate isogenic single-subjects via MixEnrich (ME; n=1 pair/genotype; FDR 5%).

	Method	129	A/J	BALB	C3H	C57BL	CAST	DBA/2	I/Ln	MRL	NZB	PERA	SM/J
Males	*limma*+FET	1034	1085	609	926	393	507	84	298	253	1317	1275	955
	SAM+FET	827	13	408	295	11	229	38	74	62	397	192	555
	GSEA	1171	1102	251	648	1181	524	19	588	21	1316	714	1203
	ME rep 1	1015	1155	934	872	865	727	269	816	530	1493	885	1182
	ME rep 2	1231	1121	649	752	1351	691	541	343	285	1050	868	817
	ME rep 3	970	1005	774	921	633	728	295	1170	394	1280	962	1045
	all 3 ME	670	752	439	554	470	427	166	276	199	857	580	578
	3 ME+*limma*	582	698	368	524	306	322	65	204	165	823	551	534
	all methods	368	5	67	159	1	127	6	15	14	283	123	277
Females	*limma*+FET	70	1422	156	342	123	1100	48	797	127	1165	515	140
	SAM+FET	40	171	102	93	71	99	91	245	73	119	134	88
	GSEA	150	811	0	62	31	752	12	230	39	1117	25	77
	ME rep 1	400	1521	603	474	440	843	355	652	575	932	719	241
	ME rep 2	465	1256	275	435	252	1235	306	528	306	1096	708	281
	ME rep 3	386	1496	361	324	450	870	273	850	318	1225	470	338
	all 3 ME	217	923	189	217	178	630	158	366	171	737	305	157
	3 ME+*limma*	53	866	118	198	85	571	41	329	104	705	236	110
	all methods	5	127	0	12	3	64	9	59	9	62	6	9

3.2. *MixEnrich identified GO-BP terms are highly convergent with reference standard*

We compared the set of GO-BP terms identified in-common across all three MixEnrich genotype replicates to the set of "orphan" GO-BP terms identified by only one MixEnrich replicate (**Fig. 1A**). Terms identified by all 3 MixEnrich replicates were highly overlapping with terms identified as significant by *limma*+FET at FDR 5%. The set of terms that were identified in common across MixEnrich replicates but not *limma* did still approach significance for the majority of cases, with a

median FDR ~20% suggesting a threshold effect. In contrast, orphan GO-BP terms were largely far from significant by *limma* analysis, achieving a median FDR ~66%.

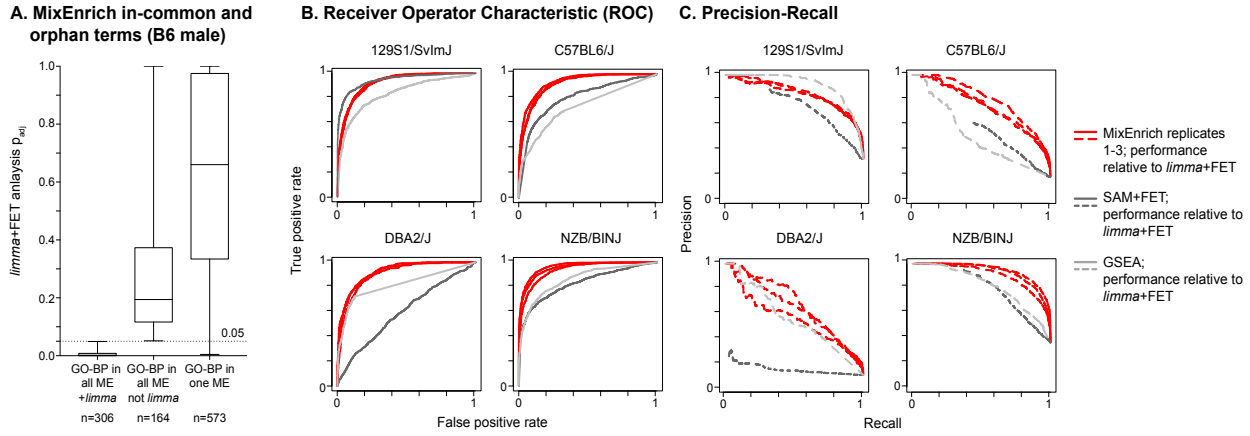


Figure 1. (A) Distribution of *limma*+FET adjusted p-values for GO-BP terms identified in common by all 3 ME replicates in B6 male mice versus orphan GO-BPs identified by a single ME replicate. (B) ROC and (C) precision-recall curves assessing the results of GO-BP enrichment for three replicate pairs analyzed for each genotype by MixEnrich (red, three lines), or cohort-derived reference sets of n=3 high fat diet versus low fat diet mice analyzed by SAM+FET (dark gray, one line) and GSEA (light gray, one line). All curves were compared to a reference standard generated *limma*+FET at FDR 5%.

Next, we compared GO-BP lists using the receiver operator characteristic (ROC). ROC is most commonly used for estimating accuracy of a classifier, based on the balance between identification of true positives relative to false positives. A curve following the diagonal with area under the curve (AUC) equal to 0.5 indicates random performance, while, a curve that hugs the Y-axis up to a square turn at the top of the plot (AUC-ROC = 1) represents perfect accuracy[39]. In this case, a 'true positive' was defined as the MixEnrich, SAM+FET or GSEA analysis identifying a GO-BP term that was also identified by *limma*+FET analysis of that genotype (FDR 5%), and a 'false positive' was defined as the MixEnrich, SAM+FET or GSEA analysis calling a significant GO-BP term that the *limma*+FET analysis did not. This should be considered an approximation as the *limma* method – though widely used and well-validated – is still likely to contain errors in the totality of the result, and selecting FDR 5% as a binary decision threshold for truth is somewhat arbitrary. In addition, for these analyses, we used exact GO-BP term matching only and did not credit similar but non-identical terms appearing on the comparison lists, which is a property we hope to investigate in future analyses. Nonetheless, MixEnrich replicates were highly accurate, capturing much of the signal detected by *limma*+FET (**Table 1**, **Fig. 1B**). Across all genotypes, the median AUC-ROC for single subject MixEnrich replicates was equal to 0.96 (males: median=0.96, 1st quartile= 0.94, 3rd quart=0.98; females: med=0.96, 1st quart=0.94, 3rd quart=0.97). In contrast, SAM+FET achieved a median AUC-ROC of 0.87 (males: med=0.84, 1st quart=0.80, 3rd quart=0.91; females: med=0.89, 1st quart=0.84, 3rd quart=0.93). GSEA scored median 0.81 for males (1st quart=0.67, 3rd quart=0.84) and 0.69 for females (1st quart=0.57, 3rd quart=0.81). **Fig. 1B** shows curves for four representative male genotypes.

We also examined the performance of MixEnrich and GSEA in terms of precision (positive prediction value) and recall (sensitivity) (**Fig. 1C**). Precision and recall are related to the true

positive and false positive measurements in the ROC but provide additional evaluative information in terms of relevance. Precision can be interpreted as the probability that a retrieved positive result by MixEnrich, SAM+FET or GSEA is a true positive, again, based on *limma* analysis as the reference. Recall assesses the ability of those methods to retrieve the complete list of relevant results (those matching *limma* analysis of that genotypes). We observed all three of the MixEnrich replicates outperforming both SAM+FET and GSEA for all genotypes. Relative to the other three strains, precision measurements for the DBA/2J strain were poor, likely reflecting the subtler phenotypic and liver transcriptomic responses in this strain, with accordingly short significant GO-BP pathway list in the *limma*+FET derived reference.

3.3. *Similarity by rank based correlation*

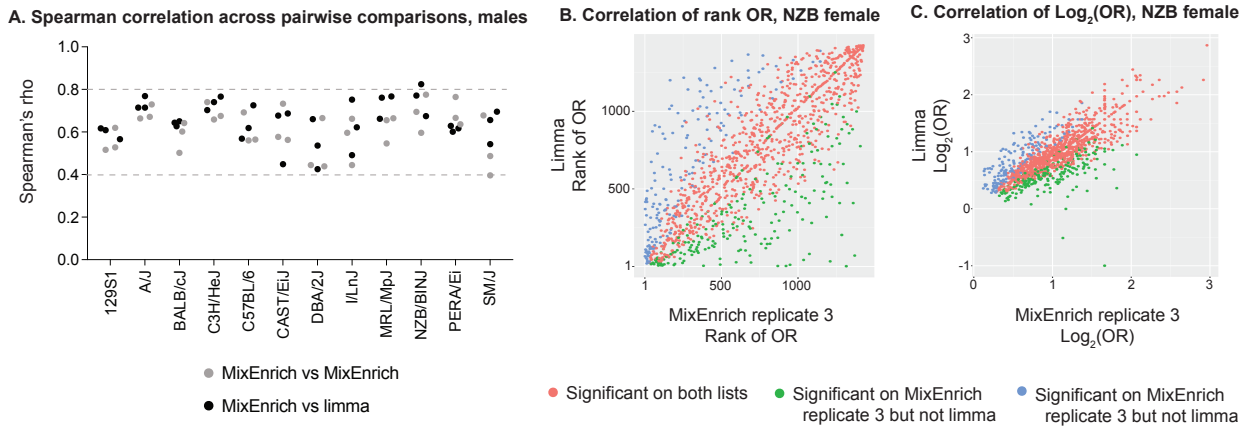


Figure 2. High correlation between rank order of GO-BP terms based on odds ratio. (A) high Spearman's correlation coefficient (rho) across all strains, only males shown. (B) Example plot of rank correlation using NZB female data, comparing MixEnrich replicate 3 to the limma analysis conducted across all NZB females, rho=0.75, $p>10^{-25}$. Data are ranked from smallest to largest, so the largest ORs are ranked at (~1500, ~1500), rather than (1,1). (B) Example plot of $\log_2$(OR) values for the same pairwise comparison.

We next examined whether the GO-BP term lists were similarly ordered between the output of MixEnrich and *limma*. More than simple overlap, it is clearly important in terms of biological or clinical significance that reproducing the strongest and most salient signals should be afforded more weight than reproducing those near the bottom, as these may be of marginal significance despite meeting the set statistical threshold. We used the R package *OrderedList*[40] to compare GO-BP rank order according to their FET-derived GO-BP odds ratios (ORs). Each MixEnrich replicate was compared to the other two and *limma*+FET, resulting in 6 pairwise comparisons per genotype. P-values across all strains were extremely low ($p<10^{-25}$), supporting the assertion that the GO-BP results of MixEnrich were highly consistent across replicates, and their order was highly consistent with the GO-BP pathway order determined by *limma*. Examining a correlation plot of the ranks (**Fig. 2B**) further supports this, as points largely follow the diagonal with the top left corner (highest ranks, sorting from lowest OR to highest OR) highly enriched for commonality. Thus, these analyses were reproducing the same dominant signals.

We also examined each pairwise comparison using Spearman's correlation coefficient (rho; **Fig. 2A**) on the rank of the FET odds ratio (**Fig. 2B**) and $\log_2$OR values themselves (**Fig. 2C**).

Correlation was high between MixEnrich genotype replicates and relative to *limma* (p-value $<10^{-20}$ for all pairwise comparisons), with no significant difference in correlation when cross-replicates were considered (MixEnrich-to-MixEnrich) versus MixEnrich-to-*limma* comparisons ($p>0.05$).

3.4. *Novel signing of GO biological process activity*

Biologically, it is important to know whether the activity in an enriched GO-BP is increased or decreased by the disease state or intervention being studied. GSEA identifies uses concordant direction of gene expression changes, reporting results with a signed enrichment score. However, DEG+enrichment (as in *limma*+FET, SAM+FET, or MixEnrich analyses) only provides an odds ratio and statistical significance rating (p-value or FDR). Altered transcripts adding to the enrichment signal may be increased in expression, decreased, or bi-directionally split. In addition, the biological activity of genes annotated to a term can include negative regulators of that process, whose increased expression logically decreases the parent process when considered as functional biology. (**Fig. 3B, 3C**)[34]. In the mouse genome, 4889 unique gene products have been annotated as having negative regulatory activity in at least one biological process context (annotated to GO:0048519 or child terms), representing 20% of GO-BP annotated genes[21]. However, despite prevalence and utility, this regulatory relationship logic remains underused by the translational bioinformatics and related communities.

We incorporated ontology relationships based on *negatively_regulates* edges into our analysis as described in **Methods section 2.6** and **Fig 3,** then used FET on 'upregulated and upregulatory' genes and 'downregulated and downregulatory' genes separately. Comparing the results from incorporation of the *negatively_regulates* terms to a set of results without this relationship changed the categorical direction of approximately 10-15% of significant GO-BP terms, primarily shifting towards the green/bi-directionally altered category (not shown).

Plotting these significant and functionally signed GO-BP terms in a heat map (**Fig. 3D**) highlights several findings. First, it is again clear that the transcriptional response to high fat diet across mouse genotypes shares some common responses, and some genotype-dependent differences. In the data shown, the most notable difference is between the relatively non-responsive DBA/2J versus the other strains, though genotype-specific blocks of terms can be seen by unsupervised Euclidean clustering. Second, it is apparent that the GO-BP terms identified as significant by all of the methods and across all three biological replicates find a lot in common, as all four columns for a given genotype are visually consistent for much of their length. Third, we can see that the majority of significant DEG+enrichment identified pathways are signed in a common direction across replicates and methods. For GSEA columns (G), we used the sign of the enrichment score which results in either a unidirectional positive upregulation (blue) or negative downregulation (yellow), without the capacity for calling bidirectionality or including repressor function. As a result, the signing/color code of GSEA identified pathways appears as an outlier for a substantial number of pathways. Interestingly, GSEA also failed to identify most of the downregulated metabolic and biosynthesis pathways that comprised about 1/5 of the results in the other methods. These mechanistic responses have been verified and confirmed in other experiments[18,37], so we consider this an omission from GSEA rather than a false positive. Looking

at trends across those pathways in our data, the high fat diet appeared to increase immune and inflammatory activity, increase response to stimulus activity, and decrease metabolic biosynthesis activity, each with some GO-BP specific and genotype-specific variability.

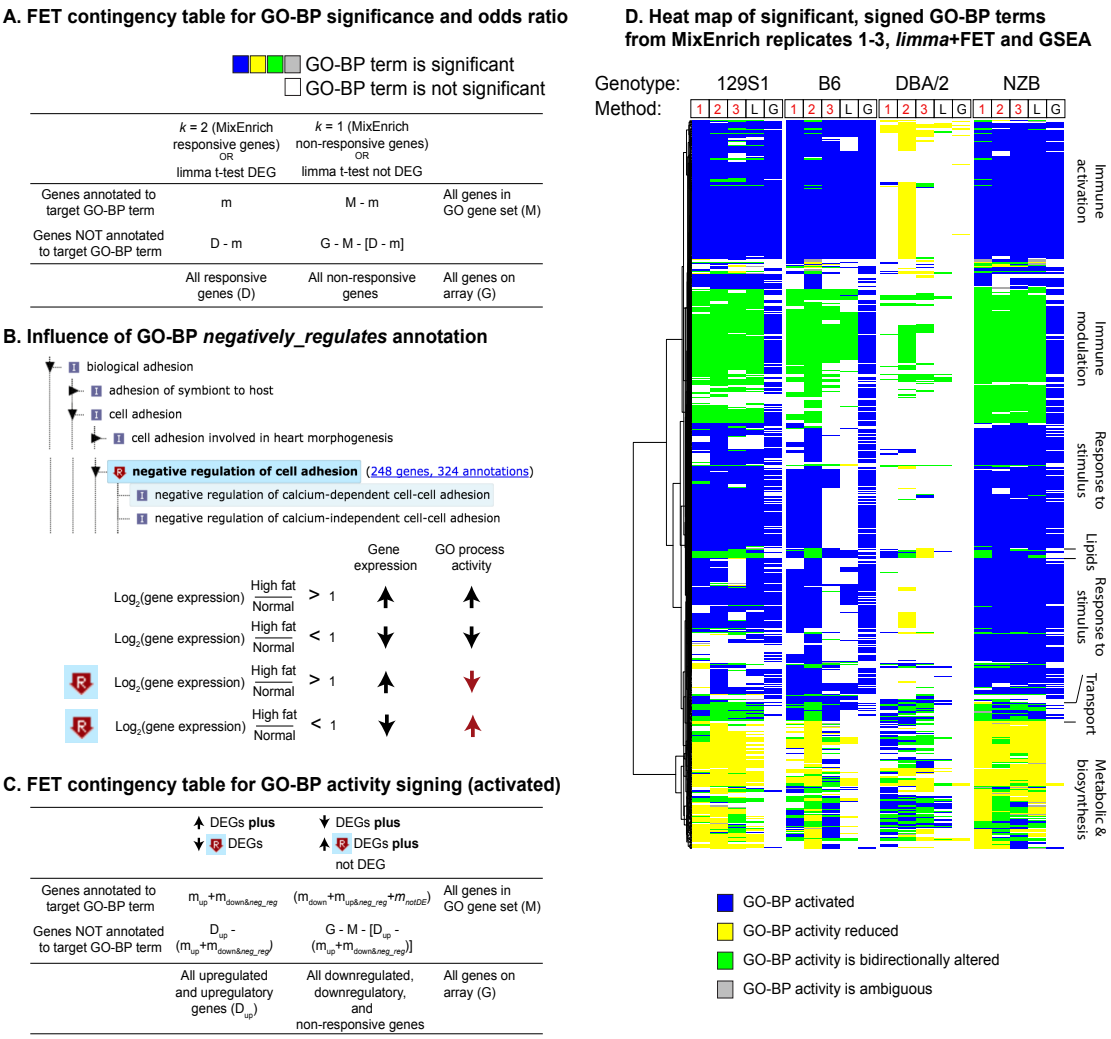


Figure 3. Incorporation of signed functional activation of GO-BP using *negatively_regulates*

3.5. *Biological interpretation of results*

We selected the strain comparison of NZB versus B6 to explore in greater detail. Across MixEnrich replicates and *limma*+FET, we found 268 common enriched GO-BP terms in response to diet with a further 190 GO-BP terms that appeared on 7/8 lists. These included many terms relating to induction of immune related responses and a reduction of cholesterol biosynthesis (**Fig. 3D**) which is consistent with what was reported in the original analyses of these data[18]. Comparing strain-specific responses that may underlie differential phenotypic responses, we found 16 GO-BP terms that appeared as significantly enriched in 4/4 NZB analyses and 0/4 B6 analyses, with 5 GO-BP terms significantly enriched in 4/4 B6 analyses and 0/4 NZB. NZB-specific responses

were suggestive of tissue remodeling in the liver, including angiogenesis related processes, homeostasis, actin and cellular/tissue differentiation. Meanwhile, B6-specific processes were all related to 'entry into host', suggesting immune or cell surface receptor mediated changes.

We next determined if there were any GO-BP terms in common between the two strains where the direction of activity was opposed. Filtering the eligible list to only those GO-BP terms that were significant across all three MixEnrich replicates for both genotypes (n=436 terms), we found only one term that was significant and activated in B6 while significant and repressed in NZB mice. Cofactor catabolic process is involved in metabolism, and the participating transcripts contributing to the enrichment signal include *Acat1*, *Aldh1l1*, *Blvra*, *Blvrb*, *Cbr3*, *Hmox1*, *Hmox2*, *Ncf1*, *Nudt7*, suggesting that the specific processes of heme metabolism, Acetyl-CoA and NADPH oxidation may all be responding differentially across strains. Changes to energetic processes including each of these is consistent with what we would expect from a high fat diet, but without external validation we do not want to draw strong conclusions at transcript resolution.

4. Limitations and future studies

Results described here demonstrate the MixEnrich method is reproducible across isogenic replicates and provides interpretable insights at an accuracy level equivalent to other leading cross-method comparative analyses. However, there are limitations. First, GO-BP is designed to capture the normal behavior of a gene product within its cellular context[14]. If a biological process exists only within a pathogenic state (i.e. cancer), or if a gene product participates in a GO-BP exclusively in the pathogenic state, annotations will not be captured and true biology may be missed. This is a limitation of all gene set methods using GO, but should be acknowledged. In well-understood systems, a custom gene set or alternative ontology may be possible and preferred. Second, regulatory relationships may be context dependent, which we have not accounted for. Bidirectional regulatory capacity is known for certain transcription factors and epigenetic modifying proteins, depending on their interaction partners and which target(s) are being investigated. Our current strategy of treating all DEGs with *negative_regulates* edges as suppressors of their parent process may over-emphasize their down-regulatory capacity. We hope to incorporate a representation of dual annotations in the future. Third, our calculation of gene set overrepresentation via FET, though common, was not thoroughly tested, and alternative statistics (i.e. Mann-Whitney, hypergeometric) may better suit these data[41]. A final constraint on applicability of MixEnrich is that it has exclusively been tested in conditions where a paired sample approach can be used. This applies to many biological questions, but not all, and accurate interpretation of results relies on the 'goodness' of that assay's experimental design.

5. Conclusions

Developing and validating new methods capable of providing insight into the transcriptomes of individuals has the potential to provide important information in aberrant, rare, highly stratified, or clinically relevant patient-level responses. In this study we examined cross-replicate and cross-method reproducibility of GO-BP signal using the paired liver transcriptomes of isogenic mice. Overall, we the MixEnrich method was successfully reproduced the same GO-BP signals as other

methods including *limma*+FET, SAM+FET and GSEA, ranking signals in roughly the same priority order as *limma*+FET. The advantage of MixEnrich is that it only requires a single sample pair and allows individualized conclusions, while the mathematics of *limma*+FET and most other comparably validated methods require at least three pairs to capture an average response.

In addition, to our knowledge, this is the first example of computational method that exploits regulatory edge relationships in signing GO-BP directionality as activated or repressed. These edges have been a component of the ontology structure since 2011[34], but primarily used to reason with logic for knowledge representation rather than incorporated into data analysis pipelines.

6. References

1. Chakravarti A. *Science.* 2011;334(6052):15.
2. Chen R, Snyder M. *Wiley Interdiscip Rev Syst Biol Med.* 2013;5(1):73-82.
3. Love MI, Huber W, Anders S. *Genome Biol.* 2014;15(12):550.
4. Marioni JC, Mason CE, et al. *Genome research.* 2008;18(9):1509-1517.
5. Kim JK, Kolodziejczyk AA, et al. *Nat Commun.* 2015;6:8687.
6. Wang H, Sun Q, Zhao W, et al. *Bioinformatics.* 2015;31(1):62-68.
7. Wang H, Cai H, Ao L, et al. *Brief Bioinform.* 2016;17(1):78-87.
8. Subramanian A, Tamayo P, et al. *Proc Natl Acad Sciences USA.* 2005;102(43):15545-15550.
9. Cummings BB, Marshall JL, Tukiainen T, et al. *Sci Transl Med.* 2017;9(386).
10. Parker JS, Mullins M, Cheang MC, et al. *J Clin Oncol.* 2009;27(8):1160-1167.
11. Cancer Genome Atlas N. *Nature.* 2012;490(7418):61-70.
12. Sparano JA, Paik S. *J Clin Oncol.* 2008;26(5):721-728.
13. Ashburner M, Ball CA, Blake JA, et al. *Nature genetics.* 2000;25(1):25-29.
14. Gene Ontology C. *Nucleic Acids Res.* 2015;43(Database issue):D1049-1056.
15. Li Q, Schissler AG, Gardeux V, et al. *BMC Medical Genomics.* 2017;10(1):27.
16. Silver LM. *Mouse Genetics: Concepts and Applications.* Oxford University Press; 1995.
17. Churchill GA, Airey DC, Allayee H, et al. *Nat Genet.* 2004;36(11):1133-1137.
18. Shockley KR, Witmer D, et al. *Physiol Genomics.* 2009;39(3):172-182.
19. Smyth GK. In: Gentleman RC, Carey VJ, et al. eds. New York: Springer; 2005:397-420.
20. Shaw DR. *Curr Protoc Bioinformatics.* 2016;56:1 7 1-1 7 16.
21. Blake JA, Eppig JT, Kadin JA, et al. *Nucleic Acids Res.* 2017;45(D1):D723-D729.
22. Piccolo SR, Sun Y, et al. *Genomics.* 2012;100(6):337-344.
23. Piccolo SR, Withers MR, et al. *Proc Natl Acad Sci U S A.* 2013;110(44):17778-17783.
24. Johnson WE, Li C, Rabinovic A. *Biostatistics.* 2007;8(1):118-127.
25. Leek JT. *Nucleic Acids Res.* 2014;42(21).
26. Pritchard CC, Hsu L, Delrow J, et al. *Proc Natl Acad Sci USA.* 2001;98(23):13266-13271.
27. Vedell PT, Svenson KL, Churchill GA. *BMC Genomics.* 2011;12:167.
28. Chick JM, Munger SC, Simecek P, et al. *Nature.* 2016;534(7608):500-505.
29. Whitney AR, Diehn M, Popper SJ, et al. *Proc Natl Acad Sci USA.* 2003;100(4):1896-1901.
30. Eady JJ, Wortley GM, Wormstone YM, et al. *Physiol Genomics.* 2005;22(3):402-411.
31. Ritchie ME, Phipson B, Wu D, et al. *Nucleic Acids Res.* 2015;43(7):e47.
32. Tusher VG, Tibshirani R, Chu G. *Proc Natl Acad Sci USA.* 2001;98(9):5116-5121.
33. *siggenes* [computer program]. Version 1.50.0: Bioconductor; 2012.
34. Mungall CJ, Bada M, Berardini TZ, et al. *J Biomed Inform.* 2011;44(1):80-86.
35. Barrington WT, Pomp D., et al Abstract presented at TAGC; July 13-17, 2016, 2016; Orlando, FL.
36. Parks BW, Sallam T, Mehrabian M, et al. *Cell Metab.* 2015;21(2):334-346.
37. Montgomery MK, Hallahan NL, Brown SH, et al. *Diabetologia.* 2013;56(5):1129-1139.
38. Jeanmougin M, de Reynies A, et al. *PloS one.* 2010;5(9):e12336.
39. Fawcett T. *Pattern Recognition Letters.* 2006;27:861-874.
40. *OrderedList: Similarities of Ordered Gene Lists.* [computer program]. Version 1.44.02008.
41. Khatri P, Draghici S. *Bioinformatics.* 2005;21(18):3587-3595.

Using simulation and optimization approach to improve outcome through warfarin precision treatment

Chih-Lin Chi[1], Lu He[2], Kourosh Ravvaz[3], John Weissert[3], Peter J. Tonellato[4,5]

[1]*School of Nursing & Institute for Health Informatics*
University of Minnesota, Minneapolis, MN, USA;

[2]*Computer Science and Engineering*
University of Minnesota, Minneapolis, MN, USA;

[3]*Aurora Health Care, Milwaukee, WI, USA;*

[4]*Department of Biomedical Informatics, Department of Pathology*
H*arvard Medical School, Boston, MA, USA;*

Zilber School of Public Health
[5]*University of Wisconsin-Milwaukee, Milwaukee, WI, USA*

Email: cchi@umn.edu, Peter_Tonellato@hms.harvard.edu

We apply a treatment simulation and optimization approach to develop decision support guidance for warfarin precision treatment plans. Simulation include the use of ~1,500,000 clinical avatars (simulated patients) generated by an integrated data-driven and domain-knowledge based Bayesian Network Modeling approach. Subsequently, we simulate 30-day individual patient response to warfarin treatment of five clinical and genetic treatment plans followed by both individual and sub-population based optimization. Sub-population optimization (compared to individual optimization) provides a cost effective and realistic means of implementation of a precision-driven treatment plan in practical settings. In this project, we use the property of minimal entropy to minimize overall adverse risks for the largest possible patient sub-populations and we temper the results by considering both transparency and ease of implementation. Finally, we discuss the improved outcome of the precision treatment plan based on the sub-population optimized decision support rules.

Keywords: Precision Medicine; Personalized Treatment; Clinical Trial Simulation; Optimization.

1. Introduction

Precision medicine's fundamental charter is to improve treatment outcomes. Generally, outcome is improved by tailoring treatment personalized to an individual's clinical and genetic characteristics. In this manuscript, we combine both simulation and optimization approaches to personalize treatment using decision support rules, which indicate which particular treatment plan maximizes outcome improvement for patients with a particular set of clinical and genetic characteristics

(throughout this manuscript, patients with a particular set of clinical and genetic characteristics are defined as a 'sub-population'). We use precision warfarin treatment plans as an example to show our approach, methods and results.

Warfarin is the most widely used anticlotting agent with a highly significant effect on reducing blood clots and preventing strokes and, thus, can dramatically decrease healthcare costs due to clotting-related events. However, administration of this drug is a challenge because of both sides of risks: stroke, if under-dosed, and bleeding, if over-dosed. Over 50 unique warfarin treatment plans have been developed to minimize the complexity of administration and both sides of risks. Warfarin treatments plans predict the initial warfarin dose and suggest a follow-up dose adjustment based on the patient's lab results, International Normalized Ratio (INR), and the patients individual clinical and/or genetic factors.

The key question to healthcare providers and hospital administrators charged with treating hundreds of thousands (if not millions) of a diverse collection of patients, is which treatment plan is the most "effective". For warfarin, no single treatment plan exists that provides the "optimal" outcome for every patient in a large diverse patient population In short, there is no "one-size fits all" treatment plan that optimally reduces adverse events for all individuals. Our approach is based on a fundamental principle, "What are the optimal sub-populations of patients, with corresponding precision-based treatment plan, that minimizes adverse events while simultaneously reduces complexity of implementation across a large diverse patient population." Our optimization "variable" is the collection of precision treatment plans derived from each individual's clinical and genetic characteristics [1,2]. We then vary the treatment plan, and definition of sub-population to create the optimal outcome. The combination of prediction (of treatment outcome as a function of treatment plan) and optimization of grouped outcome (across sub-populations) allows our approach to capture each individuals' variability; produce precision-driven treatment plans and takes into consideration the complexity of implementing optimized treatment protocols across a diverse patient population [2–4].

Specifically, we selected 5 warfarin treatment plans, each of which consists of 3 different treatment periods and protocols. We demonstrate our approach to find sub-populations and the optimal treatment plan for each sub-population. To avoid confusion, throughout this manuscript, we use the term 'treatment plan' for the 30-day course of treatment and 'protocol' for the specification of components of the 30 day treatment plan. In addition, 'personalized treatment plan' is the treatment plan optimized for an individual whereas 'precision treatment plan' is the treatment plan optimized for an entire sub-population.

2. Methods

In this manuscript, we combined all methods to produce rules to decide precision treatment plan with ~1,500,000 clinical avatars, who are simulated Aurora Health Care patients. In this method section, we first discuss the clinical avatars, clinical trial simulation, and optimization. Figure 1 summarizes 4 key steps to produce data for this work, which consists of two major components, Treatment Simulation[5] and Optimization[2].

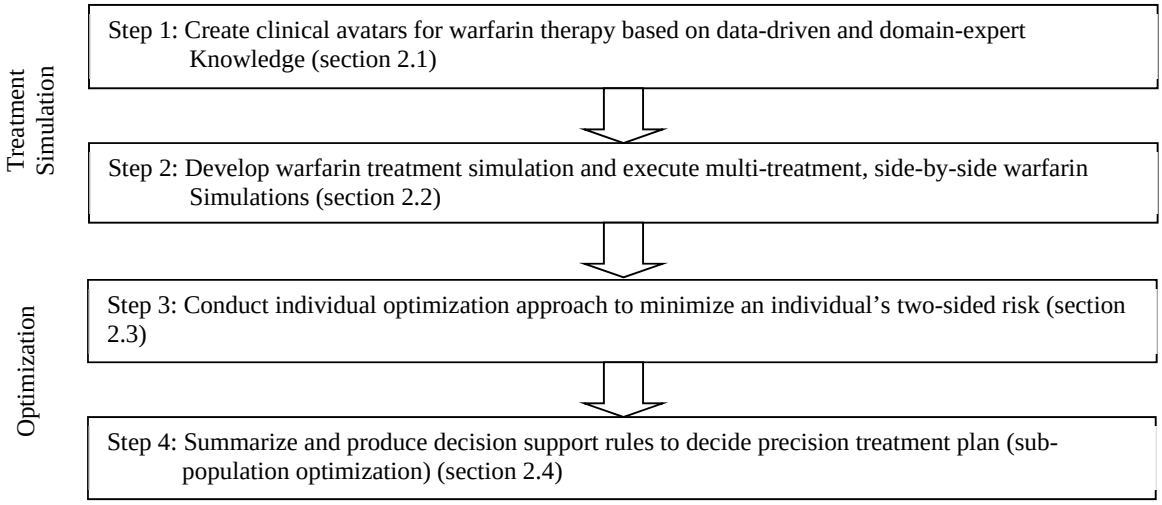


Figure 1. 4-step integrated approach to produce rules to decide precision treatment plan

2.1. *Create clinical avatars (simulated patients) for warfarin therapy based on data-driven and domain-expert knowledge*

The platform used in this study begins with extracting real patient electronic medical records (EMRs) from Aurora Health Care, a major hospital system in the Midwest. We used only de-identified patient data per IRB approval by an AHC honest broker. We extract from a diverse collection of 14,206 patients with Atrial Fibrillation (AF) over 10 years which were then integrated into that previously validated warfarin clinical trial simulation framework. The simulation framework has reproduced warfarin clinical trial results by using a PK/PD model [6] and iterative Bayesian Network Modeling (BNM) implemented in TETRAD [7].

We incorporated both data-driven knowledge and domain-expert knowledge to create clinical avatars. Specifically, (1) Distribution and causal relationships among most variables of patient characteristics were discovered from the EMRs, and (2) distribution of genetic tests, CYP2C9 and VKORC1, and causal relationships between other variables were observed from extensive literature review. We used TETRAD to take both types of input parameters to instantiate a Bayesian Network Model of the EMRs coupled with stochastic models to subsequently create clinical avatars. In order to cover variations, we create 100 times more subjects (clinical avatars) than the original subjects (EMR patients). The clinical avatars were then used as input to the following step, the warfarin Pharmacokinetic and Pharmacodynamic (PK/PD) model, which will be described in the following section in details.

2.2. *Develop warfarin treatment simulation and execute multi-treatment, side-by-side warfarin simulations*

Details of the treatment simulation are described in our preliminary work [8]. In brief, on the treatment simulation platform, a clinical avatar representing a patient, "takes" the protocol-adjusted warfarin

dose each day of a thirty-day simulation. Since International Normalized Ratio (INR) is not tested daily, the avatar's dose may not be changed until the next scheduled INR test.

In this study, we simulate treatment process for created clinical avatars based on 5 warfarin treatment plans, which are summarized in Table 1. They are AAA, CAA, PGAA, PGPGI, and PGPGA. Each treatment plan includes three protocols for three treatment periods. For example, in the PGPGI treatment plan, modified IWPC PG protocol is used for initial period (days 1 to 3) [9], Lenzini PG protocol is used for adjustment period (days 4 and 5), and Intermountain protocol is used for maintenance (days 6 to 30) period. Among all protocols, some protocols (including AHC [10], IWPC Clinical [9], and Intermoutain [11]) use only INR and others (including IWPC PG [9], Modified IWPC PG [12], and Lenzini PG [13]) use genotypic information and/or INR to adjust dosage. We note that 'AHC' is the treatment plan or protocols currently used in the Aurora Health Care and our team coded that clinical protocol and put it in our simulation settings. Other treatment plans and protocols can be found in the associated references.

Table 1. Five distinct treatment plans included in our treatment simulations.

Treatment Periods / Treatment Plans	Initial protocol (days)	Adjustment protocol (days)	Maintenance protocol (days)
AAA (Clinical)	AHC (1-2)	AHC (3-7)	AHC (8-30)
CAA (Clinical)	IWPC Clinical (1-2)	AHC (3-7)	AHC (8-30)
PGAA (Pharmacogenetics)	IWPC PG (1-2)	AHC (3-7)	AHC (8-30)
PGPGI (Pharmacogenetics)	Modified IWPC PG (1-3)	Lenzini PG (4-5)	Intermountain (6-30)
PGPGA (Pharmacogenetics)	Modified IWPC PG (1-3)	Lenzini PG (4-5)	AHC (6-30)

In practice, each simulated patient receives warfarin treatment based on the treatment plans and protocols. Their physiological response to the dosing is measured by International Normalized Ratio (INR), i.e., measurement to understand how thin the blood is. Generally, if the INR is larger or smaller than the therapeutic window (often the range $2.0 < INR < 3.0$), then the protocol calculates the correct change in dose to adjust INR into therapeutic range. In our computational simulation the INRs are predicted using a PK/PD model [6] based on clinical and genotypic characteristics of the clinical avatar.

Similar to actual treatment process, on this platform, the clinical avatar ingests the warfarin dose computed by the treatment plan or protocol, then the PK/PD model predict INR, and, then, the warfarin dose is, again, adjusted based on the INR. Typically, warfarin is taken every day, and INR test is measured based on the schedule documented on the protocol.

In this study, we simulated 30-day warfarin treatment process using five treatment plans for each clinical avatar based on adapted PODSS algorithm [3,4], which was originally developed in machine learning settings. Each treatment plan includes three protocols, each of which corresponds to a treatment period, and a different protocol uses a distinct algorithm to adjust dosage. As a result, we generate several INR values in the 30-day treatment for a protocol. Using each clinical avatars

modeled INR values, we calculated Time in Therapeutic Range (TTR; 0-100%) by way of linear interpolation between protocol based check days [14] with therapeutic range defined as 2-3. The TTR is typically used as surrogate outcome in warfarin studies to represent prevention of the both side of risks, stroke and bleeding. In general, the higher TTR, the lower the two-sided risks, and vice versa. Finally, we have 5 TTRs, each of which is associated with a specific treatment plan. These TTRs are personalized prediction results and the values are very different for various patients based on their clinical and genetic characteristics. These predicted TTR outcome can be further used for optimization to identify personalized treatment in the following steps.

2.3. *Conduct individual optimization approach to minimize an individual's two-sided risk*

In the simulation, when a treatment plan has a larger predicted TTR, the treatment plan has a higher probability and chance to actually minimize two-side of risks. Hence, the purpose is to obtain a treatment plan to maximize predicted TTR (or minimize risks). More details were discussed in our preliminary work [1]. Implement the individual optimization approach is straightforward. We first prioritize TTRs (each corresponding to a protocol) using the matrix of data developed from the above warfarin treatment simulations for each patient and then identify the protocol with the largest TTR. The treatment plan or protocol with the greatest TTR indicates the optimal reduction in the two-sided risks for a clinical avatar. As a result, based on each clinical avatars' clinical and genetic characteristics, in this process, we identify a label of treatment plan that minimizes the two-sided risks for each clinical avatar.

2.4. *Summarize and produce decision support rules to decide precision treatment plan (sub-population optimization)*

The difference between this and the last step is optimization target: an individual or a sub-population.

The advantage of this step is 'practicability', which means we do not need sophisticated program and clinical decision support system (CDSS) to take multiple variables of a patient to find the 'personalized treatment plan'. Instead, one can use simple criteria and easily follow the decision support rules created in this step to find the 'precision treatment plan'. A rule, for example, indicates, PGAA treatment plan is predicted to minimize two-sided risks for the patient sub-population, age<64.95 yrs and VKORC = A/A or G/A.

We used an entropy function implanted in a decision tree algorithm to create the rule set. The reason and details has been discussed in our previous method paper [2]. In brief, we aim to maximize overall outcomes by simultaneously optimize two targets when producing decision support rules: (1) identify a treatment plan to minimizes the two-sided risks for a sub-population and (2) identify the sub-population that maximal proportion of patients can benefit from the identified treatment plan. Due to our specifically designed data structure and entropy/impurity property, we simultaneously optimize these two targets to create decision support rules.

Table 2. Independent variables used to create the decision tree

Variable Name	**Variable Type**	**Variable Range**
Weight	Continuous	71-490
Height	Continuous	46.1-85
Age	Continuous	18-102
Race	Discrete(Nominal)	Asian/Black or African American/White
AMI	Binary	0/1
BMI	Continuous	8.7-91.3
CYP2C9	Discrete(Ordinal)	*1/*1 (value=5), *1/*2 (value=4), *1/*3 (value=3),*2/*2 (value=2),*2/*3 (value=1)
Fluvastatin	Binary	0/1
VKORC1G	Discrete(Ordinal)	A/A (value=1), G/A (value=2), G/G (value=3)
Gender	Binary	0/1
Smoker	Binary	0/1

Table 2 summarizes independent variables used for the multi-class decision tree algorithm implemented in Matlab [15,16]. All independent variables are numeric. The original variable type of CYP2C9 and VKORC1 are categorical. We transform these categorical variables into ordinal variables based on the dosage amount [17] for appropriate therapy. In the Step 3, we have created a label of optimal personalized treatment plan for a clinical avatar. Dependent variable of the decision tree is that label. The dataset consists ~1,500,000 clinical avatars, each of which includes a set of patient characteristics and the optimal treatment plan label. Now the computational problem becomes using a decision tree algorithm to identify patient sub-population with the maximal proportion of individuals that benefit from one or fewer optimal treatment plan label(s).

3. Results

3.1. *Observation of the data*

Table 3 summarizes comparison of variable distribution between EMR and clinical avatars. There is no statistically significant difference ($P < 0.05$) between the two populations for both continuous and discrete variables.

Table 3. Characteristics of the Original Aurora AF Warfarin and Aurora Clinical Avatar Populations.

Characteristic		Aurora AF Warfarin Population (mean±SD)	Aurora Clinical Avatar Population
Age	year	67.3±14.43	67.2±14.47
Weight	lb	199.24±54.71	199.24±54.6
Height	in	66.78±4.31	66.53±4.32
Gender, %	Female Male	53.14 46.86	53.10 46.90
Race, %	White African-American Asian Am. Indian/Alaskan Pacific Islander	95.18 4.25 0.39 0.18 0.0001	95.22 4.19 0.40 0.18 0.0001
Tobacco, %	No Yes	90.33 9.66	90.67 9.33
Amiodarone, %	No Yes	88.45 11.54	88.49 11.51
Fluvastatin, %	No Yes	99.97 0.03	99.98 0.02
CYP2C9, %	*1/*1 *1/*2 *1/*3 *2/*2 *2/*3 *3/*3	65.77[a] 14.6[a] 9.11[a] 6.41[a] 1.93[a] 0[a]	67.39 14.86 9.25 6.51 1.97 0
VKORC1, %	G/G G/A A/A	38.54[a] 44.02[a] 17.33[a]	38.37 44.18 17.45

[a]Aurora patient population's genotypic characteristics were derived from published genotype distributions.

Figure 2 summarizes distribution of personalized treatment plan labels (derived from the Step 3) observed in all clinical avatars. Each clinical avatar has a personalized treatment plan label optimized from one of the five treatment plans. As described in the method section, the optimal label is decided based on the highest TTR across the 5 available treatment-plan options. That optimal-treatment-plan label is a class label used in the subsequent decision tree algorithm. PGAA treatment plan (37.8%) is the majority followed by CAA (23.4%) and AAA (22.7%) treatment plans.

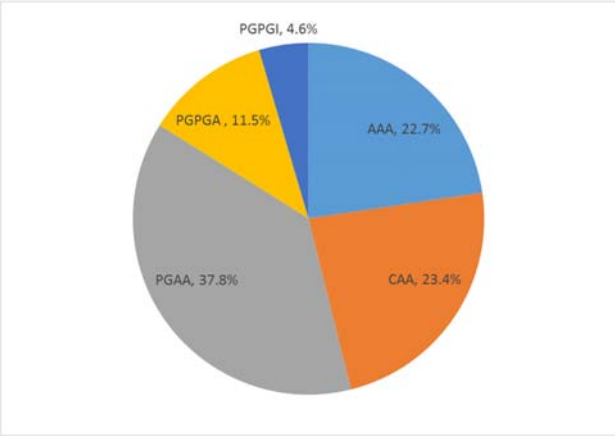


Figure 2. Proportion of personalized protocol derived from individual optimization.

3.2. *Derive personalized treatment rules from the data*

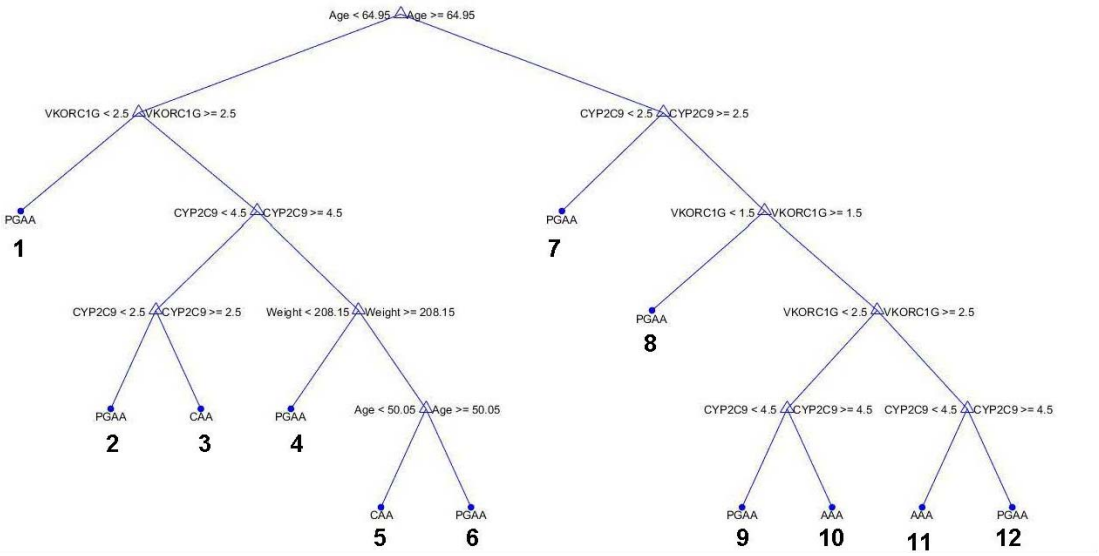


Figure 3. Personalized Treatment Rules extracted from ~1,500,000 clinical avatars by using a multi-class decision tree algorithm. The number of clinical avatars in leaf nodes 1 to 12 ranged from 19,518 to 382,823.

Figure 3 shows decision support rules derived from all those clinical avatars using multi-class decision tree algorithm. When the tree is very large, we have many complicated and a large number of rules, which results in a challenge for practical use. On the other hand, when the tree is very small, the effect of personalized treatment will diminish (extreme example is no tree at all and there is no benefit due to personalized treatment). Therefore, we empirically test and decide 12 decision support rules to maximize visualization and ease of use while maintaining performance of grouping by pruning the tree [18].

In this decision tree, we have 12 leaf nodes, each of which has a number corresponding to the nodes in the following figures. In each leaf node, we also include the label of optimal sub-population treatment plan, which is a predicted class of that node in the decision tree algorithm.

We note that the 'actual' prediction has been done in simulation and individual optimization (Steps 2 and 3). Here, the decision tree algorithm is used to summarize and produce decision support rules using prediction data created in the Steps 2 and 3. Similar to Figure 2, the optimal sub-population treatment plan for most nodes is PGAA followed by CAA and AAA.

3.3. *Analysis of treatment plans in each leaf node*

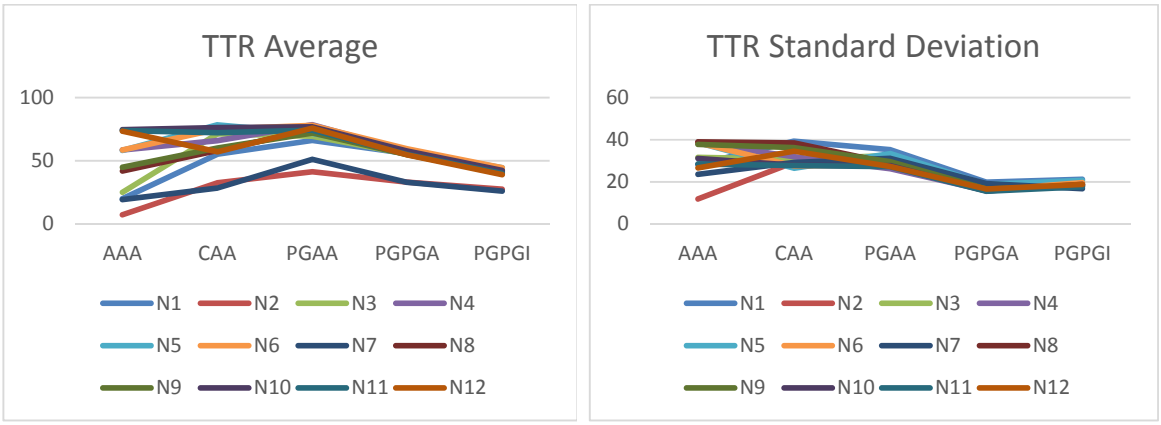


Figure 4. TTR average and standard deviation of the five protocols in different sub-populations (nodes 1 to 12). In general genetic-based treatment plans have lower variations (such as PGAA, PGPGA, and PGPGI).

Before comparing personalized with one-fit-all treatment plans, in this section, we observe TTR average and standard deviation of each one-fit-all treatment plan for the 12 leaf nodes (Figure 4). Specifically, we first follow decision rules in the Figure 3 to identify the 12 patient sub-populations. In each sub-population, we take the average and standard deviation of the 5 treatment plans from each clinical avatar belong to that sub-population. As a result, we have 5 TTR averages and standard deviations observed from the 12 patient sub-populations.

In the Figure 4, we can observe that variations for AAA and CAA are very high. Although the average TTRs for some sub-populations are high, the average TTRs for others are low. Therefore, the overall TTR is not high and, therefore, the chance to become optimal sub-population protocol is in the middle. On the other hand, although the variations for PGPGA and PGPGI treatment plans are consistently low, the TTR averages for most nodes are also consistently low. Thus, the overall TTRs are low across all patient sub-populations. Finally, PGAA treatment plan has both advantages: higher TTR average and lower variations across all patient sub-populations. Therefore, the chance to become sub-population optimal protocol is the highest. Therefore, we can expect PGAA should be the dominated sub-population protocol, followed by AAA and CAA. This expectation corresponds to the rule to decide personalized treatment plan in Figure 3, in which PGAA is the majority followed by AAA and CAA treatment plans.

3.4. *Compare personalized with one-fit-all treatment plans*

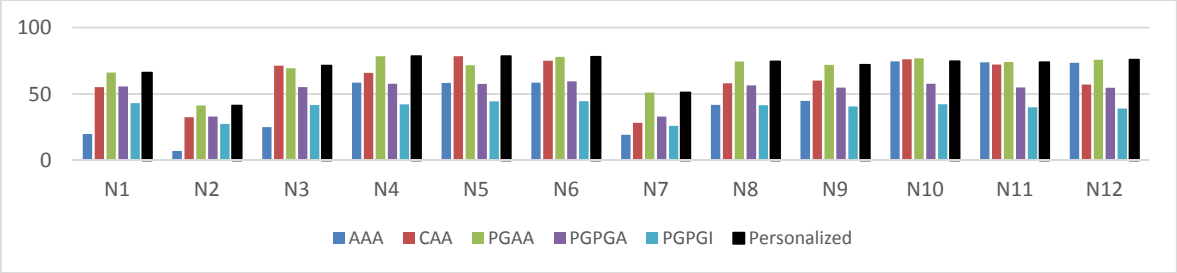


Figure 5. Comparison of TTR average between optimal sub-population and one-fit-all treatment plans.

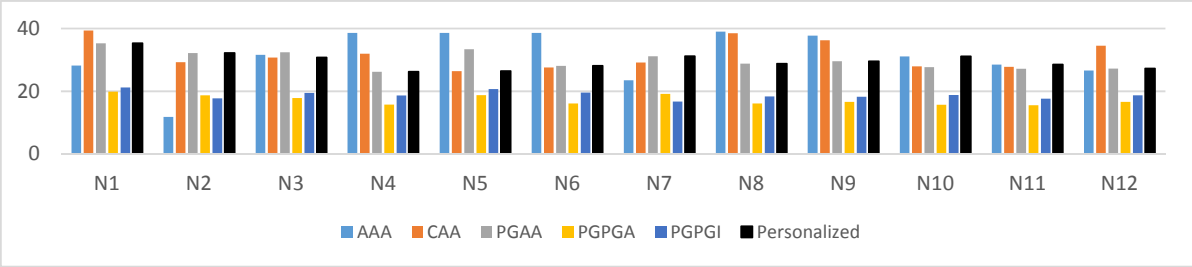


Figure 6. Comparison of TTR standard deviations between optimal sub-population and one-fit-all treatment plans.

Finally, we compare TTR average and standard deviations between the optimal sub-population and one-fit-all treatment plans in Figures 4 and 5, respectively. There are 12 sub-populations, each of which correspond a leaf node (N1~N12). The optimal sub-population treatment plan is the treatment plan (black bar) recommended in the Figure 3 for a specific type of patients. On the other hand, the one-fit-all treatment plans are the treatment plans (color bars) uniformly used on every clinical avatar belong to that sub-population. We note that only computational settings, like simulation, allow a clinical avatar receive warfarin treatment based on multiple treatment plans. From Figure 5, we can observe that TTR of the optimal sub-population treatment plan is almost the highest across all patient sub-populations. On the other hand, variations are, in general, around the middle (Figure 6). In sum, compared to one-fit-all treatment protocol, personalized treatment plan improve TTR outcome, ranged from 15% to 31% (average and median across those 12 leaf nodes are both 24%). Such outcome improvement is the result of optimizing and reusing existing treatment plan using precision-medicine approach.

4. Conclusion

Combing clinical avatars simulation, with optimization can yield clinical decision support rules with respect to dosing strategy for patients with AF. Clinical avatars are developed through BNM model based on the distribution of patient characteristics in Aurora Healthcare EMRs and genotypic characteristics from published literature (Step 1 in the Figure 1). We then simulated individual dose response and outcomes for 30-day warfarin therapy. All ~1.5-million clinical avatars underwent on five simulated treatment regimes (Step 2). Simulated outcomes were then used as input to apply individual optimization to create the label of a treatment plan that maximize TTR outcome (Step 3). Finally, a decision tree algorithm was used to identify decision support rules that minimizes overall

adverse risks (or maximizes TTR) for the largest possible patient sub-populations from these clinical avatars (Step 4). The decision support rules identify treatment plans that maximizing TTR outcome for patients with a specific type of clinical and genetic characteristics.

Leaf nodes of decision support rules demonstrate precision treatment plans for a specific type of individuals. Figures 4 and 5 further show computational evidence of average TTR outcome and variation when comparing between personalized and one-fit-all treatment plans. When comparing TTR average (Figure 5), the precision treatment plan almost has the highest TTR treatment outcome across all sub-populations from node 1 to 12. Compared with one-size-fits-all treatment plan, personalized treatment plan improved TTR outcomes ranged from 15% to 31% (average and median across 12 leaf nodes are both 24%). On the other hand, the standard deviations are about the middle compared to all other one-fit-all treatment plans. Therefore, we argue that without developing a new treatment plan and protocol, the efficacy can be improved by applying precision-medicine approach to re-use existing treatment options.

Many clinical trial simulations (CTS) were used to access cost-effectiveness[19], health economic[20], and discrete-event[19,21]. Some studies apply CTS to help understand how different genotypes influence dosing decisions[22]. In this study, we used CTS coupled with sub-population optimization to produce a 'practical' precision-medicine approach. Instead of using many patient's clinical and genetic characteristics, a care-giver can use a few variables at the point of care to optimize treatment plan. By using personalized treatment rules in the Figure 3, small, rural clinics can provide optimized treatment regimes without significant increased burden or expensive software and EMR implementation.

In general, limitations can be classified in two categories: the PK/PD model and clustering approaches. (1) Unlike machine learning, which can integrate many variables to predict INR, PK/PD models use only variables (such as partial variables in the Table 2) that were significantly relevant for prediction. (2) PK/PD model was developed with patient populations distinct from the patient population used to construct the clinical avatar model. (3) Clustering approaches, such as the decision tree algorithm, may not create rules of interest for physicians. This limitation will be address by our ongoing work, which will focus on producing various versions of rule-creation approaches. In this study, bleeding and thrombosis risks are equally important in our study. In specialized scenarios, one may be more important than the other risk. We are also working on incorporating constraints and weightings to produce these specialized rules.

Acknowledgement

We thank NIH-1R01LM011566-01 to support this work.

References

1. Chi C-L, Fusaro VA, Patil P, Crawford MA, Contant CF, Tonellato PJ. An approach to optimal individualized warfarin treatment through clinical trial simulations. In: *2010 5th Cairo International Biomedical Engineering Conference*. IEEE; 2010:116-120.
2. Chi C-L, Ravvaz K, Weissert J, Tonellato PJ. Optimal decision support rules improve personalize warfarin treatment outcomes. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2016:2594-2597.

3. Chi C-L, Street WN, Ward MM. Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. *J Biomed Inform*. 2008;41(2):371-386.
4. Chi C-L, Street WN, Robinson JG, Crawford MA. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. *J Biomed Inform*. 2012;45(6):1164-1174.
5. Ravvaz K, Weissert J, Ruff C, Chi C, Tonellato P. Personalized Anticoagulation: Optimizing Warfarin Management Using Genetics and Simulated Clinical Trials. *Circ Cardiovasc Genet*. In press.
6. Hamberg A-K, Dahl M-L, Barban M, et al. A PK–PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. *Clin Pharmacol Ther*. 2007;81(4):529-538. doi:10.1038/sj.clpt.6100084.
7. TETRAD project. http://www.phil.cmu.edu/tetrad/. Accessed February 26, 2017.
8. Fusaro VA, Patil P, Chi C-L, Contant CF, Tonellato PJ. A systems approach to designing effective clinical trials using simulations. *Circulation*. 2013;127(4):517-526. doi:10.1161/CIRCULATIONAHA.112.123034.
9. International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*. 2009;360(8):753-764. doi:10.1056/NEJMoa0809329.
10. Aurora health care warfarin proposal. *Unpubl Stand Protoc Aurora Heal Care*.
11. Anderson JL, Horne BD, Stevens SM, et al. A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). *Circulation*. 2012;125(16):1997-2005.
12. Kimmel SE, French B, Kasner SE, et al. A pharmacogenetic versus a clinical Algorithm for warfarin dosing. *N Engl J Med*. 2013;369(24):2283-2293. doi:10.1056/NEJMoa1310669.
13. Lenzini P, Wadelius M, Kimmel S, et al. Integration of genetic, clinical, and INR data to refine warfarin dosing. *Clin Pharmacol Ther*. 2010;87(5):572-578. doi:10.1038/clpt.2010.13.
14. Rosendaal FR, Cannegieter SC, van der Meer FJ, Briët E. A method to determine the optimal intensity of oral anticoagulant therapy. *Thromb Haemost*. 1993;69(3):236-239.
15. Coppersmith D, Hong SJ, Hosking JRM. Partitioning nominal attributes in decision trees. *Data Min Knowl Discov*. 1999;3(2):197-217. doi:10.1023/A:1009869804967.
16. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Chapman & Hall; 1984.
17. Dean L. *Warfarin Therapy and the Genotypes CYP2C9 and VKORC1*.; 2016. doi:10.1038/clpt.2010.13.
18. Hastie T, Tibshirani R, Friedman JH (Jerome H. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer; 2009.
19. Pink J, Pirmohamed M, Lane S, Hughes DA. Cost-Effectiveness of Pharmacogenetics-Guided Warfarin Therapy vs. Alternative Anticoagulation in Atrial Fibrillation. *Clin Pharmacol Ther*. 2014;95(2):199-207. doi:10.1038/clpt.2013.190.
20. Hughes DA, Walley T. Economic Evaluations During Early (Phase II) Drug Development. *Pharmacoeconomics*. 2001;19(11):1069-1077. doi:10.2165/00019053-200119110-00001.
21. Pink J, Lane S, Pirmohamed M, Hughes DA. Dabigatran etexilate versus warfarin in management of non-valvular atrial fibrillation in UK context: quantitative benefit-harm and economic analyses. *BMJ*. 2011;343:d6333. doi:10.1136/BMJ.D6333.
22. Salinger DH, Shen DD, Thummel K, Wittkowsky AK, Vicini P, Veenstra DL. Pharmacogenomic trial design: use of a PK/PD model to explore warfarin dosing interventions through clinical trial simulation. *Pharmacogenet Genomics*. 2009;19(12):965-971. doi:10.1097/FPC.0b013e3283333b80.

Local ancestry transitions modify snp-trait associations

Alexandra E. Fish

Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA; Departments of Biological Sciences, Biomedical Informatics, and Computer Science, Vanderbilt University, Nashville, TN 37235, USA.
Email: alex.fish@vanderbilt.edu

Dana C.Crawford

Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2530, Cleveland, OH 44106, USA
Email:dana.crawford@case.edu

John A. Capra*

Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA; Departments of Biological Sciences, Biomedical Informatics, and Computer Science, Vanderbilt University, Nashville, TN 37235, USA.
Email: tony.capra@vanderbilt.edu

William S. Bush*

Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2530, Cleveland, OH 44106, USA
Email:wsb36@case.edu

Genomic maps of local ancestry identify ancestry transitions – points on a chromosome where recent recombination events in admixed individuals have joined two different ancestral haplotypes. These events bring together alleles that evolved within separate continential populations, providing a unique opportunity to evaluate the joint effect of these alleles on health outcomes. In this work, we evaluate the impact of genetic variants in the context of nearby local ancestry transitions within a sample of nearly 10,000 adults of African ancestry with traits derived from electronic health records. Genetic data was located using the Metabochip, and used to derive local ancestry. We develop a model that captures the effect of both single variants and local ancestry, and use it to identify examples where local ancestry transitions significantly interact with nearby variants to influence metabolic traits. In our most compelling example, we find that the minor allele of rs16890640 occuring on a European background with a downstream local ancestry transition to African ancestry results in significantly lower mean corpuscular hemoglobin and volume. This finding represents a new way of discovering genetic interactions, and is supported by molecular data that suggest changes to local ancestry may impact local chromatin looping.

1. Introduction

Admixture occurs due to recent mixing of ancestral human populations, and admixed populations represent a unique opportunity to investigate epistasis, or genetic interactions, between alleles with different histories. Prior studies have shown that variants common to any one ancestral population (minor allele frequency/MAF > 5%) are typically shared between all populations,[1] though the frequency at which they occur can vary substantially among different ancestral groups.[2] Lower frequency variants (MAF < 5%) are much more likely to be population specific, and are more likely

* Authors contributed equally to this work

to occur in more ancestral populations. Both recombination rates and the sites of recombination also vary considerably by population; for example, there are more than 2,000 recombination hotspots observed in populations of West African descent, but not European descent populations.[3] Differences in both allele frequency and recombination hotspots between the ancestral populations of an admixed group result in combinations of variants that may have not been observed (or occurred very rarely) in either continental population separately. Given the extensive admixture between human populations over the past several hundred years, many allelic combinations on admixed chromosomes have had limited time to undergo purifying selection and thus may be more likely to influence human traits.

It remains unclear what role epistasis, or genetic interactions, plays in the architecture of human traits. However studies of model organisms suggest that genetic variants likely do not act in isolation, and that the genetic background of a variant may influence its phenotypic effect. For example, it is well-established that when human-derived disease-associated variants are introduced into mice, the phenotypic consequences vary between strains despite consistent environmental factors.[4] This could occur through a variety of mechanisms: strains may carry compensatory mutations that mitigate the effect of the variant, or a given threshold of genetic predisposition (i.e., burden) may be required for phenotypic effects to manifest. Within a natural population, genetic variants that mask the effect of a genomic region may permit potentially deleterious variants to arise. For example, genetic variants that are associated with decreased expression of a transcript can accumulate recently derived rare variants on the same haplotype with limited phenotypic impact.[5,6] Functional haplotypes of regulatory variants also form in which variants either cancel out one another's effects, or in which they both amplify their influence on a phenotype in the same direction.[7]

In this study, we explore admixed chromosomes for new combinations of variants with observable epistasis. As different ancestral chromosomes recombine, disease-associated alleles are placed onto new genetic backgrounds – here we specifically focus on recombination events in close physical distance to variants with established trait associations. Within a dataset of nearly 10,000 adults of African ancestry (an admixed group of European and African ancestral populations), we investigated whether transitions in local ancestry modify the effect of variants on electronic health record (EHR) derived phenotypes. We used genetic data from the Illumina Metabochip, a custom genotyping platform with dense genotyping of previously disease-associated genomic regions, to identify chromosomes with a transition in local ancestry. We then used linear regression models to evaluate the impact of local ancestry transitions on SNP-trait associations reported for African-descent populations in the NHGRI-EBI GWAS Catalog.[8]

2. Methods

2.1. *Subjects and Genotyping*

All samples used in this analysis were part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, which used Vanderbilt University Medical Center's de-identified

biorepository (BioVU) to link patient EHR data with blood-based DNA samples.[9] Details of the consent model and other human subjects issues are described elsewhere.[10] EAGLE selected 9,559 individuals for inclusion in based upon adminstratively-reported African American race,[11,12] rather than for specific health phenotypes, which minimizes ascertainment bias.[13] Samples were genotyped using the Illumina Metabochip, a custom array of almost 200,000 SNPs that targets genomic regions previously associated with type 2 diabetes, obesity, coronary artery disease, and other cardio-metabolic traits for fine-mapping purposes.[14] As part of quality control, variants were removed that did not have at least a 95% genotyping efficiency rate, or that did not vary in this dataset, leaving a total of 192,093 variants for analysis.

2.2. *Local Ancestry Determination*

Local ancestry was assigned using a two-step process: first, we phased the genotype data using SHAPEITv2[15] and the 1000 Genomes Phase 3 reference panel (available for download at https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). There were 171,439 variants that were successfully phased; when variants failed it was typically due to inconsistencies with the reference panel. We then used RFMix[16] (v1.5.4) to assign local ancestry of the phased genetic data, using a window size of 0.1 cM, and a minimum node size of 5. Phased chromosomal haplotypes were matched to ancestral population reference panels. We used all Yoruba (YRI) and CEPH/European (CEU) individuals from 1000G, phase 3v5a, representing African (AFR) and European (EUR) ancestry, respectively.

2.3. *Electronic Phenotyping and Quality Control*

Using the GWAS Catalog,[18] we identified phenotypes and their corresponding EHR trait that had previous associations to regions fine-mapped by Metabochip (Table 1). Across the EHR, individuals have multiple measures for many quantitative traits, such as height/weight/BMI[19] and low-density lipoprotein (LDL) levels. For the majority of traits, we computed the median measurement for each year with data available in the EHR, and then computed the median of these scores. For more rarely collected quantitative traits (i.e. uric acid, serum albumin, etc), we took the median value over all entries. For all phenotypes, we removed clearly non-valid scores (i.e., scores of zero or one), and then removed outliers (those scores more than three standard deviations away from the mean) as quality control.

2.4 *Statistical analyses*

Given that local ancestry is specific to a given chromosome, we performed all analyses on the level of the chromosome, rather than the individual. We used linear regression to determine whether local ancestry transitions interacted with the allele to influence the phenotypes of interest, using the following model:

$$y = A + LA + TRANS + A * TRANS + PC_{1-3} + AGE + GENDER + BMI$$

where y is the phenotype of interest; A corresponds to the allele status (0, absence of the allele; 1, presence of the allele); LA corresponds to the local ancestry at the variant (0, African ancestry; 1,

European ancestry); TRANS indicates the presence of a local ancestry transition within the Metabochip region (0, no transition; 1, no transition); A*TRANS represents the interaction term between the allele and local ancestry transition (1 indicates presence of both the allele and a local ancestry transition; 0 encompasses all other possibilities). AGE, GENDER, and the first three principal components (PC1-3) were included as covariates for all analyses. In the case of binary phenotypes, logistic regression was used. Because this analysis is designed as an investigation of known SNP-trait associations, we used a region-phenotype level Bonferroni multiple testing correction, correcting for the number of tests performed within each Metabochip region.

Table 1. GWAS Catalog traits with genetic associations in Metabochip regions. *median value was taken. In ** the median of yearly medians was taken.

GWAS Catalog Trait	Corresponding EHR Trait	Metabochip Region	Top Reported Gene
Urate levels	Uric Acid*	chr6:25235303-26141375	*SLC17A1*
Type 2 diabetes	PAGE T2D Algorithm[17]	chr11:2444094-2943115	*KCNQ1*
Red blood cell traits	Red Blood Cell Count*; Red Cell Distribution Width*	chr6:25235303-26141375	*HFE*
Iron status biomarkers	Total Iron Binding Capacity*	chr6:25235303-26141375	*HFE*
Weight	Weight**	chr16:53539509-54185787 chr18:57727147-58094636	*FTO* *MC4R*
Hematology traits	Albumin*; Alkaline Phosphatase*; Anion-gap*; Blood Urea Nitrogen*; Calcium*; Cloride*; CO2*; Creatinine*; GluBed*; Glucose*; Hgb*; Potasium*; Mean Corpuscular Hemoglobin*; Mean Corpuscular Volume*; Sodium*; RBC Count*; Red Cell Distribution Width*; Aspartate Aminotransferase*; Alanine Transaminase*; Total Bilirubin*; White Blood Count; MPV*; Platelet Count*; Total Iron Binding Capacity*	chr6:25235303-26141375	*HFE*
Mean platelet volume	MPV*	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr6:25235303-26141375	*ACAD10* *ACAD10* *ACAD10* *LRRC16A*
Height	Height**	chr7:27784039-28282062	*JAZF1*
Obesity-related traits	BMI**	chr16:53539509-54185787	*FTO*
Platelet count	Plt-Ct*	chr12:111290599-113206306	*SH2B3*
Coronary artery disease	Cases at least one ICD-9-CM Codes (410 – 414); all others were controls	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr13:110795080-111049623 chr18:57727147-58094636	*ALDH2* *ALDH2* *ALDH2* *RP11* *PMAIP1*
LDL cholesterol	First LDL-C measurement (with no mention of medication use)	chr1:109655637-110043693 chr1:109789347-109826136	*SORT1* *SORT1*
Body mass index	BMI**	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr16:53539509-54185787 chr18:57727147-58094636 chr3:122976919-123206919 chr3:123039584-123139034	*ALDH2* *ALDH2* *ALDH2* *FTO* *MC4R* *ADCY5* *ADCY5*

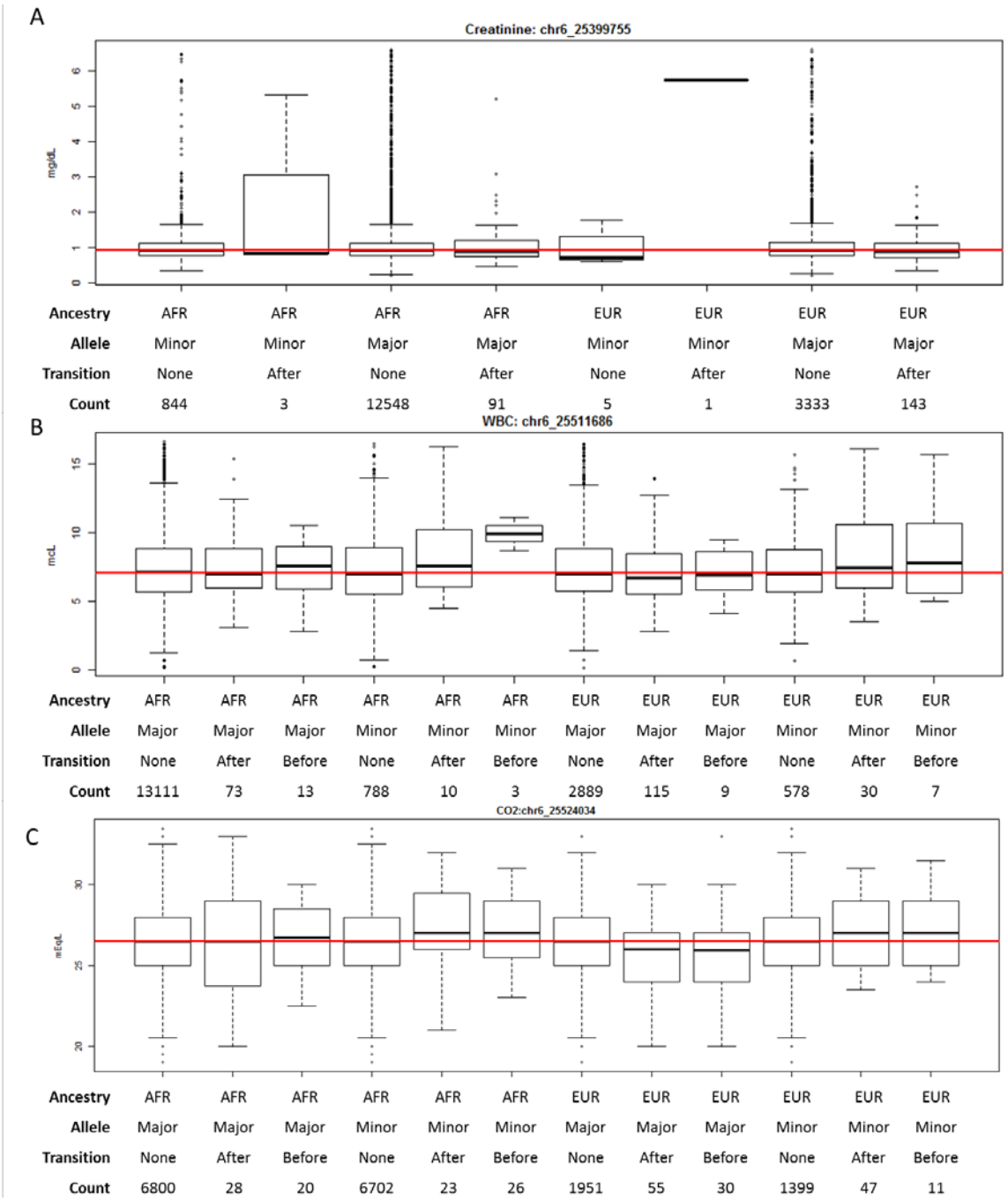


Figure 1. Stratified phenotype distributions reveal interactions between local ancestry transitions and variants regulating creatinine (A), white blood cell (WBC) counts (B), and CO_2 levels (C). Interactions are characterized by stratifying chromosomes based on: the local ancestry at the variant (EUR or AFR); the major or minor allele; and the presence and relative location (upstream/downstream) of a local ancestry transition on that chromosome within the broader Metabochip region. The number of chromosomes observed for each category is provided and reveals that the interactions for creatine and WBC are driven by a small number of chromosomes. The overall median value for each phenotype is represented with a red line. P-values for these interaction tests are given in Table 2.

3. Results

3.1. *Local ancestry transitions interact with variants to influence GWAS Catalog traits*

Using the GWAS Catalog,[18] we identified phenotypes that had previous associations to regions fine-mapped by the Metabochip. We analyzed Metabochip regions with at least 100 local ancestry transitions to provide power to detect the interaction of these transitions with risk alleles. We only considered associations that made reference to African ancestry in the study sample, were for phenotypes that could be readily derived from the EHR, and had at least 200 cases/values in the EHR. This resulted in 28 regions (Table 1), and a total of 57 trait-region pairs.

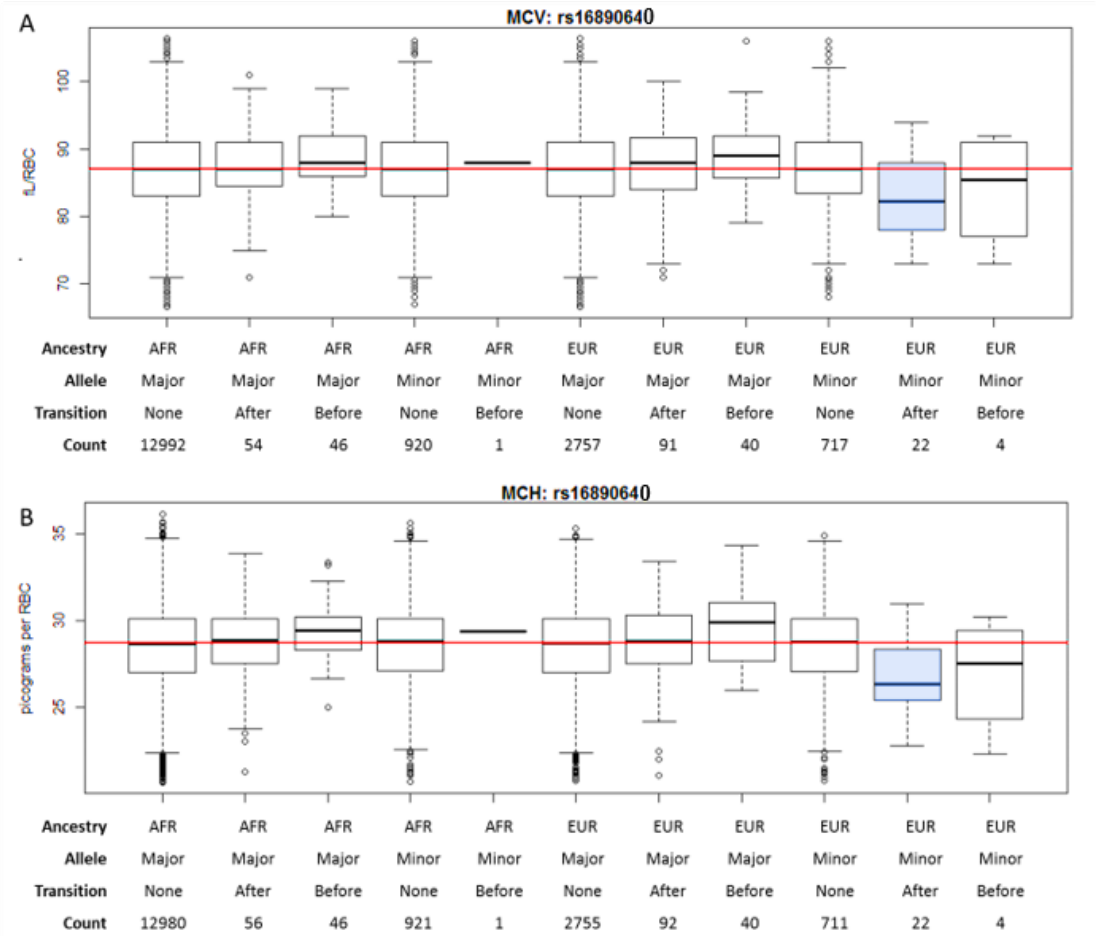


Figure 2. Chromosomes with the minor allele for rs16890640 on a European background and a downstream local ancestry transition are associated with lower MCV (A) and MCH (B). P-values for these interaction tests are given in Table 2.Only one chromosome category was significant with multiple testing corrections for each pairwise test: the minor allele of rs16890649, on a European ancestry, with a downstream local ancestry transition for MCH ($p = 0.0024$, shown in blue).

Due to both differential linkage disequilibrium (LD) structure between populations and differences in SNP coverage in the original GWAS reporting these associations, we tested all variants within the Metabochip region for association with the phenotype. For each trait-Metabochip region pair,

there was at least one nominal genetic association suggesting genetic association within the region. We next evaluated the impact of local ancestry transitions on regional SNP-trait associations.

We identified five significant interactions between local ancestry transitions and the allele across all traits, using a Bonferroni multiple testing correction for all tests within each Metabochip region (adj $P < 0.05$) (Table 2). We then characterized these interactions, grouping chromosomes together based on their local ancestry, allele, and local ancestry transition status. For two interactions, significance is driven by chromosomes that are rarely observed. The interaction for rs9467458 with creatinine levels ($P = 6.90E^{-11}$, Figure 1A) is driven by a single chromosome of European ancestry containing the minor allele, and having a downstream local ancestry transition. The interaction for variant rs4712930 regulating white blood cell (WBC) counts ($P = 8.99E^{-05}$, Figure 1B) is attributable to three chromosomes with African ancestry containing the minor allele, and having an upstream ancestry transition. While these may represent real genetic effects, given the small number of observations within our data, we did not investigate these associations further. The interaction between the variant rs1410438 and CO_2 levels ($P = 7.83E^{-05}$, Figure 1C) shows an interesting pattern in which the minor allele on either ancestral background in the context of an ancestry switchpoint is associated with higher CO_2 levels. This result points to an effect of this region, but because there is no clear biological influence of the ancestry swichpoint, this region was not further investigated.

Table 2. Significant trait associations (all within chr6:25235303-26141375 region) showing SNP by ancestry transition interactions.

Trait	Index Variant	Allele p	Switchpoint p	Local Ancestry p	Switch x Allele p
Creatinine	chr6:25399755	0.759601	2.92E-10	0.639942	6.90E-11
WBC	chr6:25511686	0.170542	0.11056	0.892657	8.99E-05
CO2	chr6:25524034	0.169619	0.048373	0.234598	7.83E-05
MCV	chr6:25577310	0.218143	0.002092	0.153851	7.12E-05
MCH	chr6:25577310	0.071935	0.002658	0.209261	1.24E-05

The two remaining significant interactions for rs16890640 and mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) are illustrated in Figure 2. Both have low numbers of chromosomes in a category; however, these low frequency categories closely resemble the sample median and do not drive the significance of the effect. To identify the categories driving the interactions, we compared each category against the rest of the population with a Mann-Whitney U test. Only one chromosome category was significant after multiple testing correction for each pairwise test; for both traits the interaction is predominantly driven by the minor allele of rs16890640 on a background of European ancestry, with a downstream transition to African ancestry ($p = 0.0024$). A composite of Manhattan plots in the context of local ancestry transitions for MCH associations on this chromosome 6 region is shown in Figure 3.

Mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) are highly correlated with one another, and consequently, the interactions strongly resemble one another. We investigated

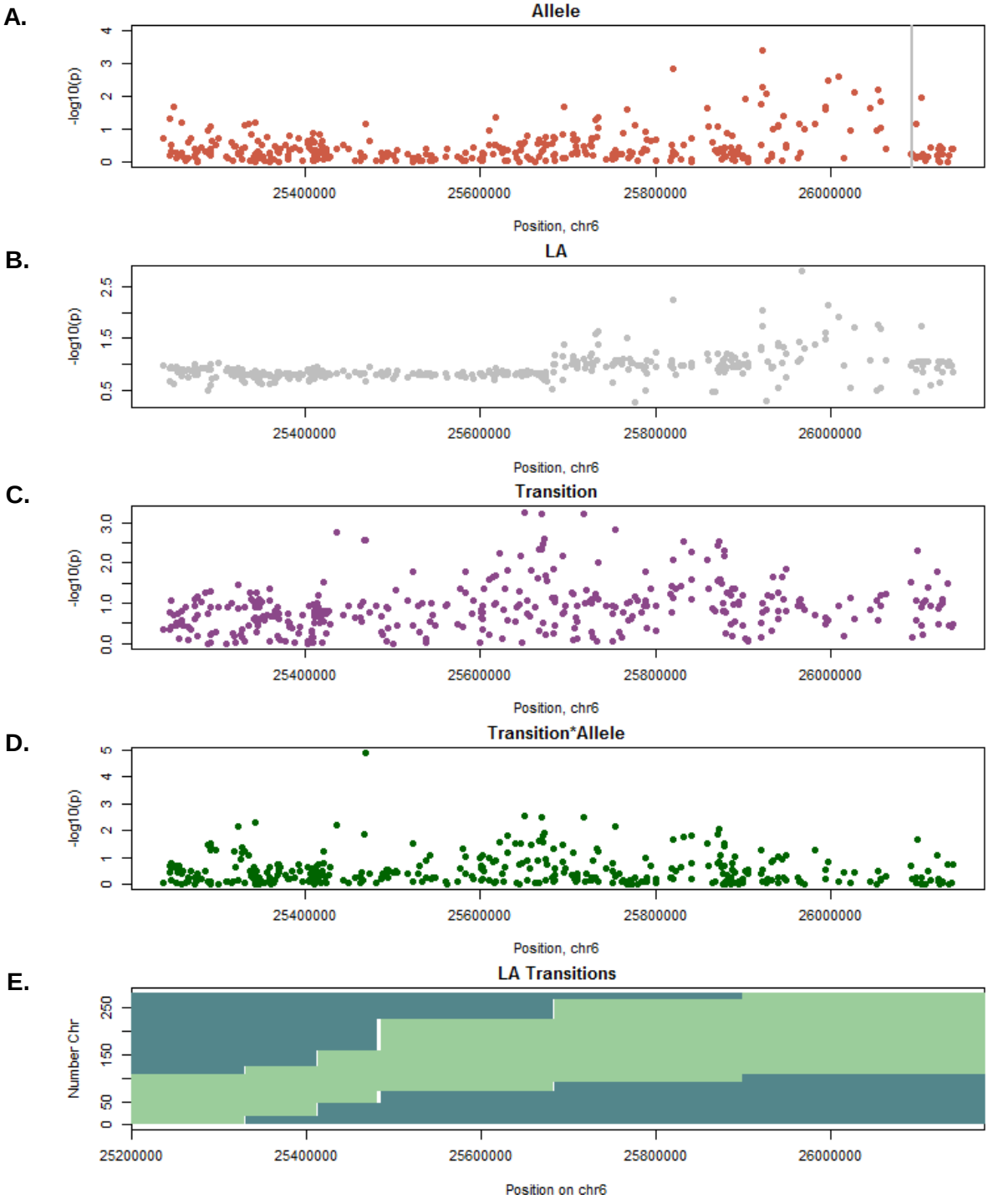


Figure 3. The association of genetic variants within chr6:25235303-26141375 with MCH interacts with local ancestry. Manhattan plots for the effect of the allele (A), local ancestry (B), presence of a local ancestry transition in the region (C), and an interaction between the allele and local ancestry transition (D) – please note differences in scale. The specific local ancestry transitions observed in this region are shown in panel E. Dark green indicates European ancestry along the chromosome; light green indicates African ancestry. rs1800562 (location marked by gray line in panel (A) has been associated with a variety of iron-related phenotypes.

the impact of the downstream transition to African ancestry by stratifying the 22 individuals from the significant chromosome category based on the location of their local ancestry transition. Local ancestry transitions occurred at three downstream points (Figure 3E). We observed a position-dependent effect wherein individuals with a transition to African ancestry at the point nearest to the variant (chr6:25481231) had lower MCH levels (Figure 4). Individuals with transitions to African ancestry at the two subsequent points began to approach the median MCH level. This suggests that the putative functional element interacting with rs16890640 is located between rs16890640 and the second transition point.

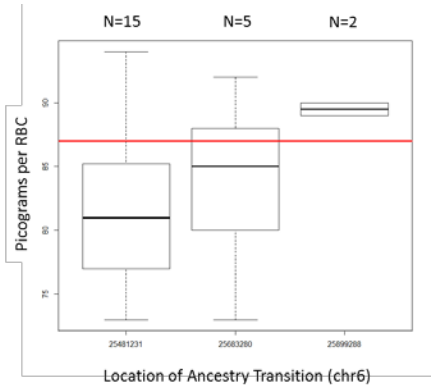


Figure 4. The effect of downstream local ancestry transitions on MCH is position-dependent. Median MCH level is shown in red. The number of chromosomes with European ancestry is provided above each boxplot.

3.2. *Local Ancestry Transitions May Affect Chromatin Looping*

To further characterize rs16890640 and explore potential biological mechanisms mediating its detected interaction with downstream local ancestry transitions, we analyzed its frequency in different populations and genomic context. rs16890640 is roughly three times more common in European-descent populations (EUR = 21%) than it is in African-descent populations (AFR = 8%) based on 1000 Genomes Project Phase 3 frequencies.[20] It occurs within an intron of *CARMIL1 (LRRC16A)*, a cytoskeleton-associated protein involved in regulation of actin polymerization and in megakaryocyte development and platelet production (Reactome Pathway R-HAS-983231). rs16890640 falls within an observed binding site for MAFK, a transcription factor relevant to hemoglobin phenotypes (Figure 5); knock out of MAFK in mice results in reduced MCV and MCH levels.[21] Additionally, it is less than 500 base pairs upstream of a predicted insulator element; however, chromatin looping patterns indicate that contacts occur on either side of this putative insulator (Figure 6). Thus, rs16890640 occurs within a plausibly relevant genomic-region, and is more frequent in Europeans.

We also identified a relatively close (within 20 kb) GWAS-catalog variant associated to a related phenotype, serum transferrin levels (i.e., the amount of glycoproteins that bind free iron).[22] Notably, this variant (rs2274089) is flanked by the two recombination peaks that could result in a local ancestry transition in the area of interest (Figure 5). The first of these recombination peaks is observed in both European (CEU) and African (YRI) descent populations;

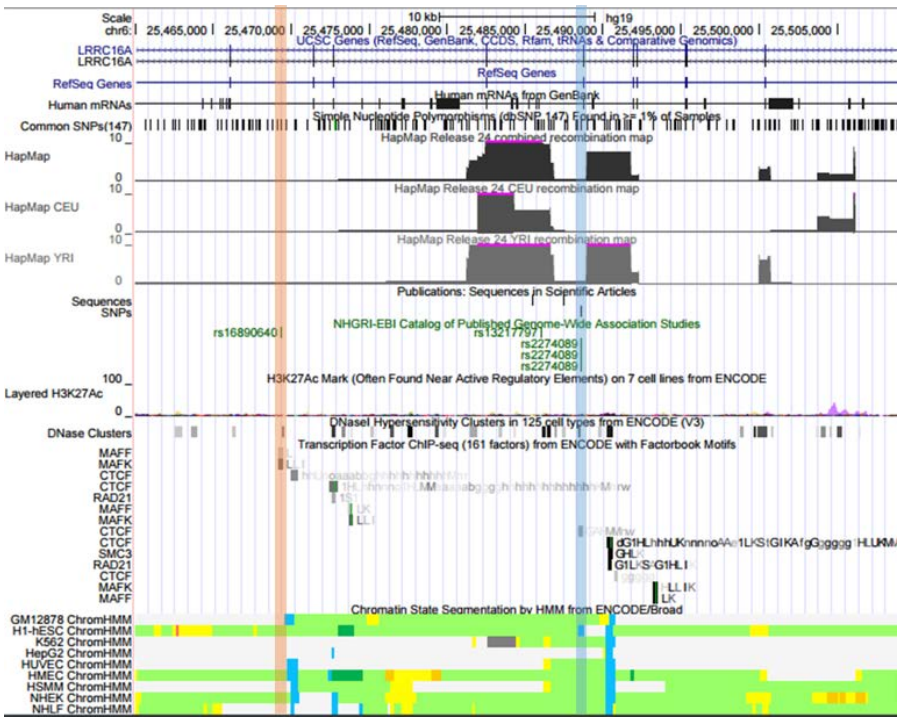


Figure 5. Ancestry-specific recombination hotspots may disrupt functional elements pertinent to MCH and MCV. rs16890640 (orange line) is located within MAFF and MAFK binding sites, and is approximately 500 bp upstream of a predicted insulator element. rs16890640 interacts with a downstream local ancestry transition occuring at one of the two African-specific recombination hotspots shown here, and is proximal to rs2274089 (blue line), a GWAS catalog variant for related traits which overlaps an insulator element.

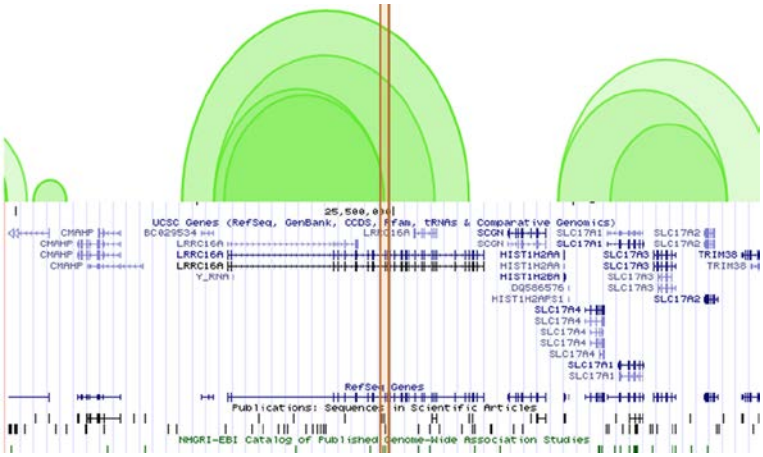


Figure 6. Local ancestry transitions may perturb regional chromatin looping patterns. The African-specific recombination peak region physically interacts with the *CARMIL1 (LRRC16A)* promoter based on ChIA-PET data for RAD21 in the GM12878 cell line. The GWAS variant rs2274089, associated with a relevant phenotype, is highlighted in orange.

however, the second recombination peak is African-specific. This African-specific recombination peak overlaps a ChromHMM predicted insulator element (Figure 5), based largely on the presence of CTCF binding. Chromatin looping data suggest this region may function as an enhancer: the

region contacts the *CARMIL1 (LRRC16A)* promoter in GM12878 (Figure 6). Regardless of whether the region is an enhancer or insulator, it is engaged in regulatory chromatin looping, and we propose that the downstream transition to African ancestry introduces a haplotype that alters the regional chromatin conformation to modify the effect of rs16890640 on MCH and MCV levels.

4. Conclusion

In this study, we hypothesized that the recombination of historically isolated ancestral haplotypes in admixed populations would result in unique combinations of genetic variants that, since they have not been subject to evolutionary pressures, are more likely to influence phenotypes relevant to human health. We investigated this hypothesis in almost ten thousand adults of African ancestry, with both EHR-derived phenotypes and genetic data from targeted regions on the Metabochip. We identified a compelling example that suggests that combinations of haplotypes from different continental ancestries may interact with one another to influence hematological traits in humans.

Chromatin looping is one biological mechanism that may explain our observed statistical interaction between a SNP and a downstream transition to African ancestry. Chromatin looping establishes "domains" in which gene regulatory activity can be confined, keeping the promoters and enhancers for one gene from influencing another. It is possible that the African-specific recombination hotspot (which overlaps putative insulator elements) has introduced low-frequency genetic variants on African-descent haplotypes that alter the insulator's function or epigenetic state. With loss of the insulator, the regulatory variant rs16890640 is then able to engage in 'off-target' effects, which ultimately reduce MCH and MCV levels. Alternatively, an African haplotype may be simply carrying another functional variant that interacts with rs16890640 to influence MCH and MCV, regardless. Ultimately, molecular validation will be required to discern between possibilities.

This approach of examining the modifying role of local ancestry in single variant association studies has several limitations. First, the resolution of local ancestry transitions is limited by the density and proximity of variants along the chromosome that are captured by the genotyping array or imputation. Secondly, we collapsed all local ancestry transitions into a single variable, regardless of where the transition occurred within the region. This introduces additional noise within the data, as not all transitions may have the same effect. Furthermore, by examining each phased chromosome separately, we do not capture *trans* effects between them. In the future, additional models of local ancestry and variants may address these limitations.

The interaction we identified provides potential evidence for epistasis influencing health-related phenotypes in humans. The variant rs16890640 is not significantly associated with the phenotype on its own – it is only in combination with the downstream transition to African ancestry that an association to the phenotype is observed. While it is possible that this haplotype conformation tags a causal variant within this region, we consider this unlikely as nearby variants did not demonstrate a strong association. Instead, it highlights that admixed populations provide a unique opportunity to investigate epistasis, as novel combinations of variants are generated and population-specific recombination hotspots may disrupt functional haplotypes. This further highlights the need to perform genetic studies within admixed populations (as the variants may not have an effect in either continental population) to address health disparities and the role of epistasis in human health.

5. Acknowledgements

Admixture methods development work was supported by the NIH grants RF1 AG054074 and U01 AG052410. Metabochip data generation was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) were obtained from Vanderbilt University Medical Center's BioVU, supported by institutional funding Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH.

6. References

1. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
2. Hinds, D. A. *et al.* Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science (80-.).* **307,** 1072 LP-1079 (2005).
3. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476,** 170–175 (2011).
4. Doetschman, T. Influence of Genetic Background on Genetically Engineered Mouse Phenotypes. *Methods Mol. Biol.* **530,** 423–433 (2009).
5. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7,** e1002144 (2011).
6. Lappalainen, T., Montgomery, S. B., Nica, A. C. & Dermitzakis, E. T. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* **89,** 459–63 (2011).
7. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24,** 1–13 (2014).
8. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
9. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84,** 362–9 (2008).
10. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* **3,** 42–8 (2010).
11. Dumitrescu, L. *et al.* Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet. Med.* **12,** 648–50 (2010).
12. Hall, J. B., Dumitrescu, L., Dilks, H. H., Crawford, D. C. & Bush, W. S. Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. *PLoS One* **9,** e99161 (2014).
13. Crawford, D. C. *et al.* Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits. *Hum. Hered.* **79,** 137–46 (2015).
14. Voight, B. F. *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet.* **8,** 1–12 (2012).
15. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10,** 5–6 (2013).
16. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93,** 278–288 (2013).
17. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform. Assoc.* **19,** 212–8 (2012).
18. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (2014).
19. Goodloe, R. J., Farber-Eger, E., Boston, J., Crawford, D. C. & Bush, W. S. Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record. *AMIA Jt Summits Transl Sci Proc* (2017).
20. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
21. Onodera, K., Shavit, J. A., Motohashi, H., Yamamoto, M. & Engel, J. D. Perinatal synthetic lethality and hematopoietic defects in compound mafG::mafK mutant mice. *EMBO J.* **19,** 1335–1345 (2000).
22. Benyamin, B. *et al.* Identification of novel loci affecting circulating chromogranins and related peptides. *Hum. Mol. Genet.* **26,** 233–242 (2017).

Coalitional game theory as a promising approach to identify candidate autism genes

Anika Gupta*, Min Woo Sun*, Kelley Marie Paskov, Nate Tyler Stockham, Jae-Yoon Jung, Dennis Paul Wall

Departments of Pediatrics and Biomedical Data Sciences, Stanford University
1265 Welch Road, Palo Alto, CA 94305, United States
Email: dpwall@stanford.edu

Despite mounting evidence for the strong role of genetics in the phenotypic manifestation of Autism Spectrum Disorder (ASD), the specific genes responsible for the variable forms of ASD remain undefined. ASD may be best explained by a combinatorial genetic model with varying epistatic interactions across many small effect mutations. Coalitional or cooperative game theory is a technique that studies the combined effects of groups of players, known as coalitions, seeking to identify players who tend to improve the performance--the relationship to a specific disease phenotype--of any coalition they join. This method has been previously shown to boost biologically informative signal in gene expression data but to-date has not been applied to the search for cooperative mutations among putative ASD genes. We describe our approach to highlight genes relevant to ASD using coalitional game theory on alteration data of 1,965 fully sequenced genomes from 756 multiplex families. Alterations were encoded into binary matrices for ASD (case) and unaffected (control) samples, indicating likely gene-disrupting, inherited mutations in altered genes. To determine individual gene contributions given an ASD phenotype, a "player" metric, referred to as the Shapley value, was calculated for each gene in the case and control cohorts. Sixty seven genes were found to have significantly elevated player scores and likely represent significant contributors to the genetic coordination underlying ASD. Using network and cross-study analysis, we found that these genes are involved in biological pathways known to be affected in the autism cases and that a subset directly interact with several genes known to have strong associations to autism. These findings suggest that coalitional game theory can be applied to large-scale genomic data to identify hidden yet influential players in complex polygenic disorders such as autism.

Keywords: Coalitional Game Theory; Autism Spectrum Disorder

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disease with strong genetic etiology. The rapid growth of genome sequencing capabilities has enabled the collection of large datasets and genome-wide association studies in the quest for causative mutations (C Yuen et al. 2017, Hardy and Singleton 2009, Hirschhorn et al. 2002, Iossifov et al. 2014, Leppa et al. 2016, Robinson et al. 2015, Sanders et al. 2015, Weiner et al. 2017, Manolio et al. 2009). Sequencing and data collection efforts such as those undertaken by the Simons Foundation Autism Research Initiative have thus far identified over 100 high confidence genes correlated to ASD and intellectual disability (Sanders et al. 2015, Abrahams et al. 2013). Expression studies have identified an additional 66 genes with a highly correlated regulatory pattern in both blood and brain samples of individuals with ASD (Diaz-Beltran et al. 2016).

Models explaining the genetic architecture of ASD include inherited and de novo alterations and have evolved to include polygenic, additive alterations (de la Torre-Ubieta et al. 2016). Despite an estimated 90% heritability of the disorder, only a small fraction of cases can be explained by known molecular causes (Abrahams and Geschwind 2008), and typical approaches to detect pertinent genes often do not account for intergenic associations. Epistatic interactions have been proposed to explain a portion of the remaining unknown genetic etiology (Phillips 2008, Coutinho et al. 2007). Due to the rarity of many alterations involved in the polygenic models, studies have largely been unable to achieve the statistical power necessary for pinpointing new genetic causes (Gratten et al. 2014, Eichler et al. 2010).

Coalitional game theory (CGT) has been suggested as an enhanced signal detection method that takes into account the combinatorial role of groups of genes in a given condition. This tactic assumes synergy within coalitions to explain a given phenotype, where individual players are represented by genes. Genes that have the greatest average marginal contribution across all coalitions–the genes that play the most in coalitions associated with the disease phenotype–are selected as significant. Previous work has applied CGT approaches for differential gene expression analysis and has successfully identified groups of genes as contributing to Alzheimer's (Vardarajan et al. 2013) and ASD (Esteban et al. 2011, Moretti et al. 2008).

We propose that CGT can be effectively applied to gene alteration data to highlight candidate ASD genes, supporting the polygenic model. We demonstrate an example on 1,965 genomes of individuals in multiplex ASD families (1616 cases and 349 controls). Identifying inherited alterations in genes not previously linked to ASD, such as those presented here, could lead to more accurate diagnoses of the disorder and targeted, proactive therapeutic development against the combinations of underlying molecular causes. This approach could be similarly applied to other complex diseases with combinatorial molecular underpinnings.

2 Methods

2.1 *Data source and preprocessing*

We analyzed 30x-coverage whole genome sequencing data from the Hartwell Foundation's Autism Research and Technology Initiative (iHART), which has amassed one of the largest human disease genome sequencing efforts to-date. Specifically, we assessed the genomes and phenotypic measurements from 756 multiplex families containing at least two children affected by ASD, unaffected parents, and zero or more unaffected siblings.

Families grouped by the phenotypes of their children are as follows (number of children with ASD, number of neurotypical children): 380 (2, 0), 243 (2, 1), 62 (3, 0), 29 (3, 1), 23 (2, 2), 5 (3, 2), 4 (4, 0), 2 (3, 3), 2 (3, 4), 2 (4, 1), 2 (5, 0), 1 (4, 4), and 1 (5, 2) families. Quality control of the sequenced data included removal of non-mendelian variants, which removed sequencing error as well as de novo mutations. Removal of all parents led to an imbalance in the number of cases and controls included in the analysis.

To test the hypothesis of the impact of inherited mutations on the ASD phenotype, we restricted our attention to inherited mutations with highest predicted impact (termed likely gene disrupting, or LGD). We filtered the alteration space to include only loss-of-function variants that had high haplotype-aware consequences (CSQ impact = high) from the variant call format files ($http://samtools.github.io/bcftools$). For each variant, we predicted the inheritance pattern based on the mother's genotype, father's genotype, and the child's genotype. This included autosomal, pseudoautosomal, and sex-linked genes.

We grouped the alterations across all alleles in each gene for each sample, combining homozygous alternate variants and compound heterozygous variants for a given gene. We also included an allele frequency cutoff (keeping only those with a frequency $\leq$0.5) in order to prevent variants with high allele frequency from masking signal. Collapsing the data from allele to gene level, we encoded these alterations into a binary matrix, with 1 indicating the presence of at least one high impact alteration in a given gene for a given sample (homozygous alternate, compound heterozygous, or both), and 0 indicating that no such alteration existed in that gene/sample combination. Only genes with a high impact alteration in at least one sample were kept in the subsequent analyses.

2.2 *Coalitional game theory method*

We applied the coalitional game theory method as described by Moretti et al. (2008) to the iHART alteration data. Coalitional game theory studies the synergy among groups of players. Let N be a finite set of players (in our case, genes). We define a coalition T to be a subset of players ($T \subseteq N$), and we select a score function $v : 2^N \rightarrow \mathbf{R}$ which assigns a score

to every possible coalition. We assume $v(\varnothing) = 0$. The fundamental concept of coalitional game theory is that players working together in a coalition may produce higher or lower scores than the sum of each of those players working individually. This means we assume there are cases such that

$$v(T) \neq \sum_{t \in T} v(\{t\}) \tag{1}$$

Our goal was to identify players who tend to increase the score of any coalition they join. To do this, we calculated the Shapley value for each player. The Shapley value ϕ_i is the marginal contribution of player i over all possible coalitions.

$$\phi_i = \sum_{T \subseteq N : i \in T} \frac{(|T|-1)!(|N|-|T|)!}{|N|!}(v(T) - v(T \setminus \{i\})) \tag{2}$$

For our analysis, we used the unanimity score given by Moretti et. al. (2007). Let m be the number of individuals in our dataset. Each individual j has a high impact alteration in a subset M_j of genes. The unanimity score $u(T, j)$ for coalition T and individual j is 1 if all M_j altered genes are contained in the coalition ($M_j \subseteq T$) and 0 otherwise. We then define our characteristic function v to be

$$v(T) = \sum_{j=1}^{m} u(T, j) \tag{3}$$

Moretti et al. (2007) show that this choice of v makes computing the Shapley values ϕ_i tractable for large datasets. The algorithm took 16.793 minutes to run for 965 genes and 1,965 samples on 1 node of a computer cluster, with 4 cores and 32GB of memory. A detailed example of Shapley value calculation can be found in the electronic supplementary material provided by Moretti et al. (2008).

2.3 *Game theory analysis*

All statistical analyses and applications of coalitional game theory were performed using Bioconductor version 3.5 (http://www.bioconductor.org/) and R version 3.4.0 (http://www.r-project.org/). We removed all genes without any alterations across samples, as the corresponding Shapley values would be 0. This reduced the feature space of genes from 13,853 to 965 genes, resulting in a final boolean matrix of 965 genes and 1,965 samples (1,616 cases and 349 controls). In order to compute the Shapley values, the boolean matrix was split into two matrices corresponding to case and control: B^{case} and $B^{control}$. We adapted the script

from Moretti et al. (2008) to compute the Shapley values using the boolean matrices.

We performed Comparative Analysis of Shapley value (CASh), a resampling-based multiple hypothesis testing procedure introduced in Moretti et al. (2008) to filter out genes with Shapley values that could be high due to chance. We used the function MTP from Bioconductor package "multtest" to compute a CASh p-value for each gene. We ran 1,000 nonparametric bootstrap re-samples with replacement on the matrices. The MTP produces unadjusted p-values for each gene, along with a bootstrap p-value, calculated via simulations. We filtered for only those genes that were significant at the 0.05 and 0.01 significance levels.

Given that our cases and controls could be related siblings and share inherited mutations, we reran coalitional game theory on randomly sampled cases and controls from each family (1 of each per eligible family). As the algorithm is run separately on the case and control binary matrices, none of the samples share familial connection when the Shapley values are computed. The gene identification process via CGT is presented in Figure 1.

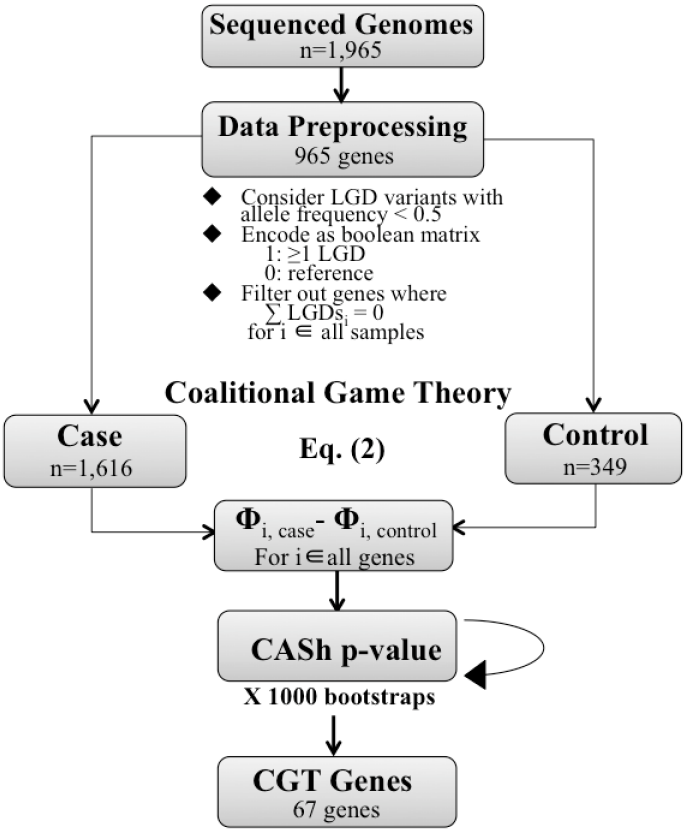


Figure 1. Data analysis flow diagram, starting from the sequenced genomes to identification of significant genes through coalitional game theory.

2.4 *Functional analyses*

To elucidate potential associations with previously-correlated ASD gene candidates, we cross-referenced the CGT gene lists with both the high confidence SFARI genes list (Abrahams et al. 2013) and a set of genes found to be significantly dysregulated in both the blood and brain of individuals with autism, the Root 66 gene list (Diaz-Beltran et al. 2016), using the functional protein network analysis tool STRING (string-db.org). We also checked for network representation of the CGT genes using the Reactome Pathway Browser (reactome.org), a free, open-source, curated and peer-reviewed pathway database. We reported pathways enriched for CGT genes with FDR<0.1.

3 Results

The filtration and binary conversion pre-processing steps yielded 1,616 cases and 349 controls with alteration information for 965 genes. We identified genes (CGT genes) as key contributors in the genetic coordination of ASD using coalitional game theory, as determined by the difference in Shapley value between cases and controls (Figure 1). Sixty-seven genes showed statistical significance at the 0.05 significance level ($p < 0.05$), with 23 of those genes significant at the 0.01 level ($p < 0.01$) (Table 1). Rerunning coalitional game theory on randomly sampled cases and controls from each of the eligible families returned the same genes and confirmed that the family structure of the dataset did not confound the results.

Table 1. Genes selected through coalitional game theory at two levels of significance. The 44 genes listed in p<0.05 are the subset of the 67 genes not in p<0.01.

Significance	Gene symbol
p-value < 0.05	A2ML1, AC008703.1, AC093911.1, ALOX15P2, ATP13A5, BORA, BPIFB5P, C12orf60, C3orf35, CARD8, CCDC26, CCDC7, CDH15, COQ10A, CTC-525D6.1, DUSP16, ERCC6L2, FAM151A, FAM81B, FLG, GBGT1, HLA-K, LGALS8, MAGEC3, MYCT1, OR2T4, OR4Q2, OR6C1, OR8B3, RBAK-RBAKDN, RP11-104E19.1, RP11-160N1.10, RP11-404K5.2, RP11-56H2.2, RP11-618I10.2, RP11-738O11.13, SLC3A1, SSPO, TCP11, TRBV6-7, TRIM48, UBXN11, YME1L1, ZNF99
p-value < 0.01	AF196972.4, AP002856.6, ATP6V1B1, C10ORF68, CDRT15P1, CTB-23I7.1, CTD-2130O13.1, CTD-2509G16.2, GEN1, KRT43P, MDP1, MPRIP, NT5C1B, OR4P4, OR5M10, OR5M11, OR8I2, PRIM2, RP11-15E18.4, RP11-283G6.4, RP11-705C15.2, SSXP3, VWA7

Cross referencing CGT genes with high confidence ASD genes extracted the known

biological functions represented by these candidate ASD genes. Nine of the CGT genes have protein products that directly interact with protein products of genes in the SFARI and Root 66 gene lists, as determined by the functional protein association networks tool STRING (Figure 2).

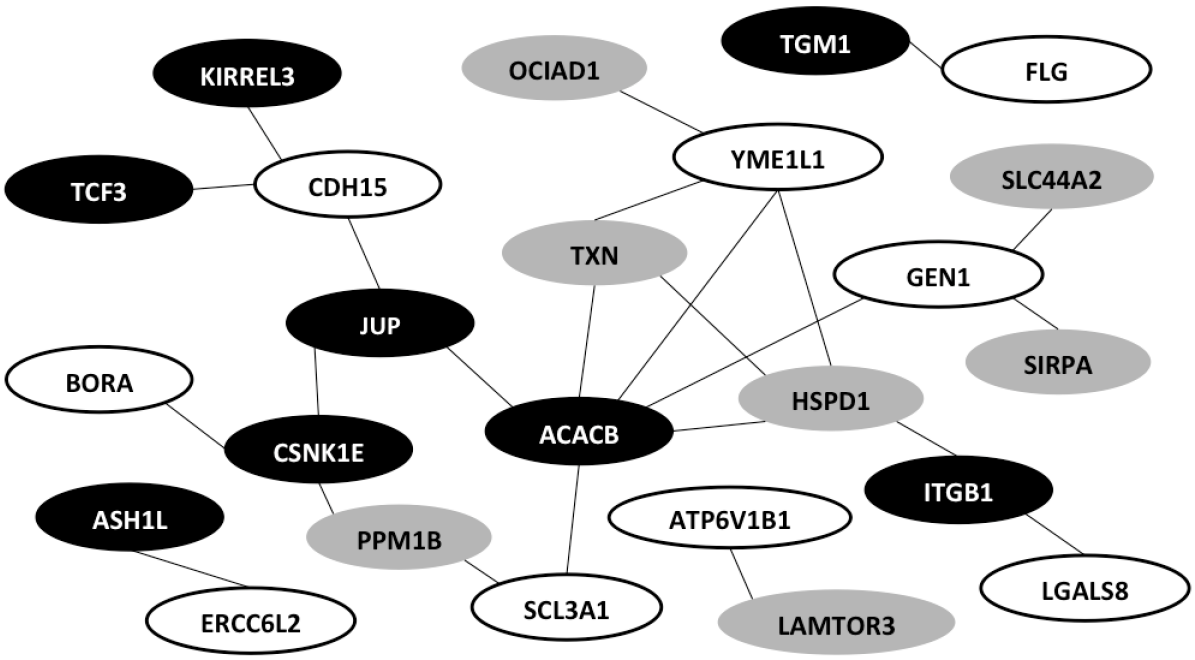


Figure 2. Functional protein interactions between coalitional game theory genes, SFARI genes, and Root 66 genes. SFARI genes are in black, Root 66 genes are in gray, and CGT genes ($p < 0.05$) are in white. Nine CGT genes have direct links to known candidate ASD genes.

Reactome pathway analysis detected 20 significant, non-overlapping functional categories (FDR < 0.1). Of note, pathways representing axon guidance in developmental biology and related neurologic disorders (FDR=0.03 and FDR=0.0047), FGFR1- and insulin receptor-mediated signaling (FDR=0.0042 and FDR=0.034), the innate immune system (FDR=0.09), and olfactory signaling (FDR=0.0252) were enriched for CGT genes. Each of these biological functions has been previously associated with ASD (Ashwin et al. 2014, Park et al. 2016, Lee et al. 2010, Peltier et al. 2007, and Goines and Van de Water 2010).

4 Discussion

Whereas classical genome-wide association studies to pinpoint genes relevant to a biological condition focus on individual genes, coalitional game theory takes into account broader interactions between groups of genes leading to a phenotype. Calculating the Shapley value and filtering for the genes with the highest average marginal contribution over all possible alteration combinations can enable detection of significant gene coaltions and thus boost biologically informative signal. This cooperative view of the alteration landscape more comprehensively accounts for polygenic complexity and may be essential for understanding ASD's genetic architecture.

In this study, we applied cooperative game theory to a large collection of whole genomes from multiplex autism families in an effort to find cooperative signal among coalitions that relate specifically to the autism phenotype. We focused our analysis on inherited gene disrupting mutations to pursue the hypothesis of ASD being a largely inherited, polygenic disorder. Our sample set consisted of 1,616 cases and 349 family-based controls (due to the exclusion of parents), where a gene was assigned a 1 if it contained at least 1 or more such mutations. By analyzing the cooperative contribution of each gene to the phenotype, we found 67 genes that significantly increased the likelihood of a coalition "winning the game" whenever they joined the coalition, where winning refers to the strength of the association to the autism phenotype. Through random sampling, we found that these genes did not appear to be an artifact of the family structure of our data.

Compellingly, 9 of these genes have published links to autism through either DNA or RNA-based analysis (Brown et al. 2015). Further supporting the role of these genes in autism, we found their protein products to be enriched for interaction with the protein products of known autism candidate genes, often interacting just one node away from hubs of connected ASD gene products. Pathway enrichment analyses revealed that signaling pathways in ASD-associated biological functions such as axon development and innate immunity are enriched for CGT genes.

Potential limitations of this work entail the exclusion of alterations in the non-coding regions, as well as the exclusion of low CSQ impact variants within genes. This limits our

ability to find links between the disease phenotype and more subtle genomic and non-coding variation. Exploring such nuanced alterations could shed light on additional high impact, potentially causative molecular states for a disease under consideration. Replication of our analyses and functional characterization will be necessary for comprehensively evaluating the biological implications of our findings. Additionally, many of the genes that were identified via CGT are pseudogenes and have not yet been well studied, making the analysis of such regions even more pressing.

Probing into the specifics of the protein-protein interactions between CGT genes and known ASD candidates, as well as into the groups of co-altered gene coalitions in case subgroups, may provide further mechanistic insights into the underlying molecular causes of ASD. Identifying similarly statistically correlated genes through additional genome-wide ASD gene alteration datasets may elucidate higher degree epistatic connections that can be targeted in both diagnosis and treatment of ASD. Stratifying patients through PCA according to their landscape of co-alterations could improve the precision of diagnosis, and knocking out groups of genes identified in functional assays could reveal potent combinations in therapeutically targeting the molecular underpinnings of ASD.

Coalitional game theory thus serves as a powerful approach to characterize epistatic interactions that may only emerge in a multi-gene model. Capitalizing on the unparalleled rate of genomes being sequenced in increasingly diverse demographic and disease populations, unconventional yet statistically sound tools such as CGT may accelerate the search for biomarkers, particularly in polygenic conditions of mental health.

5 Acknowledgments

This work was supported in part by the Hartwell Foundation award to D.P. Wall and the Hartwell Autism Research and Technology Initiative (iHART).

6 Bibliography

Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S. and Packer, A. 2013. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Molecular autism 4(1), p. 36.

Abrahams, B.S. and Geschwind, D.H. 2008. Advances in autism genetics: on the threshold of a new neurobiology. Nature Reviews. Genetics 9(5), pp. 341–355.

Ashwin, C., Chapman, E., Howells, J., Rhydderch, D., Walker, I. and Baron-Cohen, S. 2014. Enhanced olfactory sensitivity in autism spectrum conditions. Molecular Autism 5, p. 53.

Banerjee-Basu, S. and Packer, A. 2010. SFARI Gene: an evolving database for the autism research community. Disease Models Mechanisms 3(3–4), pp. 133–135.

Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R. and Murphy, T.D. 2015. Gene: a gene-centered information resource at NCBI. Nucleic Acids Research 43(Database issue), pp. D36-42.

C Yuen, R.K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., Pellecchia, G., Buchanan, J.A., Walker, S., Marshall, C.R., Uddin, M., Zarrei, M., Deneault, E., D'Abate, L., Chan, A.J.S., Koyanagi, S. and Scherer, S.W. 2017. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature Neuroscience 20(4), pp. 602–611.

Coutinho, A.M., Sousa, I., Martins, M., Correia, C., Morgadinho, T., Bento, C., Marques, C., Ataíde, A., Miguel, T.S., Moore, J.H., Oliveira, G. and Vicente, A.M. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Human Genetics 121(2), pp. 243–256.

Diaz-Beltran, L., Esteban, F.J. and Wall, D.P. 2016. A common molecular signature in ASD gene expression: following Root 66 to autism. Translational psychiatry 6, p. e705.

Dwyer, C.A. and Esko, J.D. 2016. Glycan susceptibility factors in autism spectrum disorders. Molecular aspects of medicine 51, pp. 104–114.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. and Nadeau, J.H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews. Genetics 11(6), pp. 446–450.

Esteban, E.J., Wall, D.P. 2011. Using game theory to detect genes involved in Autism Spectrum Disorder. 19(1), https://doi.org/10.1007/s11750-009-0111-6, pp. 121-129.

Goines, P. and Van de Water, J. 2010. The immune system's role in the biology of autism. Current Opinion in Neurology 23(2), pp. 111–117.

Gratten, J., Wray, N.R., Keller, M.C. and Visscher, P.M. 2014. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nature Neuroscience 17(6), pp. 782–790.

Hardy, J. and Singleton, A. 2009. Genomewide association studies and human disease. The New England Journal of Medicine 360(17), pp. 1759–1768.

Hirschhorn, J.N., Lohmueller, K., Byrne, E. and Hirschhorn, K. 2002. A comprehensive review of genetic association studies. Genetics in Medicine 4(2), pp. 45–61.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., Smith, J.D., Paeper, B., Nickerson, D.A., Dea, J., Dong, S., Gonzalez, L.E., Mandell, J.D., Mane, S.M., Murtha, M.T., Sullivan, C.A. and Wigler, M. 2014. The contribution of de novo coding mutations to autism spectrum disorder. Nature 515(7526), pp. 216–221.

Lee, E.H., Kim, Y.H., Hwang, J.S. and Kim, S.H. 2010. Non-type I cystinuria associated with mental retardation and ataxia in a Korean boy with a new missence mutation(G173R) in the SLC7A9 gene. Journal of Korean Medical Science 25(1), pp. 172–175.

Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M. and Geschwind, D.H. 2016. Rare inherited and de novo cnvs reveal complex contributions to ASD risk in multiplex families. American Journal of Human Genetics 99(3), pp. 540–554.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M. and Visscher, P.M. 2009. Finding the missing heritability of complex diseases. Nature 461(7265), pp. 747–753.

Moretti, S., van Leeuwen, D., Gmuender, H., Bonassi, S., van Delft, J., Kleinjans, J., Patrone, F. and Merlo, D.F. 2008. Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. BMC Bioinformatics 9, p. 361.

Moretti, S., Patrone, F., Bonassi, S. 2007. The class of microarray games and the relevance for index genes. 15(2), https://doi.org/10.1007/s11750-007-0021-4, pp. 256-280.

Park, H.J., Kim, S.K., Kang, W.S., Park, J.K., Kim, Y.J., Nam, M., Kim, J.W. and Chung, J.-H. 2016. Association between IRS1 Gene Polymorphism and Autism Spectrum Disorder: A Pilot Case-Control Study in Korean Males. International Journal of Molecular Sciences 17(8).

Peltier, J., O'Neill, A. and Schaffer, D.V. 2007. PI3K/Akt and CREB regulate adult neural hippocampal progenitor proliferation and differentiation. Developmental Neurobiology 67(10), pp. 1348–1361.

Phillips, P.C. 2008. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews. Genetics 9(11), pp. 855–867.

Robinson, E.B., Neale, B.M. and Hyman, S.E. 2015. Genetic research in autism spectrum disorders. Current Opinion in Pediatrics 27(6), pp. 685–691.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., Goldberg, A.P., Jinlu, C., Keaney, J.F., Klei, L., Mandell, J.D., Moreno-De-Luca, D., Poultney, C.S., Robinson, E.B., Smith, L., Solli-Nowlan, T. and State, M.W. 2015. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87(6), pp. 1215–1233.

de la Torre-Ubieta, L., Won, H., Stein, J.L. and Geschwind, D.H. 2016. Advancing the understanding of autism disease mechanisms through genetics. Nature Medicine 22(4), pp. 345–361.

Vardarajan, B. N. 2013. Identification of gene-gene interactions for alzheimer's disease using co-operative game theory. ProQuest Dissertations and Theses Global. 1179981110.

Weiner, D.J., Wigdor, E.M., Ripke, S., Walters, R.K., Kosmicki, J.A., Grove, J., Samocha, K.E., Goldstein, J.I., Okbay, A., Bybjerg-Grauholm, J., Werge, T., Hougaard, D.M., Taylor, J., iPSYCH-Broad Autism Group, Psychiatric Genomics Consortium Autism Group, Skuse, D., Devlin, B., Anney, R., Sanders, S.J., Bishop, S. and Robinson, E.B. 2017. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nature Genetics 49(7), pp. 978–985.

Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression[1*]

Binglan Li[1], Shefali S. Verma[1,2], Yogasudha C. Veturi[2], Anurag Verma[1,2], Yuki Bradford[2], David W. Haas[3,4] and Marylyn D. Ritchie[1,2]

[1]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA; [2]Biomedical and Translational Informatics Institute, Danville, PA; [3]Department of Medicine, Pharmacology, Pathology, Microbiology & Immunology, Vanderbilt University School of Medicine, Nashville, TN; [4]Department of Internal Medicine, Meharry Medical College, Nashville, TN, USA

Genome-wide association studies (GWAS) have been successful in facilitating the understanding of genetic architecture behind human diseases, but this approach faces many challenges. To identify disease-related loci with modest to weak effect size, GWAS requires very large sample sizes, which can be computational burdensome. In addition, the interpretation of discovered associations remains difficult. PrediXcan was developed to help address these issues. With built in SNP-expression models, PrediXcan is able to predict the expression of genes that are regulated by putative expression quantitative trait loci (eQTLs), and these predicted expression levels can then be used to perform gene-based association studies. This approach reduces the multiple testing burden from millions of variants down to several thousand genes. But most importantly, the identified associations can reveal the genes that are under regulation of eQTLs and consequently involved in disease pathogenesis. In this study, two of the most practical functions of PrediXcan were tested: 1) predicting gene expression, and 2) prioritizing GWAS results. We tested the prediction accuracy of PrediXcan by comparing the predicted and observed gene expression levels, and also looked into some potential influential factors and a filter criterion with the aim of improving PrediXcan performance. As for GWAS prioritization, predicted gene expression levels were used to obtain gene-trait associations, and background regions of significant associations were examined to decrease the likelihood of false positives. Our results showed that 1) PrediXcan predicted gene expression levels accurately for some but not all genes; 2) including more putative eQTLs into prediction did not improve the prediction accuracy; and 3) integrating predicted gene expression levels from the two PrediXcan whole blood models did not eliminate false positives. Still, PrediXcan was able to prioritize GWAS associations that were below the genome-wide significance threshold in GWAS, while retaining GWAS significant results. This study suggests several ways to consider PrediXcan's performance that will be of value to eQTL and complex human disease research.

Keywords: PrediXcan; GWAS; prioritization; prediction accuracy.

[1*] The project described was supported by Award Number U01AI068636 from the National Institute of Allergy and Infectious Diseases (NIAID) and supported by National Institute of Mental Health (NIMH), National Institute of Dental and Craniofacial Research (NIDCR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health. This work was supported by the AIDS Clinical Trials Group funded by the National Institute of Allergy and Infectious Diseases (AI068636, AI038858, AI068634, AI038855). Additional grant support included AI077505, AI069439, TR000445, AI054999, and AI110527.

Clinical Research Sites that participated in ACTG protocol A5202, and collected DNA under protocol A5128, were supported by the following grants from NIAID: AI069477, AI027675, AI073961, AI069474, AI069432, AI069513, AI069423, AI050410, AI069452, AI069450, AI054907, AI069428, AI069439, AI069467, AI045008, AI069495, AI069415, AI069556, AI069484, AI069424, AI069532, AI069419, AI069471, AI025859, AI069418, AI050409, AI069423, AI069501, AI069502, AI069511, AI069481, AI069465, AI069494, AI069472, AI069470, AI046376, AI072626, AI027661, AI034853, AI069447, AI032782, AI027658, AI-27666, AI058740, and AI046370, and by the following grants from the National Center for Research Resources (NCRR): RR00051, RR00046, RR025747, RR025777, RR024160, RR024996, RR024156, RR024160, and RR024160. Study drugs were provided by Bristol-Myers Squibb Co., Gilead Sciences, and GlaxoSmithKline, Inc.

1. Introduction

Genome-wide association studies (GWAS) have successfully identified disease susceptibility loci for complex traits. Yet, disease related loci discovered to date explain a small portion of the variance in disease risk[1]. It is not known whether the missing heritability is predominantly driven by variants with small effect sizes or by causal factors beyond genic regions. As a consequence, GWAS have relied on increasing sample size which increases the power to find disease-related loci and provides opportunities for rare variant analysis. However, analysis based on larger datasets consume an excessive amount of computational resources, which may not be available to everyone. The excessive number of single nucleotide polymorphism (SNP) loci in comparison to sample size leads to "the curse of dimensionality"[2]. Moreover, loci in intergenic regions may be robustly associated with complex traits, but the mechanisms behind such associations are generally not apparent.

Researchers have been trying to integrate functional genomics into GWAS in the anticipation that mechanistic studies of complex diseases will be facilitated by better interpretation of identified associations[3-6]. Much attention has been paid to the study of regulatory elements that change genes' transcriptional activities and consequently alter phenotypes. Expression quantitative trait loci (eQTLs) are one important class of such regulatory elements[7]. The Genotype-Tissue Expression (GTEx) Project[8] was initiated to identify a comprehensive set of eQTLs from different human tissues and their relationship to gene expression.

PrediXcan[9] is a computational algorithm developed to exploit GTEx data, including eQTLs identification and their relationship to complex traits. PrediXcan evaluates the aggregate effects of cis-regulatory variants (within in 1MB upstream or downstream of genes of interest) on gene expression via an elastic net regression method, and consequently, PrediXcan may identify loci with modest to weak effect sizes that do not achieve significance in variant-based association studies. In theory, PrediXcan has a greatly reduced multiple testing burden as compared to single-variant-single-trait association tests. For example, given one trait and a genotypic dataset of 10 million SNPs, there are at most about 20,000 tests for PrediXcan (~20,000 genes), but 10 million tests for single-variant-single-trait association study. Putative eQTLs and their effect sizes on gene expression level in each GTEx tissue type are available online in PredictDB (http://predictdb.org/).

Several cases have been recently identified where eQTLs are likely to play a causal role in disease by regulating gene expression[26,27]. But while more eQTLs have been identified in recent years, it remains challenging to prioritize the 'true' causal variants. Thus, as PrediXcan is designed to predict gene expression levels and prioritize GWAS results, PrediXcan can also be of great use for mechanistic studies. Here, PrediXcan performance was examined by two datasets where the PrediXcan whole blood models, the most similar tissue type to the samples, were used. One is the genotypic and transcriptomic data of the Yoruba (YRI) cohort from the 1000 Genomes Project[10]. While perhaps not the optimal dataset, it is very accessible which makes it convenient for readers to replicate this study. The other is based on the AIDS Clinical Trials Group (ACTG) protocol A5202[11,12,24], which we refer to as the A5202 cohort hereafter. A5202 cohort has a large enough sample size for evaluating the association tests (see methods) and has underwent a thorough variant-based association study[24] to compare with. To test prediction accuracy, PrediXcan's predicted gene expression levels were compared to the actual gene expression levels measured in the YRI cohort.

We also investigated possible influential factors and filter criterion to increase the possibility of identifying true predictions. As for GWAS prioritization, we carried out a transcriptome-wide association study (TWAS) based on PrediXcan predictions to obtain gene-trait associations and evaluate whether these associations prioritized the GWAS results. Our study provides insight into PrediXcan's capabilities and more importantly eQTL relationships to molecular phenotypes and disease traits, which is of great value in studying transcriptional regulation and disease pathogenesis.

2. Methods & materials

2.1. *Data preparation*

The YRI cohort from the 1000 Genomes Project was used to evaluate the prediction accuracy of PrediXcan for gene expression levels. The YRI cohort comprises 75 individuals. All specimens and 4,395,198 variants passed genotype quality control (based on Hardy-Weinberg Equilibrium ($P > 0.05$) and minor allele frequency (MAF) > 5%). From these 75 individuals, gene expression levels of 23,723 genes in RPKM (Reads Per Kilobase of transcript per Million mapped reads) were provided by the 1000 Genomes Project.

Another 1000 Genomes Project cohort, the Northern Europeans from Utah (CEU) cohort, was also included in this experiment to perform some components of the prediction accuracy test. The CEU cohort comprised 72 individuals and 3,660,275 variants after quality control (Hardy-Weinberg Equilibrium ($P > 0.05$) and MAF > 5%). But since the CEU cohort is part of the Depression Genes and Networks (DGN) cohort that was used to construct the DGN whole blood model by PrediXcan, we did not apply the DGN model to predict expression for the CEU cohort. This is the primary rationale for selecting the YRI cohort for our analyses.

Genotypic and phenotypic data from the A5202 cohort (data based on ACTG protocol A5202[11,12,24]) were used to evaluate PrediXcan's ability to prioritize GWAS results. The A5202 cohort comprises 47% European, 26% African, and 25% Hispanic Americans according to self-reported race or ethnicities. A5202 genotype and imputed data have been previously studied and reported[24]. Imputed genotypic data was quality checked using PLINK and non-ambiguous-stranded variants with imputation score > 0.7, MAF > 1%, and in Hardy-Weinberg Equilibrium ($P > 0.05$) were retained, resulting in 1221 individuals and 5,091,820 variants. Phenotypic data contained 690 continuous traits, which were based on laboratory assay results from HIV-infected patients before and after initiating antiretroviral therapy. The 690 traits were derived from plasma atazanavir pharmacokinetics, plasma efavirenz pharmacokinetics, change in CD4+ T-cell count, fasting low-density lipoprotein (LDL)-cholesterol, and fasting triglyceride data. Details about population structures, phenotypes, genotypes, and GWAS strategy are described elsewhere[24].

2.2. *Heritability Estimation*

To obtain the upper bound of how well a gene expression level can be predicted using genotypic data, we estimated the narrow-sense heritability between SNP variants and gene expression levels. Restricted maximum likelihood (REML) analysis was performed using GCTA[13] for each gene that

is included in both the PrediXcan models and the YRI cohort's gene expression data. Variant-gene relationships were retrieved from the weights table in the PrediXcan models so as to use the same exact set of variants for heritability and prediction accuracy estimations.

2.3. *Performance of gene expression prediction*

PrediXcan provides tissue-specific genotype-expression models, including 44 tissues from GTEx and 1 tissue (whole blood) from DGN[14]. As the 1000 Genomes project uses cultured cell lines derived from blood for genotypic and transcriptomic data, GTEx whole blood and DGN whole blood models were analyzed with the genotypic data from the YRI cohort to predict gene expression levels. The square of Pearson correlation (R^2) between predicted and observed gene expression levels was calculated to measure prediction accuracy. To assess directionality, the Pearson Correlation Coefficient (PCC) between predicted and actual gene expression levels was calculated and is called directionality estimates in the following context. For example, PCC is positive when predicted and observed gene expression levels both increase or decrease at the same time; PCC is negative when the predicted and observed directions are discordant. Of note, some genes had flat predicted gene expression levels across individuals whose genotypes differed. Standard deviations for these predicted gene expression levels were 0, which forced these genes to be dropped from the prediction estimation using R^2 or PCC.

To test which factors influence PrediXcan's prediction accuracy, we examined relationships between a few different model characteristics and accuracy estimates (R^2). For each predicted gene expression level, we evaluated whether the prediction accuracy is influenced by the following model characteristics: 1) the number of variants, 2) the number of variants adjusted by gene length, 3) the percentage of variants over the number of all variants in a PrediXcan model used, and 4) choice of PrediXcan models (tissue specific models). Gene length was annotated using Biofilter[25].

2.4. *Filtering for possibly more accurately predicted genes*

In most experimental data analyses, we have either genotypic data or transcriptomic data, but not both, to perform GWAS or TWAS (see method 2.5. for details). Thus, it is unlikely that we can estimate prediction accuracy or genotype-expression heritability and accordingly select more accurately predicted genes for downstream analyses. To address this issue, we explored whether it is possible to filter the gene list for a subset of more accurately predicted genes without prior knowledge of actual gene expression levels. The filter criterion we tested was based on the similarity of the predicted gene expression levels from the two whole blood models, GTEx and DGN, as predictions from different models will be the easiest to obtain for every PrediXcan users. PCC was used to measure the similarity between prediction results.

2.5. *GWAS Prioritization*

In addition to predicting gene expression levels for individuals who have SNP data but no gene expression data, we also tested PrediXcan's ability to prioritize GWAS results. Some SNP loci may be omitted from mechanistic studies because they only have modest to weak impact on traits and

thus the association signals are not strong enough to pass the multiple testing thresholds set by GWAS or phenome-wide association studies (PheWAS[15]). We were interested in whether PrediXcan could prioritize such association signals. Thus, we carried out PrediXcan followed by TWAS and compared the association hits to PheWAS (since we had multiple phenotypes). To obtain gene-trait association p-values, PrediXcan GTEx whole blood model was applied to the genotypic data from ACTG A5202 to predict gene expression levels. Then predicted gene expression levels and 690 traits were used to perform phenome-wide TWAS via PLATO[16]. Sex, age, and the first three principal components were used as covariates to adjust for sampling biases and underlying population structure. As for variant-trait association studies, to reduce computation time and burden, we only explored the variants within and close to (1MB upstream or downstream) the PrediXcan-TWAS significant genes (Bonferroni-corrected $P < 0.05$). Filtering of variants was done using Biofilter[17]. The criterion of vicinity was in accordance with the region window used by PrediXcan for expression prediction. We then carried out PheWAS using PLATO on the PrediXcan significant traits and the variants nearby PrediXcan significant genes. The association p-values of PrediXcan-TWAS and PheWAS were visualized using *ggplot2*[18] in R.

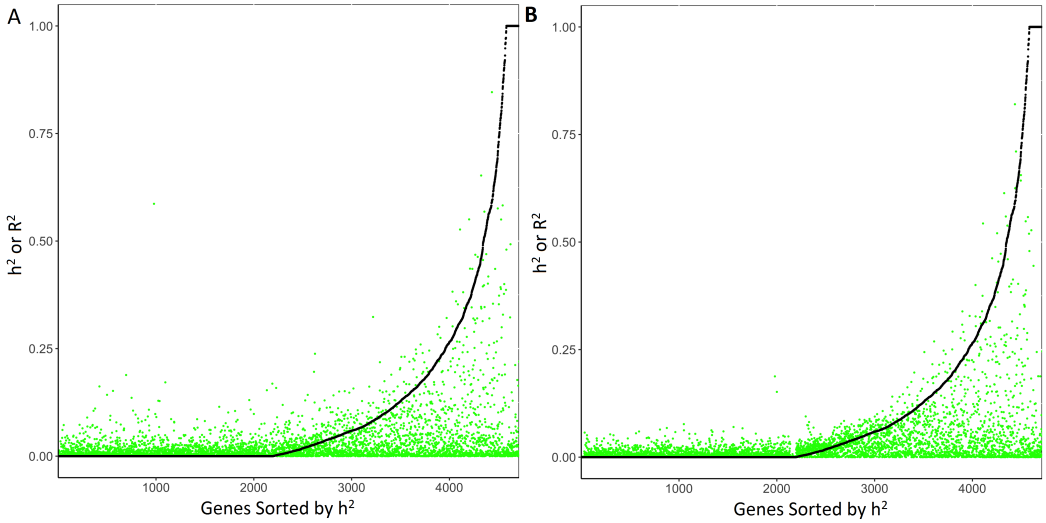


Fig. 1. Prediction performance of DGN (**A**) and GTEx (**B**) whole blood tissue model on the YRI cohort. DGN and GTEx whole blood tissue models were applied to the genotypic data from the YRI cohort. Prediction accuracy (R^2 of predicted versus observed gene expression levels; green) was compared to the narrow-sense heritability (h^2) estimates (black).

3. Results

3.1. *Prediction accuracy*

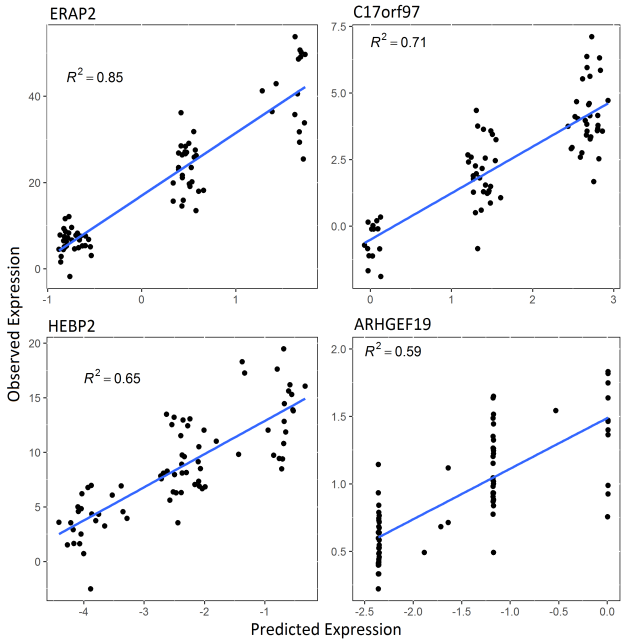


Fig. 2. Examples of well-predicted genes. These plots show the top four performing genes based on PrediXcan's prediction accuracy. Predicted gene expression levels were generated using the DGN whole blood model. Observed expression levels (in RPKM) for the YRI cohort were provided the 1000 Genome Project.

Using the genotypic data of the YRI cohort, the PrediXcan DGN and GTEx tissue models predicted expression of 11,538 and 6,695 genes, respectively. Prediction performance was evaluated using PCC and R^2 for 10,387 DGN genes and 6,127 GTEx genes, respectively (see method 2.3 for why some genes did not have estimates and the justification of using PCC and R^2). Due to the finite number of genes that were common to both models and transcriptomic data, heritability estimation was limited to 4,711 genes.

We first evaluated how well PrediXcan predictions capture the regulatory effects of variants on gene expression levels (**Fig. 1**). We found that genes with higher expression heritability were more likely to have higher R^2 values than genes with lower expression heritability. These results are consistent with what has been published in initial PrediXcan paper[9]. In theory, the better PrediXcan performs at capturing additive regulatory effects imposed by variants, the closer h^2 estimates (black line) and R^2 (green dots) should be, which was what we observed for the genes whose expression levels were influenced by genetic factors ($h^2 > 0$). These results (**Fig. 1**) suggest that PrediXcan predictions were able to capture the transcriptome/gene expression level variability.

We next sought out to evaluate PrediXcan's prediction accuracy. We found that PrediXcan's DGN and GTEx model had similar performance in predicting of gene expression. As indicated in the initial PrediXcan paper[9], PrediXcan precisely predicted gene expressions for some genes (DGN results shown in **Fig. 2**, GTEx results in supplementary figure 2), but prediction accuracy was overall unsatisfactory as most genes had accuracy estimates near 0 (**Fig. 1**). For the two whole blood models, the directionality estimates centered on zero with a small standard deviation, which suggested that most predicted gene expression levels did not correlate with the observed gene expression levels (**Fig. 3**). The GTEx model on the CEU cohort from 1000 Genomes Project performed similarly, with mean of -0.067 and variance of 0.03 (supplementary figure 3). In addition, for all three tests, about one-half of all predictions had negative correlation between predictions and observed values, which made interpretation difficult. In short, based on our evaluation, PrediXcan did not predict gene expression well when DGN and GTEx models were used as training sets to predict gene expression levels in YRI and CEU cohorts. While this finding may not be surprising, many researchers have assumed that PrediXcan could be used for this purpose. Thus, this examination was worthwhile.

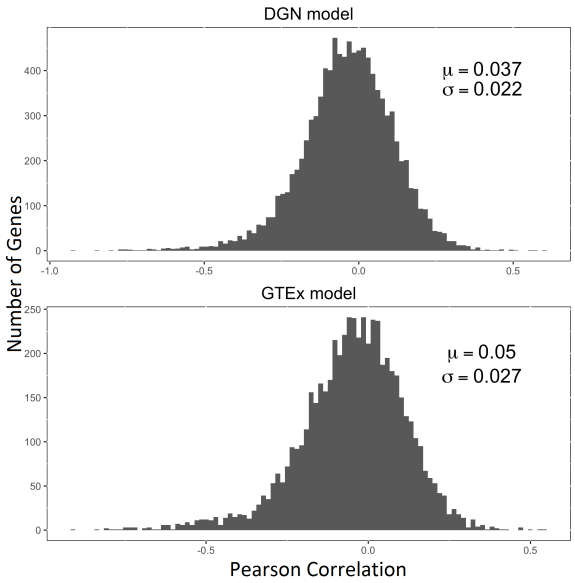


Fig. 3. Performance of prediction directionality of PrediXcan models, DGN (top) and GTEx (bottom), on the YRI cohort. Directionality was computed between predicted and observed gene expression levels.

Next, we examined factors that were responsible for predicting gene expression and more importantly which factors could improve the prediction performance of PrediXcan. We first evaluated whether prediction performance was dependent on specific model properties. For example, would prediction accuracy for a certain gene improve if more variants were included in the input genotypic data for expression prediction? To address this possibility, we explored the relationships between the prediction accuracy and three model properties: 1) the number of model variants used for prediction (**Fig. 4A**); 2) the percentage of the model variants used for prediction (**Fig. 4B**); and 3) the number of model variants used with adjustment for gene length (**Fig. 4C**). A slight improvement in prediction accuracy was apparent in these scatterplots when more variants were taken into account to predict gene expression levels. However, relationships were so weak that these model properties could not be used to favorably assess or improve PrediXcan's prediction performance.

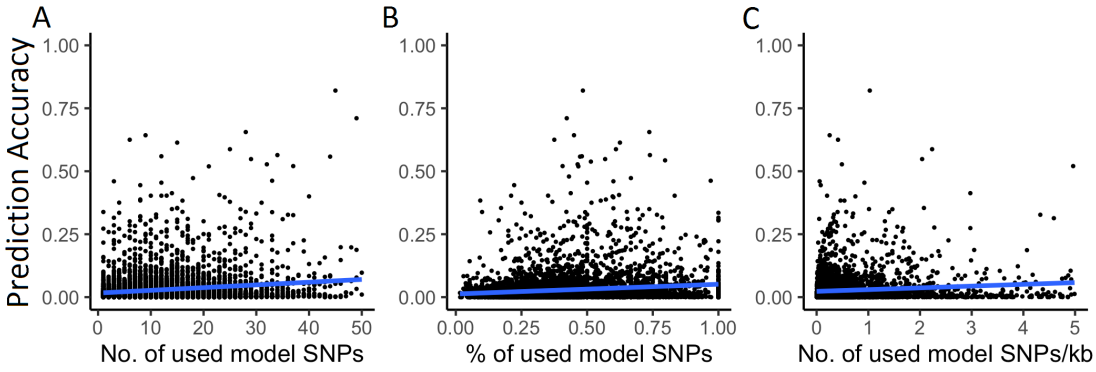


Fig. 4. Prediction accuracy has weak relationship to the model properties. R^2 was computed between observed and GTEx whole blood model predicted expressions. A few genotype-expression model properties were explored, including the number **(A)** and the percentage **(B)** of model variants used for prediction, and the number of used model variants adjusted to gene length **(C)**. But neither of them explained the unsatisfactory prediction, nor could be used as a filtering criterion.

Another potential filtering criterion, the similarity of predicted gene expression levels in the two whole blood models, was also explored. Blood is the most accessible tissue, which makes whole blood models of great practical value and their prediction accuracy critical. The fact that PrediXcan provided two whole blood tissue models offered the opportunity to examine the prediction results based on the two distinct model cohorts. If gene expression was truly regulated by genetic factors, then genotype-expression relationship would be captured regardless of the cohort, and predicted values should be the same given the same genotype data. With this assumption, we hypothesized that the predicted expression for a given gene would likely be more reliable and accurate if the predictions were similar in both whole blood models. As shown in **Fig. 5A**, we selected three sets of genes whose correlations between predicted expression were low, median, or high between the two models. If our hypothesis was correct, we would observe an increase of prediction accuracy from genes with low similarity to those with high similarity, which was indeed what we observed in **Fig. 5B**. The average of prediction accuracy increased from 0.023 to 0.084 for the DGN model and from 0.02 to 0.083 for the GTEx model. In effect, genes whose predicted expressions were more similar between models showed higher prediction accuracy in either PrediXcan whole blood model. However, the filtered results still contained genes whose predicted gene expression levels were directionally different from actual gene expression levels (figures not shown in this paper). In summary, similarity between models was a useful but not ideal filter criterion to improve prediction performance. However, the test of prediction similarity between models can be expanded to using models of different tissue types or using samples from different populations. It may also be worthwhile to investigate genes whose predictions are accurate and similar across models, which could be a good resource or reference set for future investigation of prediction accuracy. In short,

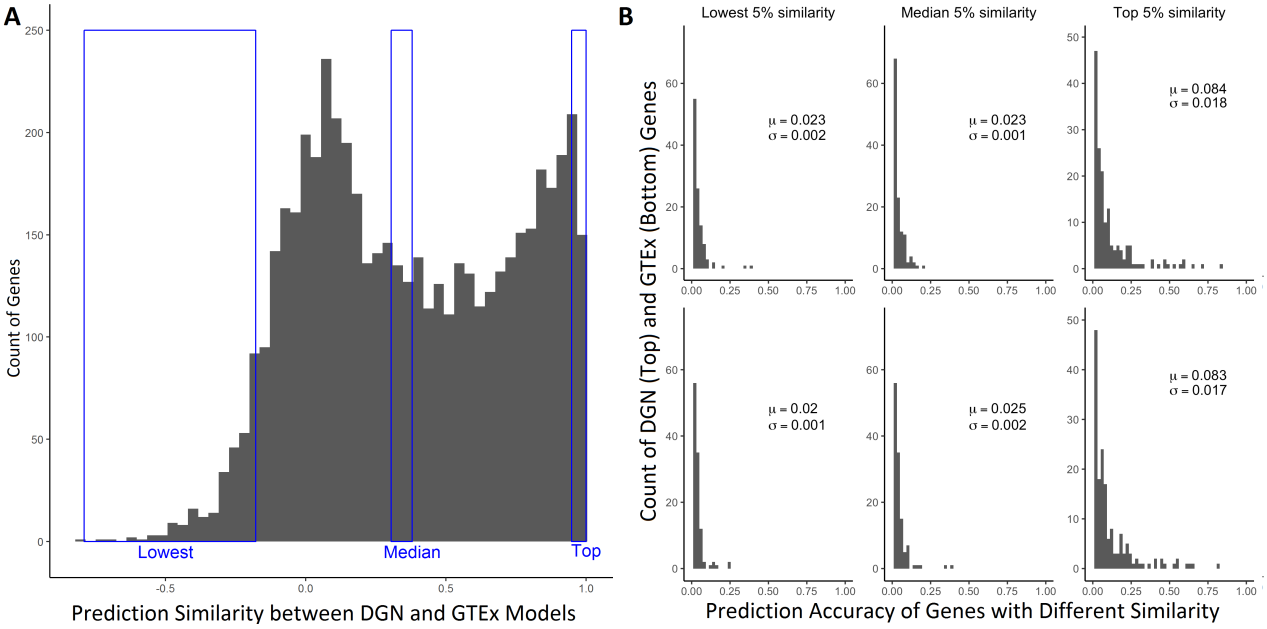


Fig. 5. Prediction similarity between two models has weak, if any, indication on prediction accuracy. Prediction similarity was measured by the Pearson correlation of predicted expressions between the DGN and the GTEx model. **(A)** Distribution of prediction similarity. **(B)** Indication of prediction similarity on prediction accuracy.

many more evaluations could be done with the PrediXcan models or the underlying GTEx data to better understand the SNP-expression relationships in different populations, different tissues, and different genes.

3.2. *Prioritizing GWAS results*

We were also interested in evaluating another use of PrediXcan – prioritization of GWAS results. We wanted to determine whether PrediXcan-TWAS could prioritize important genetic associations that could not be identified by PheWAS due to biological or statistical limitations. To address this question, variant-trait associations that were located within 1MB upstream or downstream of genes were compared to the gene-trait associations identified by PrediXcan-TWAS, using data from the A5202 cohort (**Fig. 6**). Nineteen significant genes identified by PrediXcan-TWAS ($P < 10^{-5}$) were all associated with triglyceride change from baseline to 24 or 48 weeks on treatment. For example, "tgch24_42" in **Fig. 6A** indicates the change in triglyceride from baseline (before starting HIV therapy) to week 24, and was the 42nd phenotype collected. **Fig. 6A** showed that if there were significant variant-trait associations, PrediXcan-TWAS was able to retain the significant signals ($P < 10^{-5}$). This included 3 genes, *DLEU7, DDX1*, and *NARF*. On the other hand, PrediXcan-TWAS prioritized PheWAS associations that almost reached certain significance thresholds ($P = 10^{-5}$; **Fig. 6B**). This highlighted 9 genes – *GPN3, RAP1A, TTC8, SLC5A6, ELOVL7, SUMO1, BAIAP2, OCM,* and *SPRYD4.* The remaining 7 genes had no GWAS association signals in the vicinity regions and thus were likely false positives. Loci within *DLEU7*, *DDX1*, *RAP1A*, *TTC8*, *SLC5A6*, *SUMO1*, and *SPRYD4* were related to triglycerides in previous studies[19,20] according to GRASP[28]. *DDX1* was reported to play a role in HIV-1 infection[21]. More studies are needed to see whether these genes are involved in changes in triglyceride levels on HIV therapy. Other identified genes did not have apparent connections with viral infections or triglycerides, but they could be disease related genes or simply genes that could help to fine-map causal genetic factors. In summary, we demonstrated the ability of PrediXcan to prioritize GWAS results, but the identified gene-trait associations warrant further investigation.

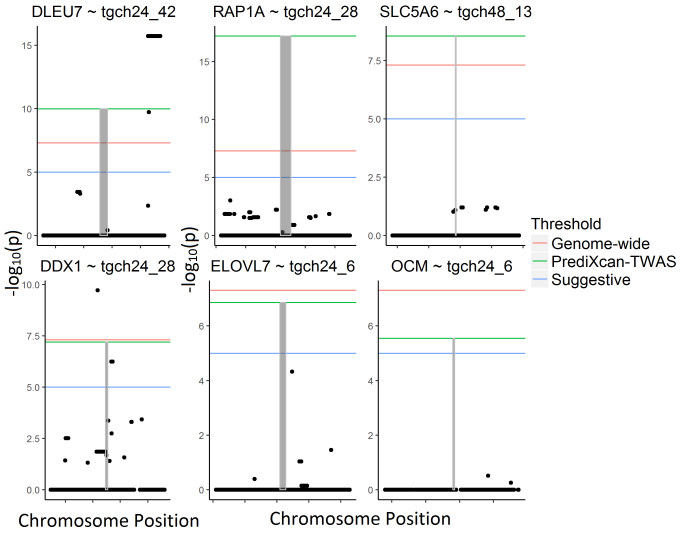


Fig. 6. PrediXcan is able to prioritize GWAS associations. ACTG A5202 imputed genotypic data after quality control was used as input for PrediXcan using GTEx whole blood model and followed by phenome-wide TWAS. Variants within 1MB upstream or downstream of PrediXcan-TWAS significant genes were used to carry out PheWAS. The figures showed the comparison of p-values between PrediXcan-TWAS associations (green line; grey shaded areas represent the size of genes) and PheWAS associations (black dots; blue and red lines denote the suggestive and genome-wide significant p-value, respectively). (A) PrediXcan-TWAS was able to replicate PheWAS results. (B) PrediXcan was able to prioritize non-significant PheWAS results.

4. Discussion

In this study, we carried out a preliminary investigation of the PrediXcan capabilities to predict gene expression levels and to prioritize GWAS signals. If PrediXcan accurately predicts gene expression from SNP data, there could be many potential uses of the algorithm such as imputation of missing transcriptomic data and exploring the biological mechanisms that link genotype to phenotype. But these future analyses are all contingent on the assumption that PrediXcan can accurately predict both the direction of a variants' effect and levels of gene expression. We tested the prediction accuracy of the two PrediXcan whole blood models, DGN and GTEx. PrediXcan was able to accurately predict gene expression for some but not all genes. The slopes of correlation between predicted and actual gene expression levels were negative for almost one-half of genes. This limited the utility of PrediXcan as a transcriptomic data imputation/prediction tool. Several model properties that we explored failed to explain the suboptimal predictions. Dr. Im and her colleagues examined tissues from GTEx and DGN and the results suggested that the local architecture of gene expression traits is simple rather than polygenic[22]. In effect, gene expression is genetically regulated by few rather than multiple eQTLs. This simple local genetic architecture of gene expression might explain why including more putative eQTLs did not improve prediction accuracy in our study. Using prediction similarity between the two whole blood models as a filter improved prediction accuracy somewhat, but did not avoid the negative linear correlation between some predicted and observed gene expression levels. When it came to prioritizing genetic association study results, PrediXcan was able to identify genes that were not significant in GWAS, and also retained significant variant-trait associations. These results were reassuring of the utility of PrediXcan. PrediXcan possessed promising features to reduce research burden by focusing on genes instead of SNPs, and map regulatory effects of distant SNPs onto responding genes, which are overlooked by most studies where only genes adjacent to SNPs are investigated.

Overall, the present study found that PrediXcan performed differently when evaluated for different functions. There are limitations to our study and PrediXcan models. First, whole blood itself is a heterogeneous tissue. And we applied the PrediXcan whole blood model to the YRI cohort whose transcriptomic data actually comes from immortalized blood cell lines. Second is the sample size and population specificity of the test cohort. The YRI cohort (75 individuals) was the most accessible cohort with both genotypic and transcriptomic data, but has a different population structure than the model cohorts from PrediXcan, either DGN or GTEx. While the GTEx cohort includes African Americans, the GTEx model did not yield better expression predictions. To better investigate the influence of population structure and sample size, we would need genotypic and transcriptomic data from multiple populations and of much larger sample sizes. If available, these datasets of different population background will also allow us to explore allelic heterogeneity and population-specific eQTLs. Third, we only evaluated the whole blood models. However, the trait of interest may be regulated by other tissue(s). For example, change of triglyceride in blood may be regulated by metabolism in liver. Thus, it is of biological interest and necessity to explore other tissue models to better understand the tissue specific SNP-expression-trait relationships in the future. Last but not least, PrediXcan is based on two assumptions, 1) loci are equivalent in their functional roles as potential eQTLs, despite the fact that loci at different functional regions may influence gene

expression via different biological mechanisms; and 2) different alleles have the same effect on gene expression. Our study did not specifically evaluate these assumptions. Investigating the relationship of locus functional regions and their roles as eQTLs depends on more detailed annotation and categorization of different types of eQTLs. On the other hand, researchers have looked into allelic expression, which could be a future development for PrediXcan's SNP-expression model design[23].

Although there are challenges, PrediXcan has illuminated a new path for GWAS – incorporating functional genomics and providing mechanistic insights for derived genetic associations. PrediXcan-TWAS results indicated that behind the association, a group of cis-eQTLs regulated gene expression and consequently affected the phenotype. More study is needed to assess PrediXcan's ability to predict gene expression levels and prioritize GWAS results, which will hopefully further our understanding of relationships between eQTLs, gene expression levels, and phenotypes or disease traits.

5. Supplementary

At http://ritchielab.psu.edu/files/PrediXcan_PSB_2018_Binglan_Supplementary_Figures.pdf can supplementary material be found.

6. Acknowledge

The authors are grateful to the many persons with HIV infection who volunteered for A5202 and A5128. In addition, they acknowledge the contributions of study teams and site staff for these protocols. We thank Paul J. McLaren, PhD (Public Health Agency of Canada, Winnipeg, Canada) for prior involvement and collaborations that used these genome-wide genotype data.

References

1. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Cho, J. H. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.
2. Van Steen, K. (2011). Travelling the world of gene–gene interactions. *Briefings in bioinformatics*, *13*(1), 1-19.
3. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., ... & Cherry, J. M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, *22*(9), 1790-1797.
4. Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nature biotechnology*, *28*(10), 1057-1068.
5. Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, *347*(6222), 664-667.
6. Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., ... & Bitton, A. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*, *40*(8), 955-962.
7. Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197-212.
8. GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648-660.
9. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., ... & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, *47*(9), 1091-1098.

10. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.
11. Sax PE, Tierney C, Collier AC, et al. Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. *N Engl J Med* 2009; 361:2230-40.
12. Daar ES, Tierney C, Fischl MA, et al. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1. *Ann Intern Med* 2011; 154:445-56.
13. Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76-82.
14. Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., ... & Urban, A. E. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, *24*(1), 14-24.
15. Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., ... & Deelman, E. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS genetics*, *9*(1), e1003087.
16. Grady, B. J., Torstenson, E., Dudek, S. M., Giles, J., Sexton, D., & Ritchie, M. D. (2010). Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 315). NIH Public Access.
17. Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 368). NIH Public Access.
18. Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
19. Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., ... & Johansen, C. T. (2010). Biological, clinical, and population relevance of 95 loci for blood lipids. Nature, 466(7307), 707.
20. Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., ... & Voight, B. F. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics, 41(1), 56-65.
21. Fang, J., Acheampong, E., Dave, R., Wang, F., Mukhtar, M., & Pomerantz, R. J. (2005). The RNA helicase DDX1 is involved in restricted HIV-1 Rev function in human astrocytes. *Virology*, *336*(2), 299-307.
22. Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., ... & GTEx Consortium. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS genetics, 12(11), e1006423.
23. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome biology*, *16*(1), 195.
24. Verma A, Bradford Y, et al. Multiphenotype association study of patients randomized to initiate antiretroviral regimens in AIDS Clinical Trials Group protocol A5202. *Pharmacogenet Genomics* 2017; 27:101-11.
25. Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (p. 368). NIH Public Access.
26. Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., ... & Mahaney, M. C. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics, 39(10), 1208.
27. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., ... & Pirruccello, J. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature, 466(7307), 714.
28. Leslie, R., O'Donnell, C. J., & Johnson, A. D. (2014). GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics, 30(12), i185-i194.

Considerations for automated machine learning in clinical metabolic profiling: Altered homocysteine plasma concentration associated with metformin exposure

Alena Orlenko[*1]; Jason H. Moore[*1]; Patryk Orzechowski[1,2]; Randal S. Olson[1]
1 - Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA;
2 - *Department of Automatics and Biomedical Engineering, AGH University of Science and Technology, Krakow, Poland*

Junmei Cairns[1]; Pedro J. Caraballo[2]; Richard M. Weinshilboum[1]; Liewei Wang[1†]
1- Department of Pharmacology and Experimental Therapeutics, Mayo Clinic
2 - Department of Endocrinology, Mayo Clinic
Rochester, MN, USA
Email: Wang.Liewei@mayo.edu

Matthew K. Breitenstein[†]
Institute for Biomedical Informatics and Center for Pharmacoepidemiology, Research, and Training, University of Pennsylvania
Philadelphia, PA, USA
Email: mkbreit@upenn.edu

With the maturation of metabolomics science and proliferation of biobanks, clinical metabolic profiling is an increasingly opportunistic frontier for advancing translational clinical research. Automated Machine Learning (**AutoML**) approaches provide exciting opportunity to guide feature selection in agnostic metabolic profiling endeavors, where potentially thousands of independent data points must be evaluated. In previous research, AutoML using high-dimensional data of varying types has been demonstrably robust, outperforming traditional approaches. However, considerations for application in clinical metabolic profiling remain to be evaluated. Particularly, regarding the robustness of AutoML to identify and adjust for common clinical confounders. In this study, we present a focused case study regarding AutoML considerations for using the Tree-Based Optimization Tool (**TPOT**) in metabolic profiling of exposure to metformin in a biobank cohort. First, we propose a tandem rank-accuracy measure to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors. Second, while AutoML, using default parameters, demonstrated potential to lack sensitivity to low-effect confounding clinical covariates, we demonstrated residual training and adjustment of metabolite features as an easily applicable approach to ensure AutoML adjustment for potential confounding characteristics. Finally, we present increased homocysteine with long-term exposure to metformin as a potentially novel, non-replicated metabolite association suggested by TPOT;

[*] Contributions equivalent for first authorship consideration
[†] Contributions equivalent for senior and corresponding authorship consideration

an association not identified in parallel clinical metabolic profiling endeavors. While warranting independent replication, our tandem rank-accuracy measure suggests homocysteine to be the metabolite feature with largest effect, and corresponding priority for further translational clinical research. Residual training and adjustment for a potential confounding effect by BMI only slightly modified the suggested association. Increased homocysteine is thought to be associated with vitamin B12 deficiency – evaluation for potential clinical relevance is suggested. While considerations for clinical metabolic profiling are recommended, including adjustment approaches for clinical confounders, AutoML presents an exciting tool to enhance clinical metabolic profiling and advance translational research endeavors.

Keywords: Clinical metabolic profiling; Automated machine learning; Confounding; Metabolomics; Pharmacometabolomics; Metformin; Homocysteine; Biobank; Precision medicine

1. Background

1.1. *Introduction to metabolomics and study motivation*

Metabolomics, the study of organic chemical signatures within a specimen, has been increasingly deployed in clinical research applications. Characterization of perturbations to the metabolome (a.k.a. phenome) hold great promise to elucidate novel biomedical insights and potential disease mechanisms. While many 'omics perspectives provide unique molecular insights, the phenome reflects biological perturbation closest to clinical phenotype manifestation. With the proliferation of biobanks [1] – where consenting patients voluntarily donate a wide-array of biologic specimens (e.g. blood, urine, saliva) to be systematically stored and utilized for research – opportunities for secondary research applications, including metabolomics, using primary specimens abound [2]. As both the science of metabolomics advances and scale of biobanks increase, clinical metabolic profiling holds increasing promise to identify novel biological insights regarding disease state, drug response, and clinical heterogeneity [3].

Metabolic profiling is a multi-step process that 1) initiates with analytical chemistry measurement {e.g. liquid chromatography (**LC**), mass spectrometry (**MS**), Nuclear magnetic resonance spectroscopy (**NMR**)}, including deployment of tandem techniques such as LC/MS, of organic compounds contained within a biological specimen; 2) algorithmic association of raw measurements with known discrete metabolites; 3) establishment of relative metabolite concentrations; and 4) concluding with statistical generation of a profile of metabolites (i.e. metabolic profile) perturbed within the phenome given an exposure of interest. Metabolic profiling of both disease [4] and drug exposures [5] have successfully identified distinct signatures. While some of these features have been shown to remain stable over time [6], some metabolites and physiologic states are known to rapidly fluctuate [7]. Further, some metabolites are known to be lipid soluble, with measured concentrations noticeable altered in patients with elevated BMI [8] – a biological rationale for potential confounding by BMI in clinical metabolic profiling. For agnostic untargeted metabolic profiling, thousands of metabolites are identified, whereas for targeted metabolic profiling, only a small group of metabolites are selected *a priori* based on hypothesized biological relevance. With untargeted metabolic profiling, distinct analytical challenges remain as thousands of potentially unique features are ascertained, frequently exceeding the number of

samples analyzed. Augmenting the metabolic profile with other 'omics perspectives only further enhances this complexity. Regardless of selected approach and application, great opportunity exists for semi-automated machine learning approaches to assist in agnostic selection and inclusion of features in metabolic profiling endeavors. Our current application is focused within the later part of the targeted metabolic profiling process, characterizing long-term exposure to the drug metformin as a monotherapy within a human biobank cohort.

1.2. *Automated Machine Learning and TPOT*

Machine Learning (**ML**) approaches hold great opportunity to enhance metabolic profiling endeavors. Tree-based optimization tool (**TPOT**), our specific tool of interest, is an Automated Machine Learning (**AutoML**) tool with recent demonstrable success. Specifically, TPOT has been observed to automatically optimize ML pipelines that match or exceed the performance of traditional supervised approaches [9, 10, 11] while requiring minimal adjustments to default parameters. The mixture of data types deployed in human metabolic profiling and expansive feature space are ideally suited for enhancement with AutoML approaches. In genomics applications [9], TPOT has delivered promising predictive performance while being demonstrably robust to mixed datatypes with large feature spaces. Given the mixture and expansive feature space of data types found in clinical metabolic profiling, we posit that AutoML approaches offer opportunity for a robust, agnostic profiling solution. However, a thorough evaluation of potential caveats and considerations for application of AutoML using TPOT in clinical metabolic profiling is necessary. This includes specific considerations regarding continuous metabolite features in a potentially expansive feature space.

In this study, we provide an annotated methodological case study applying AutoML in clinical metabolic profiling of patients exposed to metformin monotherapy. Patient data was collected previously for traditional clinical metabolic profiling endeavors [12] from patients nested within a biobank cohort [2] Highlighted within our methodological case study are the following items: 1) A focused overview of clinical metabolic profiling using automated supervised machine learning methods. Specifically, demonstrated using the TPOT tool; 2) Necessary pre-processing and analysis steps for development of a clinical metabolic profile using AutoML; 3) Current state of the art for identification of confounding characteristics using AutoML; 4) Proposed strategies to adjust for confounding characteristics of different types commonly encountered in clinical metabolic profiling; 5) Finally, propose an AutoML-based tandem rank-accuracy metric for agnostic data-driven feature selection in clinical metabolic profiling.

2. Methods

All analyses and experiments, described in-depth below, were conducted using Python programming packages Sklearn, Pandas, and Numpy. All figures were generated using Python Matplotlib and

Seaborn programming packages. TPOT v0.8 software [9,10,11] <https://github.com/rhiever/tpot> was exclusively utilized for AutoML experiments.

2.1. *TPOT overview*

Fundamentally, TPOT takes a supervised learning dataset as input and recommends a series of preprocessing, feature construction, feature selection, and ML modeling operations that maximize the predictive performance of the final ML model. We call this series of operations a pipeline. TPOT optimizes the analysis pipeline using a stochastic optimization process that begins with several simple, random pipelines (the *population*). For every iteration of the optimization process (a *generation*), TPOT makes several copies of the current best-performing pipelines in the population and then applies random changes to them, such as adding or removing an operation or tuning a parameter setting of one of the operations. These stochastic changes can have positive or negative effects on the performance of the pipelines, and as such allow TPOT to explore new analysis pipelines that were never previously considered. At the end of every generation, the worst performing pipelines are removed from the population and TPOT proceeds to the next generation. After a fixed number of generations (in this study, 1,000 generations), TPOT recommends the best-performing pipeline that it ever created during the optimization process. In this study, we present observations from the TPOT pipeline as described. For more details regarding the TPOT algorithm and tool, see [9,10,11] and the software package online at https://github.com/rhiever/tpot

2.2. *TPOT default parameters and pre-processing*

A series of pre-processing steps were initiated to evaluate sensitivity of TPOT for detection of common clinical confounders (e.g. age, gender, body mass index (**BMI**), batch effects). Supervised classification analysis was performed using an out-of-the-box TPOT deployment. A classification predictive model was generated on the full dataset containing both prioritized metabolite features and clinical covariates using the following settings: number of generation 2000, population size 1000, 5-fold cross validation on the training set, and standard accuracy as a performance metrics. Prior to TPOT analyses, the cohort was randomly stratified into separate 75% training and 25% testing datasets. A unique random seed was selected for each of the 5 independent replications. Resulting models were characterized using accuracy metrics, representing the fraction of corrected prediction with the best possible score 1.0. For each replicate, feature importance was measured and rank was assigned in accordance with importance coefficients. Ranks were summed across replicates where inverse of sum of ranks served as the metric for feature importance across experiments. Specifically, rank coefficient or r_x, where x is a feature coefficient from replicate i, n – total number of the TPOT replicates: $r_x = 1/\sum_{i=1}^{n} x$

2.3. *TPOT Analysis*

TPOT models generated predictive ranks, an approximation of relative effect size generalizable across TPOT-selected machine learning algorithms, for comparing importance of individual features. Model performance overall was evaluated using R^2, or coefficient of determination (i.e. accuracy), describing the fraction of response variance described by the model, with a maximum possible score of 1.0. In our work, we highlighted the potential utility of rank-accuracy measures deployed in tandem to guide agnostic feature selection in clinical metabolic profiling.

To evaluate TPOT's automatic adjustment capabilities in clinical metabolic profiling, we evaluated the following features for potential confounding: 1) BMI – metabolites evaluated using case-only (metformin monotherapy exposed) and stratified {split by the median value of BMI (2 groups) and common clinical thresholds (<18.5, 18.525, 25-30, ≥30)} datasets; 2) Batch effect – 8 splits were applied using TPOT classification mode with case status set as the target variable. For each replicate, feature importance was measured and predictive ranks were assigned; 3) Dose-dependent metabolite effect of metformin exposure – case-only analysis was performed where prescribed metformin daily dose and measured metformin plasma concentration were included; 4) Confounding associations (i.e. sensitivity) not identified by TPOT (i.e. insensitive) – metabolites were adjusted using either stratification or residual adjustment approaches, both of which are described in-depth below.

To demonstrate the utility of TPOT for feature selection towards ascertainment of a metabolic profile, classification TPOT analysis was then applied to a reduced-feature dataset containing only the prioritized metabolite features (i.e. potential confounding clinical covariates were removed). TPOT settings described in the previous section were utilized. Predictive ranks were deployed to aid in feature selection of metabolic profiles; metabolites were sorted in accordance to their rank coefficient and recursive feature elimination estimated the strength (e.g. accuracy score) of prediction for various consecutive combination of sorted features. Pipelines from TPOT analyses evaluated performance of various feature sets by reporting training and testing set accuracy. To understand the impact of potential confounding insensitive to TPOT, the ascertained clinical metabolite profile was replicated using a BMI-adjusted dataset, where metabolite measured concentrations were replaced with residuals from independent univariate linear regression models of individual metabolites and BMI.

3. Results

3.1. *Cohort characteristics*

Exposure to clinically stable (i.e. 'long-term') metformin monotherapy was profiled using a de-identified case-control dataset representing 546 unique patients nested within a biobank cohort. All data was previously collected for parallel metabolic profiling endeavors [12]; data was de-identified by our collaborating institution prior to release for analysis. IRB coverage for both prior research data collection (original study purpose) [Mayo Clinic 15-003347 and 08-007049] and secondary analysis

[Penn 827996] were obtained. A pre-selected panel of amine-based metabolites (n=42) were previously measured from human plasma samples using tandem LC-MS. Clinical features were previously ascertained using electronic health record(**EHR**)-based phenotyping and confirmed by manual chart review. Clinical features included common covariates (age, gender, BMI, and metabolite batch) and metformin exposure (metformin prescribed daily dose and metformin plasma concentration). Cases (n=273) included patients exposed to metformin monotherapy with type 2 diabetes having glycemic control; controls consisted of healthy normal patients with no known metformin exposure. Case and control patients were previously matched by age and gender prior to sample selection and were statistically randomized and assigned to batches prior to metabolomic measurement.

3.2. *Descriptive analyses using default TPOT parameters*

Univariate Pearson's correlations were generated for metformin exposure (case-control status) using TPOT (**Figure 1**). Metformin exposure demonstrated correlations with varying effect and direction for both metabolite (e.g. alanine and citrulline) and clinical (BMI, metformin prescribed daily dose, metformin plasma concentration) features. These results demonstrated that associations exist within the dataset, suggestive of potential confounding between metformin exposure and metabolic perturbation. Further, strong associations were identified for several metabolite-metabolite pairings. For example, tyrosine, valine, isoleucine, leucine and phenylalanine were positively correlated ($r > 0.5$) with each other. While physiologically unclear, such distinct clustering of associations might suggest potential for proximal, interrelated metabolic responses.

From the perspective of rank associations, we evaluated sensitivity of TPOT to identify potential confounding features using default parameters. Not surprisingly, increased prescribed daily dose of metformin was associated with increased relative effect (~ 5 times larger than the most predictive metabolite) for predicting metformin exposure (**Figure 2A**). Since prescribed metformin daily dose was a non-binary feature, characterized by 10 discrete values (ranging from 250 to 3000 mg), adjusting for a potential confounding effect using stratification alone – one of the common strategies to adjust for

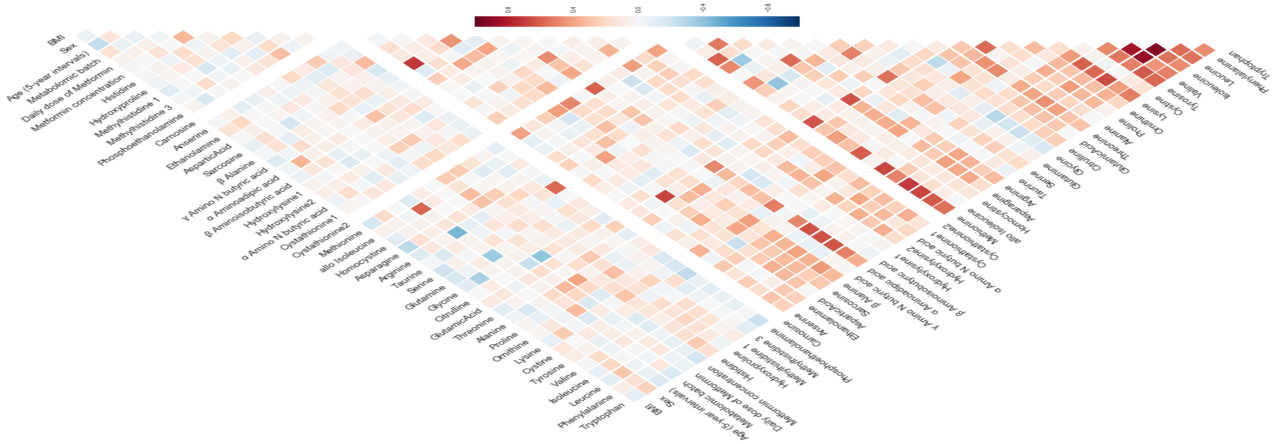


Figure 1. Pearson's correlation coefficients for metabolite and clinical features.

confounding feature in AutoML analysis – had potential to create unbalanced or underpowered subgroups and bias resulting associations. We posited, while increased dose has potential to mask identification of relevant metabolite features, inherently enhancing the biological effect of metformin exposure, it is unlikely to introduce bias by a confounding effect and can be removed from analysis. When dose and concentration of metformin exposure were removed from consideration, a more gradual distribution of rank coefficients were observed. The top three features contained the metabolites homocysteine and citrulline and clinical feature BMI (**Figure 2B**). These findings, together with existing biomedical knowledge, suggest that dose-dependent features have potential to mask important metabolite features and BMI might introduce bias due to confounding.

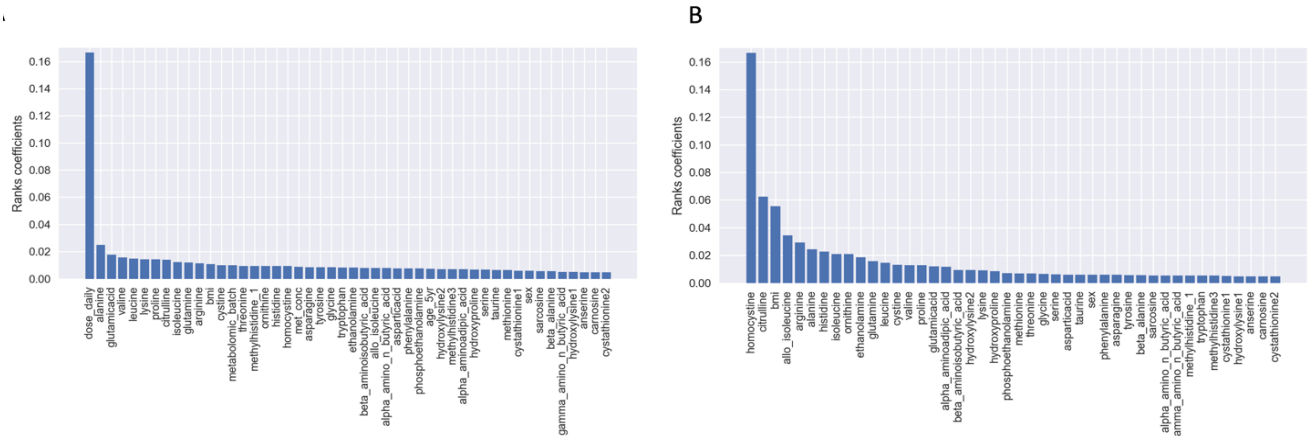


Figure 2. Metabolite and clinical feature ranks. The most predictive features have lowest values; the least predictive features have highest values. A) All metabolic and clinical features. B) All metabolic and clinical features excluding daily dosage.

3.3. *Evaluation of clinical characteristics for potential confounding*

Within the below sections, we assessed potential confounding using out-of-the-box AutoML. Specifically, we evaluatde BMI, metabolomics batch effect, and potential dose-dependent effects:

3.3.1. ***Body mass index (BMI).*** While BMI was demonstrated to be associated with metformin exposure overall, suggesting confounding, potential for within case-control status confounding was unknown. To elucidate potential within-case confounding (i.e. confounding by indication) by BMI, TPOT regression analysis was performed on a case-only dataset with metabolite and BMI features and evaluated using R^2 or accuracy (**AC**) metrics (**R^2= testing; training**). TPOT generated various regression models, including Elastic Net Regressor with built-in Cross-Validation, Extra Tree Regressor, Random Forest Regressor, and Ridge Regressor with built-in Cross-Validation. Due to low accuracy, the highest being (R^2=0.48;<0.01), we were unable to assign rank coefficients or an effect. This low accuracy suggested that either the TPOT models were insensitive to within-case BMI or that no confounding effect existed. To further elucidate a potential effect, BMI was evaluated with various splits and thresholds. However,

the highest accuracy (R^2= 0.89;neg) implicated a model likely over-fit and containing false positives. Univariate linear regression analysis was performed on the same datasets to serve as a benchmark for TPOT regression performance. In independent regression analyses of top metabolite features, model performance remained poor for alanine (R^2=0.58;null) and α-aminoadipic acid (R^2=0.62;0.42). While the models did not identify BMI associations, the existence of distinct distributions of BMI within cases and controls (**Figure 3**) suggest potential confounding and potential insensitivity of TPOT where collinearity exists.

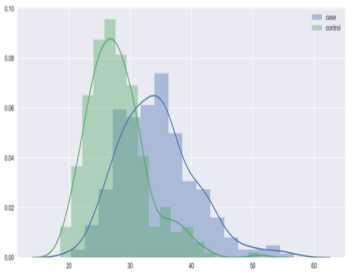


Figure 3. Distribution of BMI within case and control

3.3.2. ***Metabolomics batch effect.*** To elucidate a potential batch effect, TPOT analysis was performed stratified by batches. We ran TPOT classification analysis for each batch subset and compared performance. Overall, subsets performed very well with high accuracy for both training and testing sets, suggesting strong potential for a batch effect. For individual batch performance, we observed the following: batch 1(AC=0.90;0.87); batch 2 (AC=0.90;0.82); batch 3 (AC=0.95;0.94); batch 4 (AC=0.95;0.81); batch 5 (AC=0.92;0.90); batch 6 (AC=0.94;0.93); batch 7 (AC=0.96;0.73); batch 8 (AC=1.0;1.0). However, case-control frequencies varied within these associations, with batch 2 having 61 cases and batch 5 having only 2 cases. Unbalanced randomization between case-control selection in batch assignment likely contributed to a potential batch effect.

3.3.3. ***Metformin dose-dependent effect.*** Case-only TPOT analysis generated strong dose-dependent associations (prescribed metformin daily dose and measured plasma metformin concentrations) across several TPOT-generated models. However, when benchmarked to univariate associations, both training and testing accuracies were very low (AC<0.11), suggesting likely model overfitting. In context, these findings suggest that while dose effects may mask associations in clinical metabolic profiling due to being contained within the exposure, the observed is potentially not a true confounding effect. Further work remains to robustly identify and adjust for dose-dependent effects in AutoML.

3.4. *Obtaining metabolic profile guided by predictive ranks and tandem-rank accuracy*

In development of a clinical metabolic profile of metformin exposure, TPOT models selected Gradient Boosting Classifier and Extra Trees Classifier for classification task with accuracy scores for a training set 0.98 and above and for a testing set 0.83 and above. Distribution of rank coefficients clearly prioritized homocysteine and citrulline as top metabolite features. To ascertain additional metabolite features of potential relevance, recursive feature elimination was applied and features sorted in accordance by their rank. Training and testing accuracy of the model continued to increase (**Figure 4**) up to the addition of the top 4 important metabolite features (homocysteine, citrulline, allo- isoleucine, and arginine) and remained relatively unchanged beyond inclusion of an additional 2 metabolite features (alanine and isoleucine).

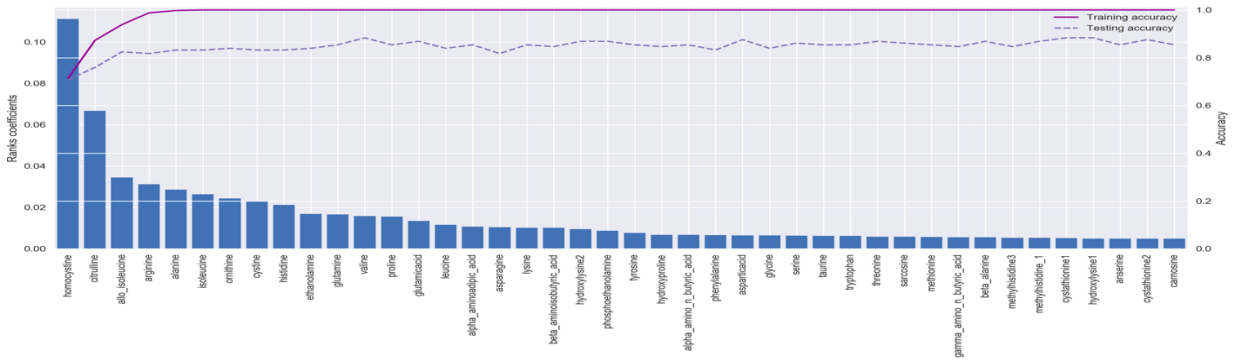


Figure 4. AutoML generated clinical metabolic profile for exposure to metformin guided by tandem-rank accuracy measure. Sorted histogram of predictive power for metabolite inverse (for ease of interpretation) sum of ranks (blue bar), training set accuracy, (solid magenta line), and testing set accuracy (dashed magenta line) describe relative feature effect size and model performance.

In this analysis, the rank metrics provided information about relative importance of metabolic variables with respect to their predictability of the outcome (metformin exposure) variable. The accuracy metrics provided an estimate of the model performance on the various subsets of the features and help to distinguish the most discriminative features in the dataset. Together rank and accuracy generated a statistical support for distinguishing metabolites that show differential response to exposure. In this study, the tandem metrics demonstrated that homocysteine was consistently identified as our top TPOT-recommended feature, with a much larger magnitude of effect than other features.

3.5. *Proposed adjustments for confounding bias in AutoML analyses*

In the previous section, we identified features with potential to bias AutoML-based clinical metabolic profiling endeavors due to confounding effects. We identified that TPOT is potentially insensitive to identification of low-effect confounding features, and that high-effect features may mask potentially relevant prioritized metabolite features. As such, manual adjustment for potential confounding features might be necessary. Further, as is often required for epidemiological inquiry, select feature adjustment might be required to rule out a suspected confounding effect.

To adjust for confounding in AutoML, we suggest two data type-dependent adjustment strategies: 1) For continuous values, we propose that residuals obtained from independent linear regressions [13] (e.g. between metabolites and BMI) be obtained prior to AutoML analyses. This aims to be consistent with approaches appropriately address confounding in multivariate statistical analyses [14]. In our application, the residual index was computed independently for each metabolite as the residual from the simple linear regression of metabolite variable on the confounding variable (i.e. BMI). The residual distances of individual points from regression line then served as the estimators of metabolites. Our adjustment for BMI slightly enhanced our homocysteine association – demonstrated to be increased in plasma concentration with metformin exposure in multivariate linear regression – with the model accuracy remaining comparable. Metabolite features originally ranked below homocysteine in unadjusted analysis consistently remained below homocysteine, but were slightly modified by rank

order and magnitude. We recommend sensitivity analysis and regression diagnostic methods to select a proper regression model for adjustment in AutoML applications. 2) For categorical data types, we recommend stratification, where independent analyses are conducted and findings (e.g. means) presented in aggregate. In this approach, AutoML generates feature importance coefficients for each subset and then transformed into ranked coefficients (described in Section 2.2) where mean of ranks are calculated over all subsets. However, stratification is known to be negatively impacted by low sample power. Frequently, well-powered strata produce more accurate estimates than relatively lower powered strata. To supplement this deficiency, we suggest weighted mean, particularly Mantel-Haenszel [15], where strata are prioritized be statistical power. When applied, stratum-specific adjusted relative risk estimates can be calculated, providing an overall summary measure of effect. Approach described for categorical data types might be most aptly applied in future research to elucidate potential metformin dose-dependent effects.

4. Discussion

In this study, we demonstrated AutoML considerations using TPOT for metabolic profiling of exposure to metformin monotherapy in a biobank cohort. Our two major informatics contributions include: 1) tandem rank-accuracy measures to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors, and 2) residual training and adjustment of metabolite features in AutoML analysis. Both our informatics contributions and identified metabolite associations contribute to precision medicine knowledge.

4.1. *Considerations and adjustments for confounding features*

In our analysis, we demonstrated that while AutoML is a potential powerful tool for clinical metabolic profiling, specific considerations and adjustments might need to be applied for potentially confounding characteristics. Correlation and agnostic TPOT analyses demonstrated that daily dose and BMI had strong-to-medium associations with metformin exposure. While not a focus of our analysis, this dataset is uniquely suited for future evaluation of dose-dependent effects using AutoML and other analysis approaches. For datasets with potential confounding, we proposed two data type-dependent adjustment strategies: 1) stratification for categorical features, and 2) independent residual identification and application for continuous features.

4.2. *Future study design considerations*

AutoML methods (and ML methods in general) can be sensitive to the dataset quality in terms of sample size and sample structure. A distinct strength of our study was the well-powered ($n = 546$) dataset with targeted metabolites. Conversely, datasets with a small sample size ($n < 50$ samples) can often lead to overfitting, especially when the dataset has high variance due to random noise. Here we had 546 samples and 42 metabolites, which is considered a good ratio of features to samples to avoid high

variance problem. However, in untargeted clinical metabolic profiling studies, where the number of metabolites exceeds the number of samples 10 or even 100-fold, this high variance is a common problem. In this scenario, even high accuracy scores could be unreliable without deploying an alternative strategy. One suggested approach to avoid this pitfall is to apply feature selection methods before running AutoML analysis. Relief-based algorithms, recursive feature elimination and regularization techniques are among the most common approaches used to treat overfitting and is directly applicable in future untargeted clinical metabolic profiling endeavors.

Imbalanced datasets with one phenotype overrepresented (i.e. more controls than cases) can also cause bias in AutoML and ML classification tasks. Several adjustments could be made including: 1) changing performance metrics to one that can give more insight into the accuracy of the model than traditional classification accuracy (e.g., area under the ROC curve, precision and recall); 2) resampling data using either addition of copies of samples from the under-represented phenotype or removal of samples from the over-represented phenotype. Imbalanced datasets can also make it difficult to apply stratification approaches to control for confounding. A potential weakness of our study is that batch effect could not be reasonably incorporated into adjustment approaches due to imbalance. Ensuring balanced observation batches is a critical consideration for study design in future clinical metabolic profiling studies.

Finally, datasets with a large percentage of missing values can be problematic for AutoML and ML methods. Another strength of our study includes that all features were fully populated. Many ML methods cannot handle missing values by default, so a common approach is to replace all missing values in a column with the median or mean value of that column (for all columns with missing values), or even replace all missing values with a fixed value (e.g., -99 or 0, depending on the feature). However, replacing missing values in such a manner, especially when there is a large percentage of missing values can introduce noise into the dataset and bias analysis. Thus, we recommend taking a thorough approach when replacing missing values in a dataset for AutoML and ML.

4.3. *Increased homocysteine*

Beyond our specific informatics contributions, we demonstrated the utility of AutoML to enhance multi-omic perspectives in pursuit of precision medicine knowledge. In this study, we present increased homocysteine with long-term exposure to metformin as a potentially novel metabolite association suggested by TPOT; an association not identified in parallel clinical metabolic profiling endeavors. While warranting independent replication, our tandem rank-accuracy measure suggests homocysteine to be the metabolite feature with largest effect, and corresponding priority for further translational clinical research. Residual training and adjustment for a potential confounding effect by BMI only slightly modified our initial association. Elevated homocysteine levels are clinically associated with vitamin B12 and folate deficiency [16] – we suggest future consideration for potential clinical relevance and independent replication. Elevated homocysteine is also associated with some increased posited risk

for atherosclerotic disease [17], potentially cancer [18], and depression [19], but has insufficient evidence to suggest consideration as a clinical predictor or biomarker.

While indeed, the maturation of metabolomics science and proliferation of biobanks are exciting, combining with the expansive clinical perspectives offered by EHR linkages offer unprecedented opportunity. We posit that EHR perspectives of phenotypic divergence combined with metabolic variation are poised to become powerful facets advancing clinical translational science. Judicious application of ML and AutoML approaches will become increasingly powerful in multi-omic research.

5. Conclusion

AutoML is an exciting tool holding great promise to enhance clinical metabolic profiling and advance translational research endeavors; considerations are recommended, including adjustment approaches for clinical confounders. Our identified association of increased homocysteine with long-term metformin exposure warrants independent replication and evaluation for potential clinical relevance.

Acknowledgments. This research was made possible with generous support from the Mayo Clinic, Center for Individualized Medicine. The Institute for Biomedical Informatics, University of Pennsylvania and Mayo Clinic Cancer Genetic Epidemiology Training Program (R25 CA092049) further supported this research.

References

1. F. S. Collins and H. Varmus, *N. Engl. J. Med.* 26, 372 (2015).
2. J. E. Olson, E. Ryu, K. J. Johnson, et. al., *Mayo. Clin. Proc.* 88, 9 (2013).
3. C.H. Johnson, J. Ivanisevic, and G. Siuzdak, *Nat. Rev. Mol. Cell Biol.* 17, 451–459 (2016).
4. T. Kind, V. Tolstikov, O. Fiehn, and R.H. Weiss, *Anal. Biochem*. 363, 185-195 (2007).
5. R. Kaddurah-Daouk and R. M. Weinshilboum, *Clin. Pharmacol. Ther.* 95, 2 (2014).
6. D. M. Rotroff, N. O. Oki, X. Liang, et. al., *Front. Pharmacol.* 7 (2016).
7. M. S. Monteiro, M. Carvalho, M. L. Bastos, P. Guedes de Pinho, *Curr Med Chem.* 20, 2 (2013).
8. W. Chai, S. M. Conroy, G. Maskarinec, et. al., *Nutr. Res.* 30, 4 (2010).
9. R. S. Olson, R. J. Urbanowicz, P.C. Andrews, et. al., *Appl. Evol. Comput.* 123-137 (2016).
10. R. S. Olson, N. Bartley, R. J. Urbanowicz and J. H. Moore, *Proc. Genet. Evol. Comput. Conf.* 485-92 (2016).
11. A. Sohn, R. S. Olson, and J. H. Moore, *Proc. Genet. Evol. Comput. Conf.* 489-496 (2017).
12. M. K. Breitenstein, R. Berger, S. Bos, et. al., *Clin. Pharmacol. Ther.* 99, S30 (2016).
13. R. McNamee, *Occup. Environ. Med.* 62, 7 (2005).
14. E. Suzuki, T. Mitsuhashi, T. Tsuda, E. S. Yamamoto., *Am. J. Epidemiol.* 27, 2 (2017).
15. A. Mannocci, *Ital J. Public Health*. 15, 6 (2012).
16. J. Selhub, M. S. Morris, and P. F. Jacques, *Proc. Natl. Acad. Sci. U.S.A.* 104, 50 (2007).
17. A. Schaffer, M. Verdoia, E. Cassetti, et. al., *Thromb. Res.* 134, 2 (2014).
18. Cancer D. Padovani, A. Hessani, F. T. Castillo, et. al., *Nat. Commun.* 7, 13386 (2016).
19. H. Tiemeier, H. R. Van Tuijl, A. Hofman, et. al., *Am. J. Psychiatry* 159, 12 (2002).

Addressing vital sign alarm fatigue using personalized alarm thresholds*

Sarah Poole[1], and Nigam Shah MBBS PhD[1]

[1]*Center for Biomedical Informatics Research, Stanford University*
Stanford, CA, United States

Alarm fatigue, a condition in which clinical staff become desensitized to alarms due to the high frequency of unnecessary alarms, is a major patient safety concern. Alarm fatigue is particularly prevalent in the pediatric setting, due to the high level of variation in vital signs with patient age. Existing studies have shown that the current default pediatric vital sign alarm thresholds are inappropriate, and lead to a larger than necessary alarm load. This study leverages a large database containing over 190 patient-years of heart rate data to accurately identify the 1st and 99th percentiles of an individual's heart rate on their first day of vital sign monitoring. These percentiles are then used as personalized vital sign thresholds, which are evaluated by comparing to non-default alarm thresholds used in practice, and by using the presence of major clinical events to infer alarm labels. Using the proposed personalized thresholds would decrease low and high heart rate alarms by up to 50% and 44% respectively, while maintaining sensitivity of 62% and increasing specificity to 49%. The proposed personalized vital sign alarm thresholds will reduce alarm fatigue, thus contributing to improved patient outcomes, shorter hospital stays, and reduced hospital costs.

Keywords: Alarm Fatigue, Vital Signs, Personalized Medicine, Random Forest

1. Introduction

Vital sign monitors are an important component of inpatient care, as they provide timely alerts to clinical staff in response to extreme vital sign values[1,2]. These vital sign alarms are intended to be a safety net in the provision of patient care, but their management in the inpatient setting is a significant patient safety issue[3,4]. Efforts to characterize vital sign alarms have shown that 64-99% of the alarms that sound are not clinically actionable[5]. The high proportion of unnecessary alarms has led to provider desensitization, also known as alarm fatigue[6,7]. This has been shown to increase nurse response time to subsequent alarms in both the short and the long term[8,9], increasing the risk to patients and contributing to adverse patient events and, in some cases, patient mortality[7,10]. In 2013, the Joint Commissions issued Sentinel Event Alert #50 to draw attention to widespread alarm fatigue in hospital settings[10], and the subsequent 2014, 2015 and 2016 National Patient Safety Goals urged hospitals to prioritize alarm system safety and ensure that alarms on medical equipment are heard and responded to on time[11-13].

* This work was supported by the Stanford Clinical and Translational Science Award (CTSA) to Spectrum (UL1 TR001085). The CTSA program is led by the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Multiple approaches have been taken to the problem of alarm fatigue, including implementing standards for checking and changing electrocardiography lead wires and electrodes[14-17], escalating alarms to pages sent directly to clinical staff[18], adding delays to alerts to avoid alarming for very short periods of extreme values[14,18,19], implementing standard time series filters[20], and combining alarms to obtain a more general measure of patient deterioration[21]. However, more work needs to be done to address the risk to patient safety from alarm fatigue[6].

Alarm fatigue is particularly prevalent in the pediatric setting, due to the high level of variation in vital signs with patient age[5,6]. Two recent studies have compared default age-based pediatric vital sign alarm thresholds from hospitals with the observed vital signs in each age group. Both studies found that default heart rate alarm thresholds fall near the 50th percentile of patient data, leading to an unnecessarily large alarm load[22,23]. These studies conclude that patient data can successfully be used to choose more appropriate thresholds for vital sign alarms, and initial efforts in this direction have been promising[23]. While the thresholds produced by these studies partially account for the expected change in vital signs with age, the performance of such default thresholds is limited, since vital signs are known to change smoothly and continuously with age[24], rather than displaying the 'step' changes that result from using discrete age groups.

Existing work addressing alarm fatigue includes only limited evaluation of the safety and efficacy of the proposed measures. This is due to the lack of large sets of gold-standard labels that indicate when alarms are crucial for patient safety and optimal outcomes, and when alarms are unnecessary and should be suppressed[5,22,25]. As a result, evaluation has typically focused on maximizing the number of alarms suppressed, with no consideration given to the appropriateness of this suppression.

This study aims to produce improved default vital sign alarm thresholds by extending the previous work in two important ways. Firstly, models are trained to find optimal default vital sign alarm thresholds, given data available at admission, on a patient-by-patient basis, rather than using patient groups defined by discrete age categories as is currently standard. Secondly, evaluation of the resulting patient-specific alarm thresholds found is conducted, by using non-default alarm thresholds as silver-standard personalized thresholds, and by using the presence of clinical events to indicate clinical concern. Heart rate alarms are used as a proof of concept in this manuscript, as heart rate threshold alarms are very common and have been shown to have a low specificity[19,26,27].

An important distinction of this study is the use of heart rate data for training the model, rather than using a set of labeled alarms. Although a large set of alarms are available, they are lacking gold-standard labels to indicate which were unnecessary and which were crucial for patient safety and optimal outcomes. As a result, using historical alarm data to learn optimal thresholds for each patient is not possible. Instead, we take a step back and aim to develop an alarm system from first principles. Since the goal of these vital sign alarms is to indicate when concerning vital sign measurements are seen, we aim to learn the 1st and 99th percentiles of HR seen during the first day of each patient's stay. The use of the 1st and 99th percentiles for these thresholds was chosen carefully. Other studies have

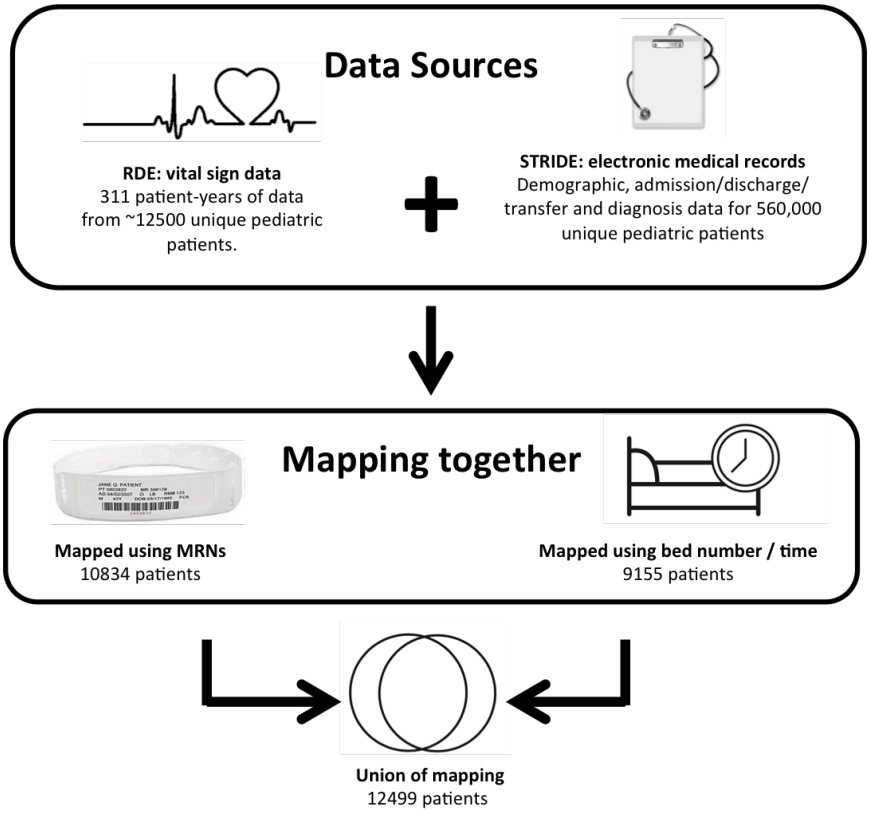


Figure1: Merging of RDE and STRIDE data using patient Medical Record Numbers (MRNs) where available, and using bed number and time where a unique patient was recorded as occupying the bed at the time of interest. Instances where the two mapping schemes gave different patients were removed.

used 5th and 95th percentiles for alarm thresholds[28] where default thresholds are chosen for large groups of patients. Due to inter-patient variability in vital signs, the conservative 5th and 95th percentiles are chosen to ensure that very extreme values in patients who have abnormally high or low vital signs are not missed. Since this study produces thresholds at a patient-specific level, this inter-patient variability does not need to be considered, and wider percentiles can be used to improve alarm specificity. Non-default alarm thresholds are used to evaluate the choice of 1st and 99th percentiles for use as personalized alarm thresholds.

2. Methods

2.1 *Data*

Two main sources of data were used for this study. The Philips Research Data Export (RDE) system at Stanford's Lucille Packard Children's Hospital (LPCH) has been recording vital sign waveforms for every patient that has had their vital signs monitored, both in intensive care units and on floor units, for the past several years. An extract from this system, containing 3.5 years worth of data (5 December 2012 - 20 April 2016) has been made available for research purposes. This extract contains once per minute average heart rate and respiratory rate, as well as records of any vital sign alarms that

were triggered. These data have been combined with data from the electronic health record, obtained through the Stanford Translational Research Integrated Database Environment (STRIDE)[29]. STRIDE contains patient demographics and clinical data including ICD9 codes and medication records. These datasets were linked using patient medical record numbers, or using data showing which patient was in a specific bed location at the time data is available. Figure 1 shows this initial mapping process, and Table 1 shows the characteristics of the final cohort.

For each patient represented in the RDE dataset, the first 24 hours of their stay was isolated and processed for use in this model. All data within this 24-hour period was considered, regardless of whether it was continuously recorded or included periods of missing vital sign data. Patients with less than one hour of data in the first 24-hours of monitoring were excluded from the analysis. Of the remaining patients, 83% of patients had data spanning at least one day, and the remaining patients had data over a mean of 15.3 hours. Four values were extracted for each patient: the mean, standard deviation, 1st percentile, and 99th percentile of the heart rate data available in this 24-hour period.

Table 1: Characteristics of patient cohort

		Count	Percentage
Total number of patients:		8,507	
Total HR alarms triggered:		1,930,493	
	Low	693,516	35.69%
	High	1,236,977	64.1%
Mean HR observations per patient:		14385 minutes (9.98 days)	
Standard deviation of HR observations per patient:		29942 minutes (20.8 days)	
Demographic breakdown:			
Gender:			
	Male	3,921	46.1%
	Female	4,586	53.9%
Ethnicity:			
	Hispanic	2,121	24.9%
	Not Hispanic	3,339	39.3%
	Unknown	3,047	35.8%
Race:			
	Asian	767	9.0%
	Black	150	1.8%
	Native American	6	0.1%
	Other	1,937	22.8%
	Pacific Islander	193	2.3%
	Unknown	3,087	36.3%
	White	2,367	27.8%
Age (years):			
	Min	0.00	
	Median	1.85	
	Mean	4.96	
	Max	17.98	

2.2 *Outcome*

There are two outcomes of interest for each patient, corresponding to the high and low alarm thresholds. The proposed ideal values for these are the 1st and 99th percentiles of the patient's observed heart rate over the first day of hospitalization. To allow for future extensions of this work, each patient's heart rate over the first day of hospitalization is modeled as a lognormal distribution parameterized by the mean and standard deviation of the heart rate, and the 1st and 99th percentiles are obtained from this lognormal model. Figure 3 shows that the 1st and 99th percentiles of the patient's heart rate are able to be accurately recovered using the mean and standard deviation of heart rate in a lognormal distribution. Two models are built, with outcomes of mean heart rate and standard deviation of heart rate. The outputs of these models are then used as the parameters of patient-specific lognormal distributions, from which the expected 1st and 99th percentiles of the patient's heart rate are found. Evaluations are performed on these resulting 1st and 99th percentiles, as these are the proposed alarm thresholds.

2.3 *Imputing weight*

Including patient weight in the model was a prime consideration, as weight is known to impact heart rate. Weight data was not available for all patients, and for some patients weight at the time of the vital sign recording was not available. An imputation process was developed for weight data, using standard pediatric growth charts. Growth charts were used to find which percentile the patient fell into for their age at the time that weight was recorded. The growth charts were then used to determine the weight that the patient would be at the time of vital sign recording, assuming that they remained in the same percentile. If patients had multiple weights recorded, the mean percentile was used. 603 patients had no weight data recorded, so were assumed to be at the 50th percentile of weight for their age. The percentile found for each patient was also used as an input to the model.

2.4 *Diagnosis Information*

The STRIDE dataset includes diagnosis related groups (DRGs)[30], which are designed to group patients according to the medical services they receive, but can also be used to provide a rough grouping by clinical complaint. 45 DRGs were present as admit diagnoses in the cohort of interest. DRGs that contained less than 10 observations were combined into an 'OTHER' group, leaving a total of 22 distinct DRGs. This categorical variable was converted to 22 variables with Boolean values, with the constraint that for a given sample only one of the values can be set to 1.

The floor departments at LPCH are arranged such that patients with particular care needs are grouped together. For example, one floor unit typically houses patients with cardiac issues, while patients with pulmonary-related problems are cared for in another unit. The department in which the patient is located was used as a feature in our model, as it provides a rough grouping according to diagnosis.

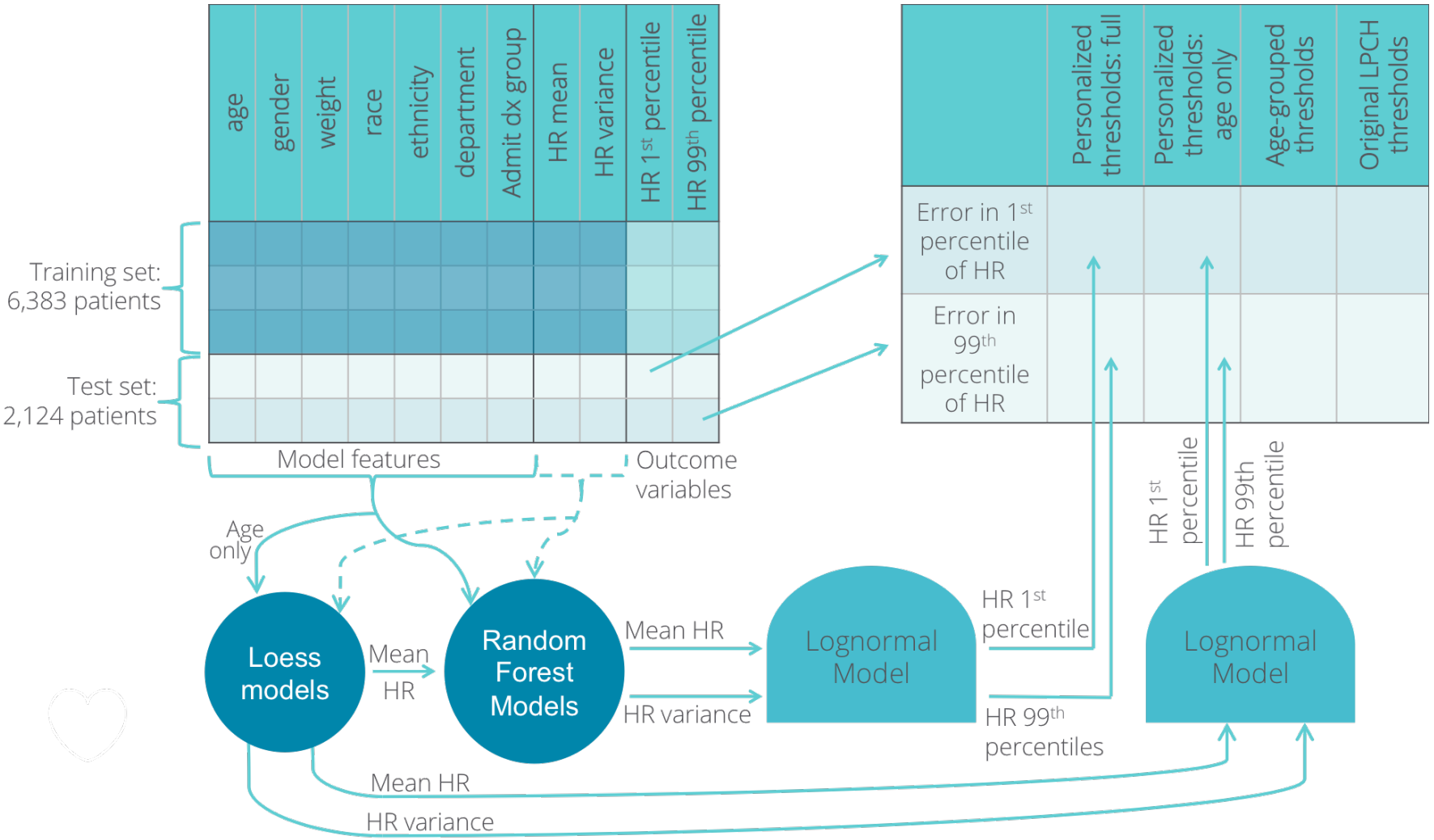


Figure 2: Schematic of the model training process. First the training set is used to fit loess models to the outcomes of mean heart rate and heart rate variance, using age as the only feature. The outputs of these models are used as the parameters of a lognormal model to estimate the 1st and 99th percentiles of heart rate. These resulting estimates are proposed as personalized thresholds: age only. The output of the loess model fitting to mean heart rate is also used as a feature in a pair of random forest models, one fit to the outcome of mean heart rate, and the other fit to variance of heart rate. These random forest models models also have gender, weight, race, ethnicity, hospital department, and admit diagnosis group (DRG) as additional features. The mean and variance of heart rate for each patient, as predicted by the random forest models, are used as the parameters of a lognormal model, allowing the 1st and 99th percentiles of heart rate to be estimated. The trained models are used to estimate the 1st and 99th percentiles of heart rate for patients in the test set, which can then be compared to the actual values observed over the first day of monitoring. The previously used original LPCH thresholds and age-grouped thresholds can also be compared to the observed 1st and 99th percentiles of heart rate.

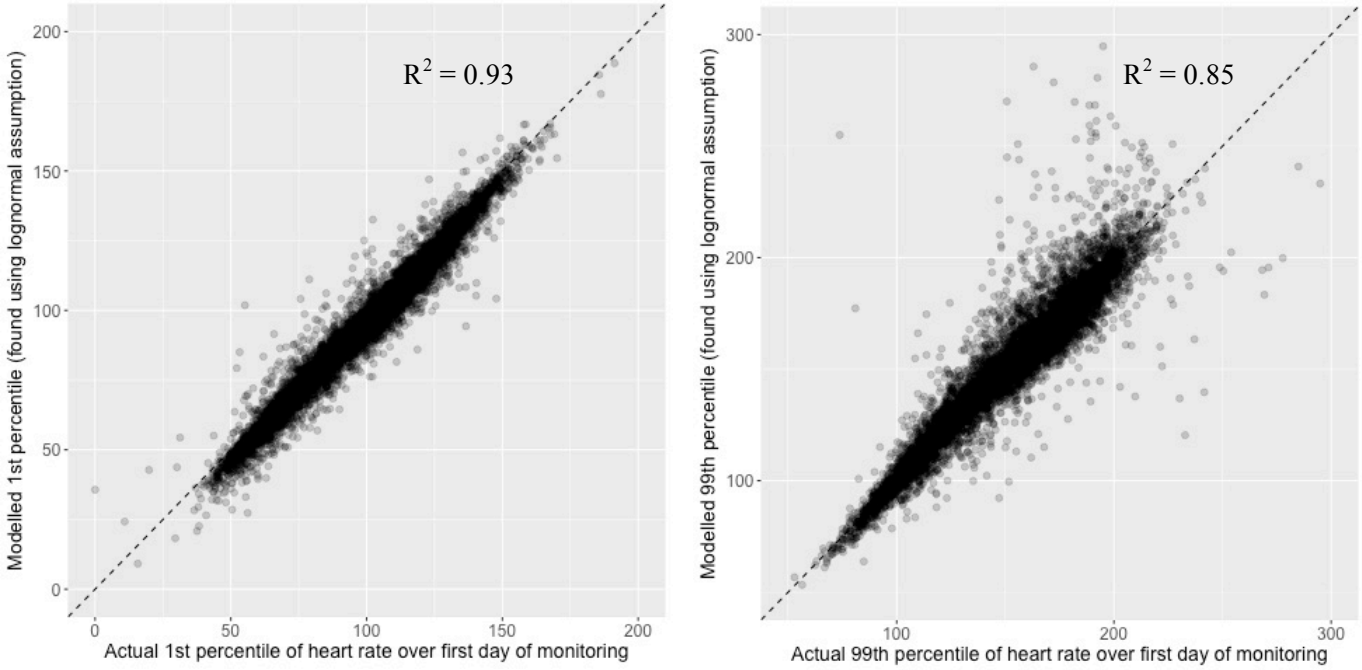


Figure 3: Error using mean and standard deviation of heart rate with lognormal assumption to find 1st (left) and 99th (right) percentile of heart rate.

2.5 *Training*

The combined RDE and STRIDE data set was randomly split into training and testing cohorts using a 75%/25% split at the patient level, resulting in cohorts of 6,383 and 2,124 patients.

As previously described, two models were trained: one to identify the mean heart rate, and the second to identify the standard deviation of the heart rate. The output of these models is used in a lognormal distribution to calculate the 1st and 99th percentiles of heart rate, which are proposed for use as the alarm thresholds. Two sets of these two models were trained. Figure 2 describes the training process. First, loess models[31] were used to capture nonlinear variation in the mean and variance of heart rate with age. The thresholds calculated from the output of these models are referred to as 'personalized thresholds: age only' thresholds. The output from these models was used as inputs to two random forest models (one each for mean heart rate and standard deviation), along with additional demographic (age, weight, gender, ethnicity and race) and diagnostic features (DRG and hospital department). A random forest model was chosen to avoid bias that would be introduced by a linear model.

2.6 *Evaluations*

The results can be evaluated directly by comparing the modeled 1st and 99th percentiles of the vital signs to the actual values. We include comparisons with the original LPCH vital sign thresholds, and age-based thresholds previously described in [32,33]. A record of the alarms that sounded is available, so the number of alarms that would have been suppressed if the predicted 1st and 99th percentiles were

used as thresholds can be found. However, evaluation of the clinical meaningfulness of these results is difficult, as gold-standard labels indicating whether alarms were meaningful are not available.

To estimate the appropriateness of the proposed alarm thresholds, we use the dataset of alarms that sounded in LPCH to find alarm thresholds that are not the default values, indicating that clinical staff manually chose this threshold for the patient. The non-default alarms from patients in both the training and the test set were able to be compared to the proposed thresholds without biasing the result of the evaluation, since the actually used thresholds that were used in practice were not input to the models. A total of 727 and 2,242 alarms with non-default settings were found for patients in the test set and the training set respectively.

As a second estimate of the appropriateness of the proposed alarm thresholds, we looked for significant clinical events in the 4 hours following an alarm. The label of 'clinically meaningful alarms' implies that to meet this criterion, some clinical action should have been taken in response to the alarm. Two lists of clinical events were formulated through consultation with clinicians and clinical experts, and are shown in Table 2. The presence of a clinical event from list A is considered to imply that the alarm was clinically meaningful, while an event from list B implies that the alarm was unnecessary. If events from both list A and list B occur in the 4-hour period following the alarm, this is considered ambiguous and no label is assigned to the alarm. A discharge event where the patient dies within 30 days is treated as a discharge to end of life care. 6.9% of all alarms recorded were assigned a label using this process (8.3% of low alarms and 6.1% of high alarms).

Table 2: Clinical events used to indicate whether an alarm was clinically meaningful (list A) or unnecessary (list B).

List A (indicates clinically meaningful alarm)	**List B (indicates unnecessary alarm)**
Patient death	Patient discharged
Patient transferred to higher acuity unit	Patient transferred to lower acuity unit
Manual change of alarm thresholds to become more conservative	Manual change of alarm thresholds to become less conservative
Patient discharged to end of life care	

A record of the alarms that sounded is available, so the number of alarms that would have been suppressed if the predicted 1st and 99th percentiles were used as thresholds can be found. The status of the alarms if the new thresholds were used is compared to the labels created using the clinical events in Table 2 to obtain estimates of the sensitivity and specificity of the alarms. These values are not available to evaluate any other vital sign alarming method, so comparisons with previous methods are not possible. We are also unable to evaluate the performance of the original or age-based LPCH thresholds, as these were used to trigger the alarms.

3. Results

As shown in Figure 3, using the mean and standard deviation of a patient's heart rate over a 24-hour period as parameters in a lognormal distribution gives an accurate estimate of the 1st and 99th percentiles of the heart rate over this period. This shows that a lognormal model is well suited to the distribution of an individual patient's heart rate over a 24-hour period.

Figure 4 shows that a model with a single variable of continuous age is able to recover the 1st and 99th percentiles of heart rate more closely than the age-based thresholds previously developed at LPCH. Adding additional demographic and diagnostic features slightly decreases the variance in the error. Figure 4 also compares the vital sign thresholds to the thresholds of non-default alarms in the data set. The low error suggests that the use of 1st and 99th percentiles as threshold values is an appropriate one. The continuous age only model has a similar error to the age-based thresholds, while adding demographic and diagnostic features decreases this error.

Table 3 shows that over 50% of low heart rate alarms would be suppressed using the proposed thresholds, as well as upwards of 35% of high heart rate alarms, depending on the threshold scheme.

Using clinical events to infer labels for the alarms allowed us to estimate sensitivity and specificity of the proposed thresholds, shown in Table 4. Previous studies have shown that the specificity of heart rate alarms ranges from 1% to 36%[5], suggesting that our proposed alarm thresholds would improve the specificity of heart rate alarms. Since it is not possible to measure false negative alarms, no studies have been conducted to determine the sensitivity of existing vital sign alarms, however this is generally considered to be extremely high, close to 100%.

Table 3: Percentage of alarms suppressed using proposed thresholds

	% low alarms suppressed	% high alarms suppressed
Personalized thresholds: age-only	53.1%	35.2%
Personalized thresholds: full	50.5%	44.1%

Table 4: Performance metrics of proposed thresholds calculated using presence of clinical events to label alarms.

	Sensitivity	Specificity	Positive Predictive Value
Personalized thresholds: age-only	0.67	0.44	0.072
Personalized thresholds: full	0.62	0.49	0.079

4. Discussion

This study has shown that the 1st and 99th percentiles of observed heart rate over the first day of an inpatient stay are able to be predicted using a random forest with demographic and diagnostic features.

The comparison of the predicted 1st and 99th percentiles to the non-default alarm settings that were used in practice gives insight into the appropriateness of using the output from these models as alarm thresholds. Despite not being trained with the non-default alarm settings as inputs, the models recover these values well.

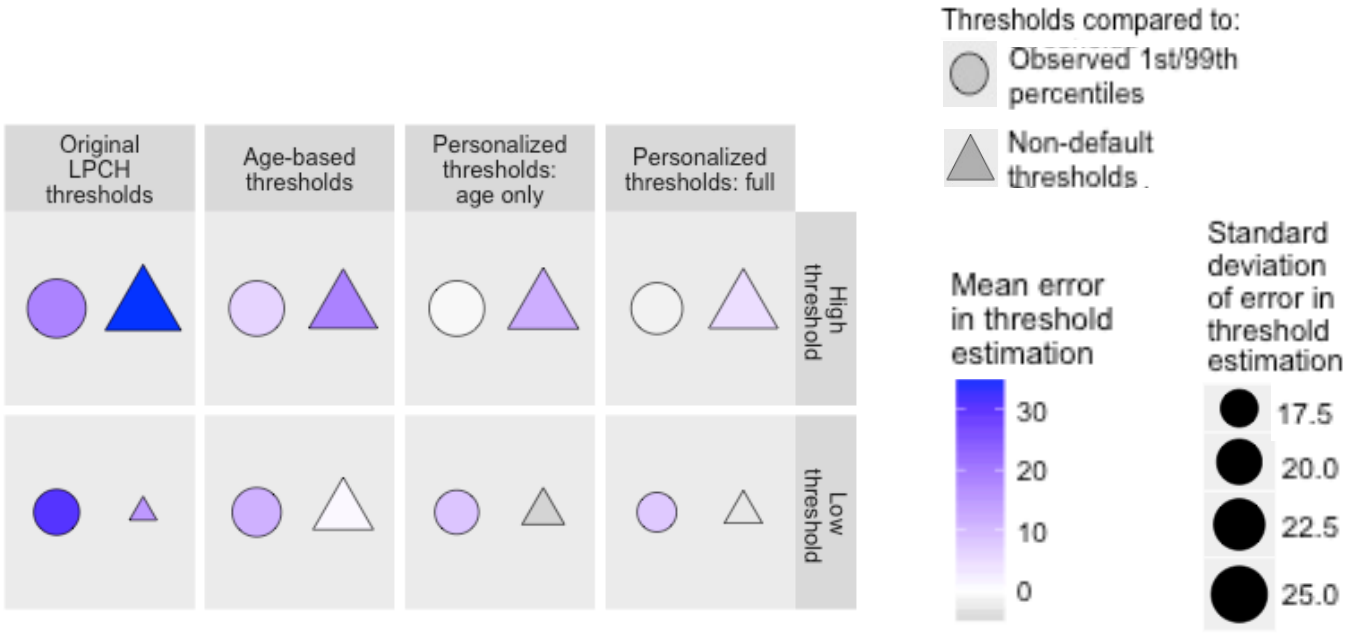


Figure 4: Comparison of alarm thresholds with the 1st (for low thresholds) and 99th (for high thresholds) percentiles of heart rate observed over the first 24 hours of monitoring (circles), and comparison of alarm thresholds with the recorded non-default thresholds (triangles).

As shown in Table 4, the specificity of the proposed thresholds is higher than that of current heart rate alarms, but the sensitivity of the proposed thresholds is likely to be lower than the sensitivity of current alarms. In theory, this increases the chance that truly concerning heart rates will fail to sound an alarm, which could lead to negative outcomes for the patient. However, an alarm sounding is of no help to the patient in distress if it is not responded to, as may happen in a situation where clinical staff are suffering from alarm fatigue[34]. While studies have shown that various forms of alarm fatigue can increase nurse response time[8,9], no studies have quantified the effective sensitivity of alarms given the presences of alarm fatigue. We propose that the reduced number of alarms that will sound if these personalized thresholds are adopted (see Table 3) will reduce the problem of alarm fatigue, and that this reduction in the desensitization of health care providers will reduce the instance of negative patient outcomes related to missed vital sign events, despite the lower expected alarm sensitivity.

Limitations of this study include the lack of gold standard alarm labels to evaluate our proposed alarm thresholds. The evaluation methods used in lieu of gold standard labels (comparing to patient 1st and 99th percentiles, comparing to non-default alarm limits, and using clinical events to infer alarm labels) improve upon previous studies that have lacked any evaluation, but are still limited. For example only 7% of alarms could be labeled using clinical events. This also limits the accuracy of the performance metrics reported for the proposed alarm thresholds.

In conclusion, this study presents a model to accurately identify the 1st and 99th percentiles of an individual's heart rate during their first day of vital sign monitoring, using demographic and diagnosis features as input to a random forest. This is a proof of concept that personalized alarm thresholds can be learned, and demonstrates promising results for use of such personalized thresholds to reduce false alarms and address alarm fatigue. Patient-specific alarm thresholds represent a first step towards personalized medicine, and the resulting reduction in alarm fatigue will improve patient outcomes while also contributing to lower healthcare costs.

Acknowledgements

The authors would like to thank Eric Helfenbein, Principal Scientist at Philips Healthcare, for his help in obtaining the RDE dataset. We also thank Stephen Pfohl and Dr. Tavpritesh Sethi, both members of Stanford's Biomedical Informatics Research group, for their helpful feedback on the manuscript.

References

1. Force C. Impact of clinical alarms on patient safety: a report from the American College of Clinical Engineering Healthcare Technology Foundation. *J Clin Eng* 2007.
2. Chopra V, McMahon LF, Jr. Redesigning hospital alarms for patient safety: alarmed and potentially dangerous. *JAMA* 2014; **311**(12): 1199-200.
3. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. *AACN Adv Crit Care* 2013; **24**(4): 378-86; quiz 87-8.
4. Bridi AC, Louro TQ, da Silva RC. Clinical Alarms in intensive care: implications of alarm fatigue for the safety of patients. *Rev Lat Am Enfermagem* 2014; **22**(6): 1034-40.
5. Paine CW, Goel VV, Ely E, et al. Systematic Review of Physiologic Monitor Alarm Characteristics and Pragmatic Interventions to Reduce Alarm Frequency. *J Hosp Med* 2016; **11**(2): 136-44.
6. Lawless ST. Crying wolf: false alarms in a pediatric intensive care unit. *Crit Care Med* 1994; **22**(6): 981-5.
7. Varpio L, Kuziemsky C, MacDonald C, King WJ. The helpful or hindering effects of in-hospital patient monitor alarms on nurses: a qualitative analysis. *Comput Inform Nurs* 2012; **30**(4): 210-7.
8. Bonafide CP, Lin R, Zander M, et al. Association between exposure to nonactionable physiologic monitor alarms and response time in a children's hospital. *J Hosp Med* 2015; **10**(6): 345-51.
9. Bonafide CP, Zander M, Graham CS, et al. Video methods for evaluating physiologic monitor alarms and alarm responses. *Biomed Instrum Technol* 2014; **48**(3): 220-30.
10. Sentinel Event Alert Issue 50: Medical device alarm safety in hospitals.
11. Commission TJ. The Joint Commission Announces 2014 National Patient Safey Goal. 2014.
12. Commission TJ. 2015 Hospital National Patient Safety Goals. 2015.
13. Commission TJ. 2016 Hospital National Patient Safety Goals. 2016.
14. Dandoy CE, Davies SM, Flesch L, et al. A team-based approach to reducing cardiac monitor alarms. *Pediatrics* 2014; **134**(6): e1686-94.
15. Albert NM, Murray T, Bena JF, et al. Differences in alarm events between disposable and reusable electrocardiography lead wires. *Am J Crit Care* 2015; **24**(1): 67-73; quiz 4.
16. Cvach MM, Biggs M, Rothwell KJ, Charles-Hudson C. Daily electrode change and effect on cardiac monitor alarms: an evidence-based practice approach. *J Nurs Care Qual* 2013; **28**(3): 265-71.
17. Sendelbach S, Wahl S, Anthony A, Shotts P. Stop the Noise: A Quality Improvement Project to Decrease Electrocardiographic Nuisance Alarms. *Crit Care Nurse* 2015; **35**(4): 15-22; quiz 1p following

18. Cvach MM, Frank RJ, Doyle P, Stevens ZK. Use of pagers with an alarm escalation system to reduce cardiac monitor alarm signals. *J Nurs Care Qual* 2014; **29**(1): 9-18.
19. Gorges M, Markewitz BA, Westenskow DR. Improving alarm performance in the medical intensive care unit using delays and clinical context. *Anesth Analg* 2009; **108**(5): 1546-52.
20. Tsien C. Reducing False Alarms in the Intensive Care Unit: A Systematic Comparison of Four Algorithms. 1997.
21. Hu X, Sapo M, Nenov V, et al. Predictive combinations of monitor alarms preceding in-hospital code blue events. *J Biomed Inform* 2012; **45**(5): 913-21.
22. Goel VV, Poole SF, Longhurst CA, et al. Safety analysis of proposed data-driven physiologic alarm parameters for hospitalized children. *J Hosp Med* 2016; **11**(12): 817-23.
23. Jacques SJ, Fauss EK, Sanders JA, et al. Patient-centered design of alarm limits in a complex pediatric population. *Health and Technology* 2016.
24. Fleming S, Thompson M, Stevens R, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *Lancet* 2011; **377**(9770): 1011-8.
25. de Waele S, Nielsen L, Frassica J. Estimation of the patient monitor alarm rate for a quantitative analysis of new alarm settings. *Conf Proc IEEE Eng Med Biol Soc* 2014; **2014**: 5727-30.
26. Tsien CL, Fackler JC. Poor prognosis for existing monitors in the intensive care unit. *Crit Care Med* 1997; **25**(4): 614-9.
27. Schoenberg R, Sands DZ, Safran C. Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. *Proc AMIA Symp* 1999: 379-83.
28. Karnik A, Bonafide CP. A framework for reducing alarm fatigue on pediatric inpatient units. *Hosp Pediatr* 2015; **5**(3): 160-3.
29. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009; **2009**: 391-5.
30. Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. *Med Care* 1980; **18**(2 Suppl): iii, 1-53.
31. Cleveland WS. LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. *The American Statistician* 1981; **35**(1): 54.
32. Goel V, Poole S, Kipps A, et al. Implementation of Data Drive Heart Rate and Respiratory Rate parameters on a Pediatric Acute Care Unit. *Stud Health Technol Inform* 2015; **216**: 918.
33. Kipps AK, Poole SF, Slaney C, et al. Inpatient-Derived Vital Sign Parameters Implementation: An Initiative to Decrease Alarm Burden. *Pediatrics* 2017.
34. Bridi AC, da Silva RC, de Farias CC, Franco AS, dos Santos Vde L. [Reaction time of a health care team to monitoring alarms in the intensive care unit: implications for the safety of seriously ill patients]. *Rev Bras Ter Intensiva* 2014; **26**(1): 28-35.

Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses [*]

Samir Rachid Zaim[†], Qike Li[†], and A. Grant Schissler[†,‡]
Ctr for Biomed. Informatics & Biostatistics, Dept of Medicine, Grad. Interdisciplinary Prog. in Statist., The University of Arizona, 1657 E. Helen Street, Tucson, AZ, 85721, USA
Email: samirrachidzaim@email.arizona.edu, qikeli@email.arizona.edu, grant.schissler@gmail.com

Yves A. Lussier[§]
Center for Biomedical Informatics & Biostatistics, Dept of Medicine, Cancer Center, BIO5 Institute, The University of Arizona, 1657 E. Helen Street, Tucson, AZ, 85721, USA
Email: yves@email.arizona.edu

Recent precision medicine initiatives have led to the expectation of improved clinical decision-making anchored in genomic data science. However, over the last decade, only a handful of new single-gene product biomarkers have been translated to clinical practice (FDA approved) in spite of considerable discovery efforts deployed and a plethora of transcriptomes available in the Gene Expression Omnibus. With this modest outcome of current approaches in mind, we developed a pilot simulation study to demonstrate the untapped benefits of developing disease detection methods for cases where the true signal lies at the pathway level, even if the pathway's gene expression alterations may be heterogeneous across patients. In other words, we relaxed the cross-patient homogeneity assumption from the transcript level (cohort assumptions of deregulated gene expression) to the pathway level (assumptions of deregulated pathway expression). Furthermore, we have expanded previous **single-subject** (**SS**) methods into cohort analyses to illustrate the benefit of accounting for an individual's variability in cohort scenarios. We compare SS and **cohort-based** (**CB**) techniques under 54 distinct scenarios, each with 1,000 simulations, to demonstrate that the emergence of a pathway-level signal occurs through the summative effect of its altered gene expression, heterogeneous across patients. Studied variables include pathway gene set size, fraction of expressed gene responsive within gene set, fraction of expressed gene responsive up- vs down-regulated, and cohort size. We demonstrated that our SS approach was uniquely suited to detect signals in heterogeneous populations in which individuals have varying levels of baseline risks that are simultaneously confounded by patient-specific "**genome -by-environment" interactions** (**G×E**). Area under the precision-recall curve of the SS approach far surpassed that of the CB (1st quartile, median, 3rd quartile: SS = 0.94, 0.96, 0.99; CB= 0.50, 0.52, 0.65). We conclude that single-subject pathway detection methods are uniquely suited for consistently detecting pathway dysregulation by the inclusion of a patient's individual variability.

http://www.lussiergroup.org/publications/PathwayMarker/

Keywords: pathway, gene set, biomarkers, single-subject, cohort, precision medicine, kMEn, n-of-1

[*] This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)

[†] These authors contributed equally to this work
[‡] Work completed at The University of Arizona, author now at University of Nevada, Reno
[§] Corrsponding Author

1. Introduction

Recent precision medicine initiatives have led clinicians, patients, and investors to expect improved clinical decision-making anchored in genomic data science. Conventionally, precise prognostication and therapeutic decision-making relies on assays measuring the expression or activity of specific molecules driving a pathophysiological mechanism implicated in disease progression or drug response. To extend conventional biomarker discovery in the post-genome era, the NIH has invested more than $2.5 billion/year in hypothesis- and data-driven "biomarker" grants (>30,000 grants in 25 years) [1]. Yet, in the last decade, only a handful of new single-gene product biomarkers have been translated to clinical practice [2, 3] in spite of considerable discovery efforts deployed and a plethora of transcriptomes available in the Gene Expression Omnibus. This may be due, in part, to the challenging FDA requirements for biomarker qualification, which has conventionally required a high level of evidence on the degree of biological understanding between a qualified biomarker and the predicted pathophysiology or drug response [4]. Perhaps, the community has exhausted the reductionist approach for identifying one gene product expression associated to the prognosis or therapeutic response of complex diseases. Further, could it be that, as anticipated by statistical geneticists a decade and a half ago [5] and newly rediscovered [6], diseases of complex genetic inheritance (**complex diseases**) are not often amenable to the single gene biomarker reductionism that has worked so well for Mendelian diseases? Rather than modifying the FDA evidentiary criteria for biomarker qualification, we and others postulate that a paradigm shift is required for integrative or systems biology approaches to enable new types of biomarker discovery [7-10]. To address this biomarker dilemma, we propose to use two strategies jointly: (1) the discovery of pathway-level composite biomarkers consisting of multiple gene products that are combined in a stated algorithm to reach a single interpretive readout[**], and (2) the use of **single-subject** (**SS**) (isogenic) analytics to recover an effect size and statistical significance and thereafter aggregating these signals across subjects.

Why utilize SS analytics rather than DNA sequencing for pathway-level biomarkers? In practice, a single-subject transcriptome or proteome may be easier to interpret as it provides the downstream additive effects of genomes and proteomes, and thus there could, in principle, be more similarities between transcriptomes than genomes of distinct individuals suffering from a complex disease and responding similarly to a drug. Precision medicine has advanced primarily through DNA sequencing. Unsurprisingly, most DNA sequences remain uninterpretable: Snyder's group identified >130,000 very rare or private single nucleotide variants not previously observed in HAPMAP [11]. However, gene product expression cannot easily be annotated as normal or dysregulated on a single subject; therefore, a personal reference transcriptome or proteome should be designed ideally in isogenic conditions with a specific cell type in a specified environmental and known epigenetic context. Fortunately, the biomedical informatics and bioinformatics research community is responding to this growing need for identifying the best prognosis and therapeutic response for a specific individual with a paradigm shift in gene product analyses.

[**] Guidance for Industry and FDA Staff Qualification Process for Drug Development Tools. 1/2014 U.S. Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research (CDER). https://www.fda.gov/downloads/drugs/guidances/ucm230597.pdf

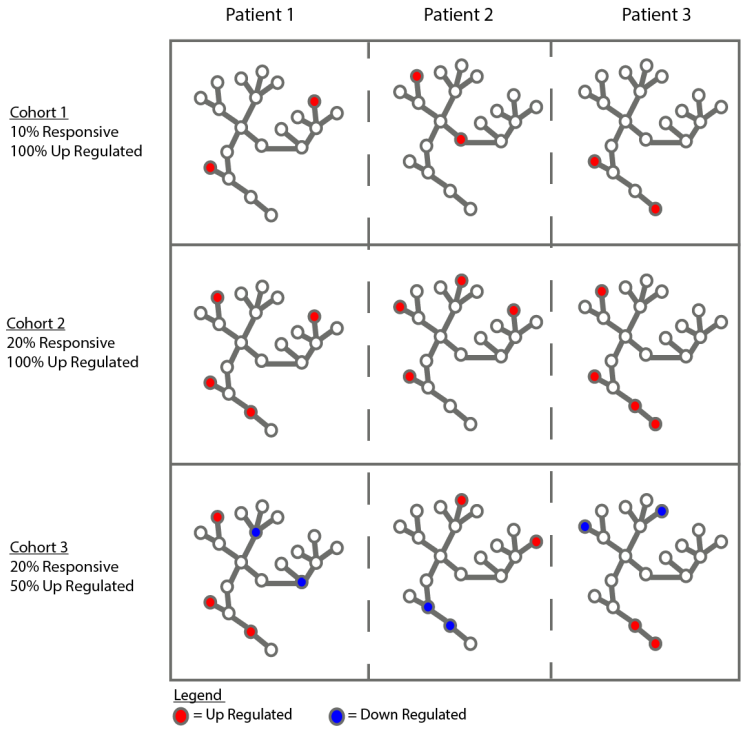


Fig. 1. Pathway dysregulation across various biological conditions. Each graph represents a patient (within three cohorts of three subjects), illustrating the same biological pathway for each patient. The nodes are the genes in a pathway. The colored nodes are responsive genes in each subject, and their color denotes the direction of dysregulation. The three rows represent various scenarios we examined in this simulation study to define a cohort.

Statistical and clinical frameworks are being developed for single-subject (**n-of-1**) interpretation of transcriptomes and transcriptomic responses, such as single sample pathway transformations [12, 13], and comparing their results to pathways expressed in differentially expressed genes discovered by conventional statistics. Newer studies have been designed to discover differentially expressed features in a single subject (gene products and pathways) and are based on reference transcriptome-based interpretations [14, 15], two paired samples [16, 17], or individual time expression series [18, 19]. None of these studies, nor related ones we recently reviewed[††], attempted to quantify how well the discovered single-subject gene set/pathway signal could aggregate across distinct subjects without an underlying assumption of having the same gene products differentially expressed.

Implicitly, these **single-subject** (**SS**) methods differ from conventional **cohort-based** (**CB**) statistics as they are devoid of cross-subject assumptions and could provide the framework for a common pathway-level biomarker across subjects stemming from the summative effect of distinct polymorphisms, distinct epigenetics, and distinct transcriptomes in each subject. We hypothesized that we could conduct a proof-of-concept simulation to establish that conditions of operations for discovery of a common biomarker are feasible in practice. Therefore, we designed a simulation

[††] Vitali F, Li Q, Schissler AG, Berghout J, Kenost C, Lussier YA*. Developing a 'personalome' for precision medicine: emerging methods that compute clinically interpretable effect sizes from single-subject omics. Brief Bioinform, Accepted.

study to identify pathway-level effect size and statistical significance within subject and then used descriptive statistics across-subject to find common pathways. We utilize the n-of-1-*pathways* kMEn method on two paired samples for its simplicity. Our goals are i) to understand the robustness of single-subject methods in heterogenic and heterogeneous expression scenarios across subjects that are ill-suited for conventional cohort-level discovery methods (e.g., paired T-test as a control), and ii) to demonstrate the benefit of including biological pathways as part of what constitutes a reference systems-level biomarker. As **Figure 1** shows, in dysregulated pathways, patients with the same condition may have different genes responsive to a stimulus when compared to paired samples (e.g., before and during therapy; cancer vs control tissue), making conventional differential expression or classification tasks inherently difficult when searching for a common gene product signal across subjects.

2. Methods

2.1. *Datasets*

An RNA-seq dataset was downloaded from GTEx[‡‡] and filtered to include only brain tissue samples. The resulting dataset contained 1,632 brain samples of distinct human individuals and 18,327 measured genes. This dataset was used to estimate average gene expression for patients in our simulation. Since donors had varying numbers of replicates, only donors with at least 8 replicates were kept to reliably estimate their sampling distributions (see Section 2.2). This criterion resulted in a reduction from 97 to 87 distinct patients. Gene Ontology Biological Processes (**GO-BP**)[20, 21] groups genes into their respective pathways (gene sets). The GO-BP dataset was downloaded in June 2015 using the *org.Hs.eg.db* package from Bioconductor[22].

2.2. *Parameter estimation: modeling heterogenic human paired samples*

Each gene's expression distribution parameters for each patient were estimated using the method of moments technique [23]. Our model assumes the Negative Binomial – NB(μ,θ) – distribution, and the GTEx dataset was used to estimate each gene's mean expression, μ, and its dispersion parameter, θ, where the dispersion parameter connects the mean to the variance as follows:

$$\sigma^2 = \mu + \mu^2 / \theta \quad (1)$$

When the variance was less than the mean and its distribution was consequently under-dispersed compared to the Poisson, we conservatively defined the gene expression to follow a Poisson(μ) distribution. A fold-change multiplier, K, was used to generate the responsive genes in dysregulated pathways, and K followed a Uniform(3,5) distribution to ensure separation between responsive and non-responsive genes. For non-responsive genes, K=1. Equations (2) and (3) show that the updated NB distribution for a gene, G_i , is actually a discrete mixture distribution where (2) is the underlying sampling distribution if the dysregulated gene is up-regulated with

[‡‡] https://gtexportal.org/home/datasets: RNA-Seq Data: GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct

probability *p*, and (3) is the sampling distribution if it is down-regulated with probability *1-p,* where $p \in [0,1]$.

$$G_i, \text{ up-regulated} \sim \text{p} * \text{NB}(\mu_i * K, \theta_i) \tag{2}$$

$$G_i, \text{ down-regulated} \sim \text{(1-p)} * \text{NB}(\mu_i * K^{-1}, \theta_i) \tag{3}$$

Equations (4) and (5) show the Poisson distribution for an under-dispersed gene, G_i.

$$G_i, \text{ up-regulated} \sim \text{p} * \text{Poisson}(\mu_i * K) \tag{4}$$

$$G_i, \text{ down-regulated} \sim \text{(1-p)} * \text{Poisson}(\mu_i * K^{-1}) \tag{5}$$

To establish heterogeneity, we assumed each patient as a distinct population with its own patient-specific population parameters. Therefore, for any given subject j, eqs. (2-5) become:

$$G_{ij} \sim \text{p} * \text{NB}(\mu_{ij} * K, \theta_{ij}) + \text{(1-p)} * \text{NB}(\mu_{ij} * K^{-1}, \theta_{ij}) \tag{6-7}$$

$$G_{ij} \sim \text{p} * \text{Poisson}(\mu_{ij} * K) + \text{(1-p)} * \text{Poisson}(\mu_{ij} * K^{-1}) \tag{8-9}$$

This results in N=87 distinct distributions from which we sample *n* patients without replacement, ensuring patient-specific baseline expression levels and modeling a heterogeneous population.

2.3. *Simulation Parameters*

Table 1. Simulation Parameters generating 54 distinct scenarios (1000 simulations/scenario)

Parameter	Notation	Values
Gene Set Size	*m*	40 200
Fraction Responsive within gene set	*r*	5% 10% 25%
Fraction Responsive up regulated	*p*	25% 50% 100%
Number of patients in cohort	*n*	10 20 30

Table 1 shows the different conditions of interest (54 combinations) that span our study. The gene set size parameter was chosen to analyze how the fraction of responsive genes within the gene set affects the detection ability in small gene sets (e.g., 5% responsive results in 2/40 genes responsive) vs. large gene sets (5% responsive results in 10/200 genes responsive). The fraction responsive within the gene set parameter was chosen to model the effect of randomly selecting *r* genes to be responsive in a dysregulated pathway in each patient. Clearly, when the fraction increases, the chances of the same gene being responsive across all patients increases. Similarly, the fraction responsive up-regulated was conceived to model the effect of randomly choosing the direction of dysregulation for the responsive genes in that pathway, such that even if the same gene is responsive across patients, their direction of dysregulation might not be. Finally, the number of patients in the cohort parameter was chosen to examine the needed size of a cohort for detecting a dysregulated pathway when its signal is reflected via the summative effect of the genes within it. The graphs in **Fig. 1** illustrate how varying the parameters affects dysregulated pathways across patients in a cohort.

Table 2. Algorithm

For each parameter combination, replicate the following 1000 times.

1. *Dataset Generation:* Simulate *N* Paired-Transcriptomes (representing normal, tumor) using heterogenic gene distributions described above.
   a. For each normal transcriptome, simulate gene expression levels by randomly sampling from each patient's baseline distribution.
   b. For each tumor transcriptome, first generate a normal transcriptome, then generate the positive dysregulated pathway as follows:
      i. Choose a gene set size, *m*, and randomly sample a pathway of size *m* from GO-BP.
      ii. Randomly choose *r* genes from the selected pathway.
      iii. For each of the *r* genes, sample from its dysregulated distribution such that, each *r* dysregulated genes, G_i follows a discrete mixture distribution where
         1. $G_{ij} \sim p * NB(\mu_{ij} * K, \theta_{ij}) + (1-p) * NB(\mu_{ij} * K^{-1}, \theta_{ij})$, or
         2. $G_{ij} \sim p * Poisson(\mu_{ij} * K) + (1-p) * Poisson(\mu_{ij} * K^{-1})$ if the gene is under-dispersed
      iv. For all remaining genes (i.e. genes not in the gene set), these genes remain unaltered and follow the patient's baseline distribution.
   c. For each tumor transcriptome generate a control pathway by randomly sampling a pathway of size *m* (from GO-BP) and leave its expression values unaltered such that the genes in the control pathway follow the patient's baseline distribution.
2. *Cohort-Based Analysis:* Compute a paired t-test for the paired samples across each gene product and detect differentially expressed genes (DEGs), labeling a gene DEG if nominal $p < .05$. Using the DEGs and GO-BP, conduct an enrichment test using Fisher's Exact Test (FET)[26] to obtain the FET pathway prediction for the positive and control pathways, respectively. Adjust p-values for multiple hypothesis testing using FDR_BY [25].
3. *Single-Subject Analysis:* Perform an N-of-1-*pathways* kMEn analysis to obtain a pathway prediction (a pair of p-values – one for the positive and one for the control pathway) for each patient. Utilize the median of the positive and control pathway predictions to serve as an aggregate cohort-level result. Adjust p-values for multiple hypothesis (FDR_BY[25]).

2.4. *Pathway dysregulation detection methods*

Table 2 details the workflow of this simulation study (**Fig. 2**). We generate *n* heterogenic transcriptomes, each corresponding to one subject (heterogenic conditions between patients). We then generate from each patient distribution a paired transcriptome thus creating noise in isogenic conditions, in which we further modify a pathway as follows. First, we randomly select a gene set from a real GO-BP pathway of size *m*, randomly select which annotated genes among this gene set will be responsive according to parameters of **Table 1**, and we sample from their dysregulated distributions (**Eq. 6-9**) to generate a positive (dysregulated) pathway. Then, we select a second gene set from a distinct existing GO-BP, of size *m* as well, as an unaltered pathway to use as our control. Finally, we apply the SS and CB pathway detection pipelines, and then compare and evaluate them. We note that the single-subject approach aggregates the p-values by taking the sample median, as the sample median provides a simple yet robust location estimator in small sample sizes, which provided us with the flexibility of experimenting with sample sizes of $n < 10$ [24]. Furthermore, Benjamini and Yekutieli's (**FDR_BY**) approach is used for false discovery rate correction[25].

2.5. *Precision recall calculations*

In this study, we evaluate the SS and CB approaches using precision-recall plots. Provided a given threshold, the equations for precision and recall are:

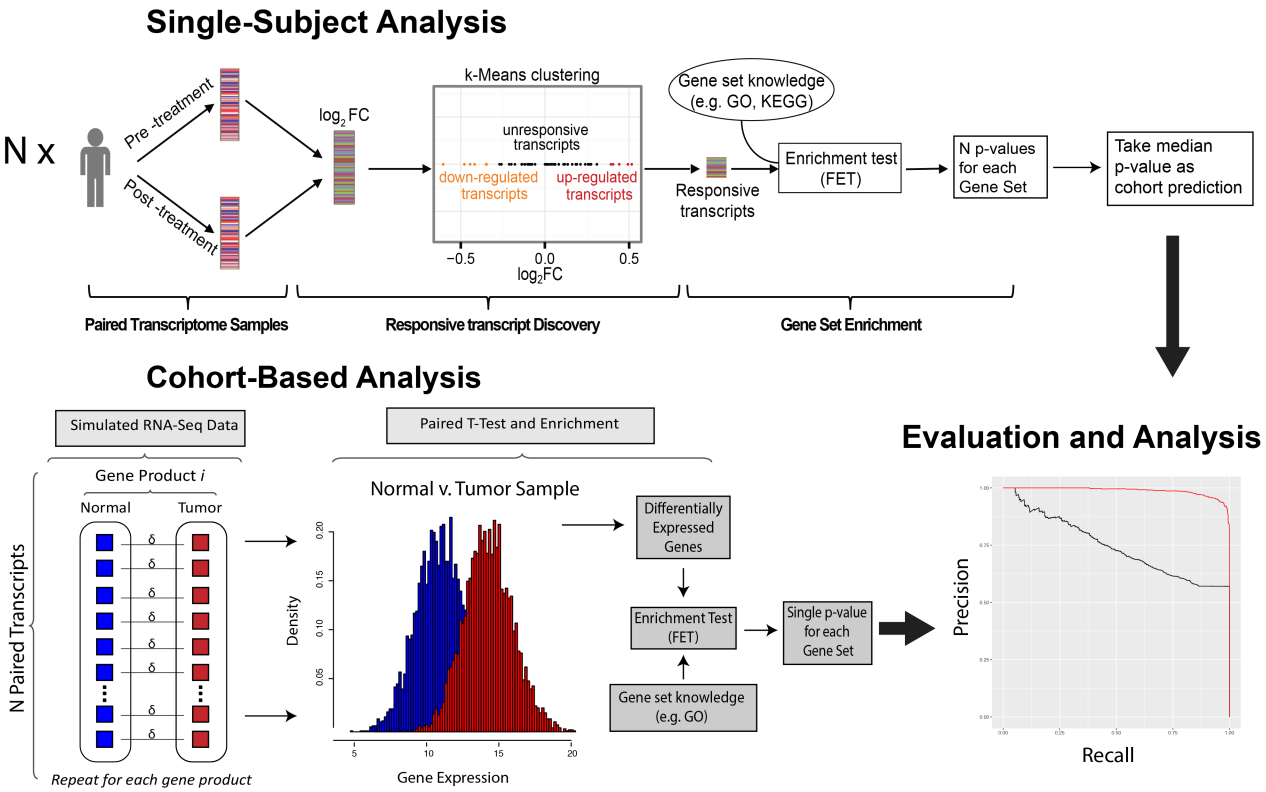


Fig. 2. Workflow: Single-subject (SS) and cohort-based (CB) pipelines. (SS; top) Given the simulated values as input, the transcript expression measurements for each of the *N* paired transcriptomes are used to calculate the fold change (K) between paired samples. Next, the genes are clustered into three groups to define responsive transcripts (RTs), and then an enrichment test is conducted using Fisher's Exact Test (FET). This produces N p-values (one for each patient in the cohort), and the median p-value is taken as the kMEn-cohort prediction. (CB; bottom) Using the same simulated data, alternatively examined CB approach employs a paired t-test to find differentially expressed genes (DEGs) followed by an enrichment test using FET, resulting in a single pathway prediction utilizing all samples. Both approaches are then compared by inspecting their precision-recall curves.

$$precision = \frac{tp}{tp + fp} \quad recall = \frac{tp}{tp + fn} \tag{10-11}$$

Each of the 54 combinations results in a pair of precision-recall curves that are used to compare SS to CB approaches. The R *ggplot2* package[27] was used to construct the precision-recall plots.

3. Results

Fig. 3 depicts the precision-recall curves, grouping them by their parameters to highlight the effects of each the parameters individually and holistically. The greatest difference in performance between the SS approach and the CB technique occurs when responsive genes are fully bi-directional (i.e., equally expressed in both directions; Fraction Responsive up-regulated, p = 50%)

or when the same genes are not consistently responsive across pathways (fraction responsive with in gene set = 5%). The smallest gap in performance between these methods occurs when the fraction responsive within gene set is high (as genes are more likely to be responsive consistently across patients) and, in some cases, when the precision-recall curves are overlapping. Increasing

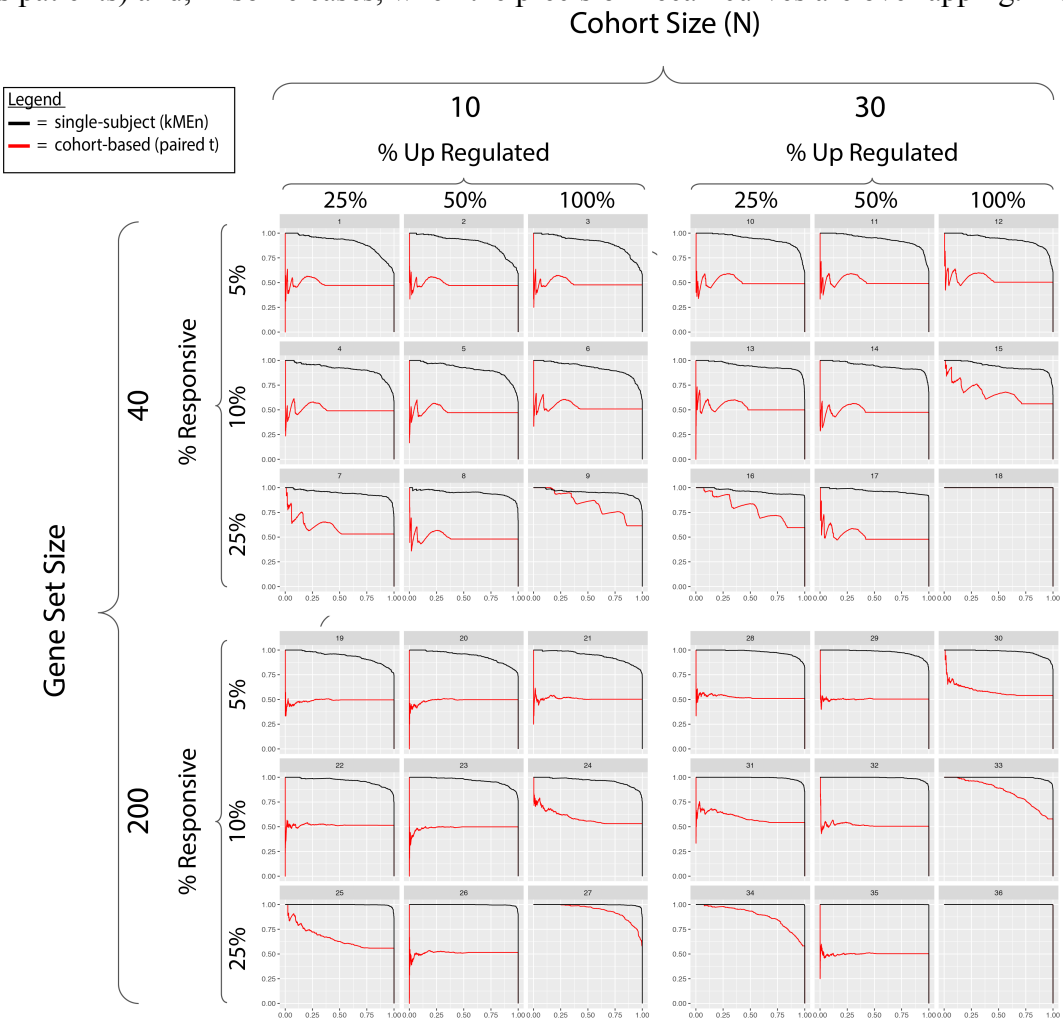


Fig. 3. Cross-subject aggregation of single-subject pathway predictions (kMEn) robustly detects signals while cohort-based method (Student's paired t-test) fails on heterogeneous conditions. SS kMEn method applied to paired samples of one subject works in isogenic conditions by design, which explains how pathway signals can thereafter be aggregated across subjects in spite of heterogenic noise confounding the conventional cohort-based method. Each subject simulation comes from a distinct transcriptome sampled from GTEx, creating heterogenic conditions between subjects. Each seed sample from GTEx is modified according to parameters in **Table 1** generating distinct scenarios. The four sets of panels characterize distinct scenarios in the simulation and are organized in blocks. Within each block, there are various levels of 'signal quality' and across blocks there are different combinations of cohort and pathway sizes. The 9 **precision-recall curves (PR)**, within each block, represent the performance of the SS (black) and CB (red) approaches at various levels of genes responsive in a pathway as well as at various levels of bi-directionality. PR area under the curve (AUC) of SS surpasses that of CB (1st quartile, median, 3rd quartile: SS = 0.94, 0.96, 0.99; CB= 0.50, 0.52, 0.65). Each block represents how both methods perform when varying the gene set size or the cohort size. Omitted are the

results for N=20 to promote visualization and they are highly similar to the N=30 scenarios. Using GO-BP2017, we simulated test cohorts (n=3 subjects) and obtained comparable accuracies.

the pathway size and the sample size also improves the detection-ability of both approaches though the marginal benefit of increasing each parameter is much larger for the CB approach.

The panels in **Fig. 3** allow for visually assessing the effects of varying multiple parameters simultaneously. For example, increasing the number of responsive genes in the pathway compensates for adding bi-directionality into the mix (and *vice versa*), although the SS approach still detects the signal at a much higher rate than the CB approach. Furthermore, increasing the pathway size and/or the cohort size improves the performances of both approaches in most cases. The simulation settings where the CB method is comparable to the SS approach is when the signal is strongest (% responsive = 25%), N is large, and there is little or no bi-directionality in gene expression levels. This shows that outside of this specific condition, even CB approaches that can handle bi-directionality will still be underpowered (in varying levels) vis-à-vis an SS approach.

4. Discussion

As mentioned in Section 3, two of the biggest indicators of whether the t-test would fail are pathways with different genes responsive across patients (Fraction responsive within gene set = 5%) and pathways with genes equally expressed in both directions (Fraction responsive up regulated = 50%). Not surprisingly, one of the biggest differences in performances occurs with full bi-directionality with a method like the t-test, and methods like DEGSeq address this [28]. However, as illustrated in **Fig. 1**, when the signal lies at the pathway-level, different genes are responsive in different patients (as well as potentially their direction of dysregulation). This means that in a cohort of three patients, the same gene in a dysregulated pathway could be (responsive, up-regulated) in Patient 1, (responsive, down-regulated) Patient 2, or non-responsive in Patient 3, rendering a CB approach nearly unusable. Therefore, decreasing the fraction responsive within gene set parameter shows how a CB approach greatly underperforms when the true signal lies at the pathway level and it attempts detecting it through genes not consistently responsive across patients. In addition, heterogeneous baseline risks add an extra layer of complexity that CB approaches are not equipped to handle since an up-regulated responsive gene in Patient A might have a lower expression level than the same gene, non-responsive in Patient B. These factors, individually and in aggregate, make an SS method uniquely suited for detecting diseases in individuals when patient-specific factors harm CB approaches and when we allow biological pathways to represent a reference systems-level biomarker. Finally, taking a consensus of the SS predictions results in a robust cohort prediction that can consistently detect converging gene set signals in heterogenic populations via the summative effect of altered gene expression.

4.1. *Limitations and future studies*

One of the major challenges in simulation studies is the inclusion of noise and the effects introduced into the analysis; here we used a single model source. Of note, each patient of a cohort in the simulation is seeded by a distinct transcriptome distribution from GTEx and noise is generated implicitly by the algorithm on the entire transcriptome of each paired sample, creating

isogenic noise within patient as well as heterogenic noise across subject conditions *ab initio*. In future studies, real data will be utilized to estimate the fraction responsive (5%-25%) and fraction upregulated (25%-100%) parameters according to the type of diseases. Currently the wide range of simulation of these parameters likely spans multiple distinct unrelated biology and should be clarified (e.g. Mendelian diseases vs cancer vs diabetes)."

The scope for this proof of concept was also limited to one single-subject and one cohort-based approach. Since the kMEn algorithm and the enriched paired t-test are by no means the only SS and CB approaches, respectively, we foresee potential follow-up studies with multiple SS [3] and multiple CB methods [29] in order to find which techniques within these two frameworks are best suited to handle this type of data. A more comprehensive analysis would then allow us to make broader claims with respect to the feasibility of detecting diseases using biological pathways as biomarkers in heterogeneous patient populations.

With this simulation study, we demonstrate the benefits of expanding the definition of a biomarker by illustrating biological conditions in which the 'true' signal is not detectable at the gene level, and must, therefore, be pushed upstream to the pathway level. As **Fig. 3** shows, the CB method achieved comparable performance in only 6 out of the 36 simulation conditions. Unless an infrequent "niche" scenario is present, this (and potentially other CB methods with the same drawbacks) will fail to consistently detect diseases whose signals are found at the pathway level. Expanding SS methods into cohort studies and allowing for pathways to serve as a reference biomarker in disease detection have the potential to offer more tools for detecting diseases in cases where existing methods have failed to provide consistent success.

Clearly, the exhaustive conventional biomarker discovery effort to identify a single gene product consistently dysregulated in each patient with complex disorders yields infrequent results at best. Moreover, the difficulty increases when within-subject biological replicates are not available either due to limited tissue availability or invasive tissue-sampling procedures among other cost-preventive limitations. Despite the decreasing costs associated with advancing RNA-seq technologies, the incentives still favor sequencing more subjects rather than obtaining multiple biological replicates per subject. Future studies should test in human datasets (both with and without subject-specific biological replicates) using various experimental conditions, mitigating geneset enrichment inflation due to inter-transcript correlations [32], to understand the frequency of the proposed scenario of heterogeneous signal within a pathway across patients. While kMEn's algorithm requires a large transcriptome, democratizing pathway-level biomarkers as an affordable qPCR assay can be attained with self-contained approaches [31,33].

5. Conclusion

As medicine continues to shift towards precision medicine and the n-of-1 framework, it will be necessary to consider novel approaches for effectively qualifying biological pathways for FDA approval as composite biomarkers[30]. We provide evidence via this proof-of-concept study that, under certain conditions, this may be the optimal way of detecting pathway mechanisms associated to the prognosis or drug response of complex diseases, as the signal may consistently

aggregate at the pathway level in each subject in spite of a distinct subset of transcript dysregulation across subjects.

This simulation was developed to show the potential advantages of using a pathway as a biomarker using the 'N-of-1-*pathways*' framework [31] and that **single-subject** (**SS**) approaches (expanded into cohort studies) can provide certain advantages over conventional cohort-based techniques. We demonstrated that our SS approach was uniquely better suited to detect signals in heterogeneous populations in which individuals have varying levels of baseline risks that are simultaneously confounded by patient-specific "genome –by– environment" interactions (**G×E**).

Finally, these approaches should, in principle, scale to other quantitative 'omics measures such as proteomics or metabolomics. Future studies should consider aggregating pathway signals across multiple 'omics measures in heterogeneous conditions across patients using strong systems biology modeling of a single subject for consistency of multiscale signal within patient (e.g., reverberation of a pathway-level signal from DNA to mRNA to protein). The success of precision medicine demands advancing genome-anchored clinical decision-making and having the courage to challenge failed or unproductive data analytics models. A handful of statistical geneticists has long anticipated that epistasis, pleiotropy, and systems biology principles be incorporated for effectively modeling genomics data. This proof of concept brings us closer to realizing their vision in transforming the biomarker discovery process.

6. Acknowledgements

The author would like to thank Dr. Colleen Kenost for manuscript revisions.

References

1. Ptolemy, A.S. and N. Rifai, *What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema.* Scand J Clin Lab Invest Suppl, 2010. **242**: p. 6-14.
2. Fuzery, A.K., et al., *Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges.* Clin Proteomics, 2013. **10**(1): p. 13.
3. Pavlou, M.P., E.P. Diamandis, and I.M. Blasutig, *The long journey of cancer biomarkers from the bench to the clinic.* Clin Chem, 2013. **59**(1): p. 147-57.
4. Rifai, N., M.A. Gillette, and S.A. Carr, *Protein biomarker discovery and validation: the long and uncertain path to clinical utility.* Nat Biotechnol, 2006. **24**(8): p. 971-83.
5. Ritchie, M.D., et al., *Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.* Am J Hum Genet, 2001. **69**(1): p. 138-47.
6. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.
7. McDermott, J.E., et al., *Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data.* Expert Opin Med Diagn, 2013. **7**(1): p. 37-51.
8. Moore, J.H., *A global view of epistasis.* Nat Genet, 2005. **37**(1): p. 13-4.
9. Massague, J., *Sorting out breast-cancer gene signatures.* N Engl J Med, 2007. **356**(3): p.294-7.
10. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* Nat Rev Genet, 2014. **15**(1): p. 34-48.

11. Chen, R., et al., *Personal omics profiling reveals dynamic molecular and medical phenotypes.* Cell, 2012. **148**(6): p. 1293-307.
12. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.
13. Yang, X., et al., *Single sample expression-anchored mechanisms predict survival in head and neck cancer.* PLoS Comput Biol, 2012. **8**(1): p. e1002350.
14. Liu, R., et al., *Identifying critical transitions of complex diseases based on a single sample.* Bioinformatics, 2014. **30**(11): p. 1579-86.
15. Drier, Y., M. Sheffer, and E. Domany, *Pathway-based personalized analysis of cancer.* Proc Natl Acad Sci U S A, 2013. **110**(16): p. 6388-93.
16. Li, Q., et al., *kMEn: Analyzing noisy and bidirectional transcriptional pathway responses in single subjects.* J Biomed Inform, 2017. **66**: p. 32-41.
17. Li, Q., et al., *N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes.* BMC Med Genomics, 2017. **10**(Suppl 1): p. 27.
18. Wu, S. and H. Wu, *More powerful significant testing for time course gene expression data using functional principal component analysis approaches.* BMC Bioinformatics, 2013. **14**: 6.
19. Martini, P., et al., *timeClip: pathway analysis for time course data without replicates.* BMC Bioinformatics, 2014. **15 Suppl 5**: p. S3.
20. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nat Genet, 2000. **25**(1): p. 25-29.
21. *Gene Ontology Consortium: going forward.* Nucleic Acids Research, 2015. **43**(D1):D1049-56.
22. Gentleman, R.C., et al., *Bioconductor: Open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**.
23. Casella, G. and R.L. Berger, *Statistical inference*. Vol. 2. 2002: Duxbury Pacific Grove, CA.
24. Rousseeuw, P.J. and S. Verboven, *Robust estimation in very small samples.* Computational Statistics & Data Analysis, 2002. **40**(4): p. 741-758.
25. Benjamini, Y. and D. Yekutieli, *The Control of the False Discovery Rate in Multiple Testing under Dependency.* The Annals of Statistics, 2001. **29**(4): p. 1165-1188.
26. G., U.J., *Fisher's Exact Test.* Journal of the Royal Statistical Society, 1992. **155**(3): p. 395-402.
27. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis.* Springer, 2009.
28. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics, 2010. **26**(1): p. 136-8.
29. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biology, 2010. **11**(10): p. R106.
30. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
31. Gardeux, V., et al., *'N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine.* J Am Med Inform Assoc, 2014. **21**(6): p. 1015-25.
32. Schissler AG, et al., *Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation.*. Stat Methods Med Res, 2017 Jan 1:962280217712271.
33. Schissler AG, et al., *Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival*, Bioinformatics. 2015 **10**;31(12):i293-302.

Analyzing metabolomics data for association with genotypes using two-component Gaussian mixture distributions

Jason Westra
Department of Statistics, Iowa State University
Ames, IA 50011, United States
Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, United States
Email: jwestra@iastate.edu

Nicholas Hartman
Department of Biological Statistics and Computational Biology, Cornell University
Ithaca, NY 14853, United States
Email: ngh32@cornell.edu

Bethany Lake
Department of Mathematics and Statistics, Elon University
Elon, NC 27244, United States
Email: blake@elon.edu

Gregory Shearer
Department of Nutritional Sciences, Pennsylvania State University
University Park, PA 16801, United States
Email: gcs13@psu.edu

Nathan Tintle
Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, United States
Email: Nathan.tintle@dordt.edu

Standard approaches to evaluate the impact of single nucleotide polymorphisms (SNP) on quantitative phenotypes use linear models. However, these normal-based approaches may not optimally model phenotypes which are better represented by Gaussian mixture distributions (e.g., some metabolomics data). We develop a likelihood ratio test on the mixing proportions of two-component Gaussian mixture distributions and consider more restrictive models to increase power in light of *a priori* biological knowledge. Data were simulated to validate the improved power of the likelihood ratio test and the restricted likelihood ratio test over a linear model and a log transformed linear model. Then, using real data from the Framingham Heart Study, we analyzed 20,315 SNPs on chromosome 11, demonstrating that the proposed likelihood ratio test identifies SNPs well known to participate in the desaturation of certain fatty acids. Our study both validates the approach of increasing power by using the likelihood ratio test that leverages Gaussian mixture models, and creates a model with improved sensitivity and interpretability.

Keywords: Metabolomics; Gaussian Mixture Distributions; Fatty Acids

1. Introduction

Genome-wide association studies (GWAS) continue to be viewed as a standard approach to evaluating the genetic component of a variety of diseases and other phenotypes of interest [1]. Standard approaches to the analysis of genotype associations with quantitative phenotypes use linear models.

As suggested in Tintle et al. [2], bimodal distributions are frequently observed in continuous phenotype samples of metabolites, challenging the normality assumption needed in many existing GWAS analysis approaches. For example, red blood cell fatty acid levels have been found to contribute to coronary heart disease [3]. As outlined in Tintle et al. [2], it is biologically reasonable to consider one's fatty acid levels as coming from a mixture of Gaussian distributions, with each of the two or three mean fatty acid levels determined by genetics, and variation around the mean level determined by other factors (e.g., diet; lifestyle). While the standard way of analyzing fatty acids follows the typical GWAS linear model approach, in cases where the distribution does not appear to be normally distributed, a log transformation is sometimes used [4]. However, this log transformation may fail to accurately capture the true distribution of the genotypic and phenotypic data since it ignores the biological reasoning for observing a non-normal distribution. It may be more powerful to directly model the normal mixture distribution and then test for genotype-phenotype association.

Recently, Kim et al. proposed a likelihood ratio test to test for association between copy number polymorphisms (CNP) with quantitative phenotypes and case control outcomes which followed a mixture of Gaussian distributions [5]. The likelihood ratio test evaluates possible differences in the mixing proportions of the Gaussian components by different copy number. Kim et al. showed that the likelihood ratio test was more powerful than a 2 x d chi-squared test with d equal to the number of CNP categories when the underlying data was from a mixture distribution.

We propose adapting the Kim et al. likelihood ratio test to the standard genotype-phenotype testing situation for phenotypes which are distributed as a mixture of Gaussian distributions, like some metabolomics data (e.g., fatty acid levels). We will provide a theoretical framework for the likelihood ratio test, evaluate its performance on simulated data and then apply it to a real set of fatty acid data from the Framingham Heart Study.

2. Methods

2.1. *Notation*

Let X be a quantitative phenotype that follows a two-component Gaussian mixture distribution. Thus, $X \sim \pi N(\mu_1, \sigma^2) + (1-\pi)N(\mu_2, \sigma^2)$ where π is the mixing parameter of the Gaussian components. Let μ_1 and μ_2 be the mean parameters such that $\mu_1 \neq \mu_2$, and we assume a common variance σ^2 for both components. We assume $\pi = p_{01}(n_0/\mathrm{N}) + p_{11}(n_1/\mathrm{N}) + p_{21}(n_2/\mathrm{N})$ where p_{t1} (t = 0, 1, 2) is the proportion of genotype t in the first component of the mixture distribution, n_t (t = 0, 1, 2) is the number of individuals with genotype t, and N is the total number of individuals. We consider the null hypothesis $\mathrm{H}_0 : p_{01} = p_{11} = p_{21}$ and the alternative H_a: at least one is not equal. Let $p_{\phi i} = p_{0i} = p_{1i} = p_{2i}$ $(i = 1,2)$ (see *Figure 1* for a visual representation). Let x_{tb} (b = 1, 2, …

n_t) and (t = 0, 1, 2) be a random variable representing the phenotype for individual b who has genotype t, and let w be a vector of all x_{tb}. Across all the components, the mixing proportion for genotype t must sum to 1 such that $p_{t1} + p_{t2} = 1$ (t = 0 , 1, 2).

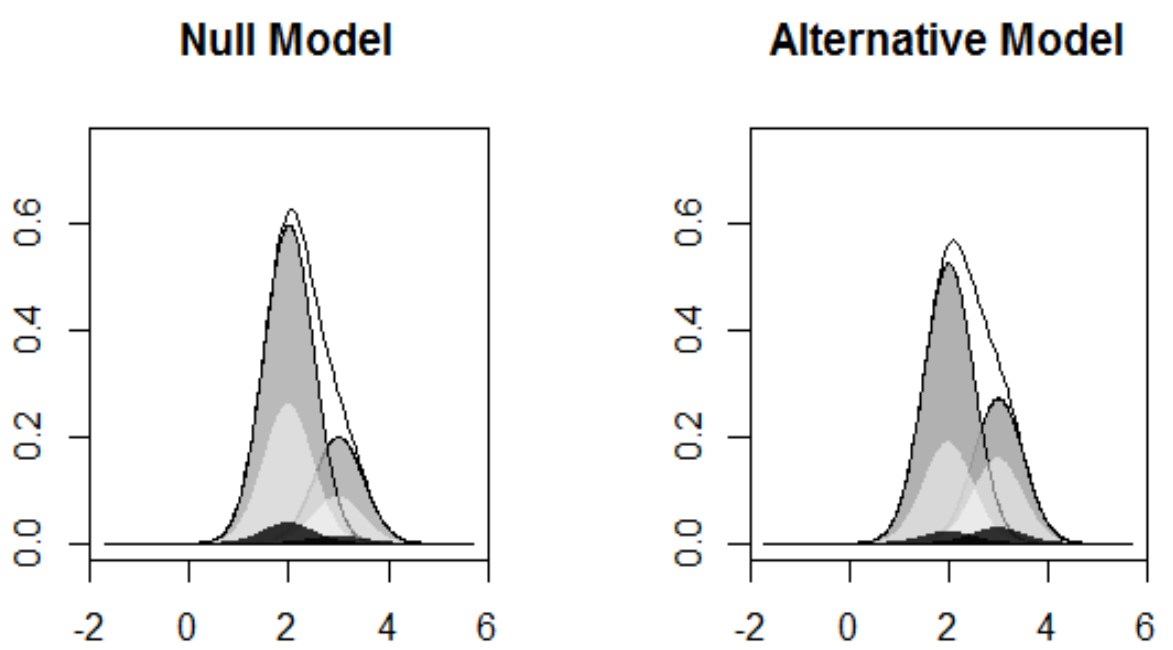


Figure 1 visually illustrates the null and alternative models. The black, light grey, and dark grey two-component mixture distributions are the phenotype distributions for the less common homozygote, the heterozygote and the more common homozygote, respectively. In the null model, 75% of the observations in each genotype are in the component with the smaller mean. In the alternative model, the mixing proportion for the component density with the smaller mean varies across genotypes.

2.2. *Likelihood functions*

2.2.1. *Null and alternative likelihood function*

The likelihood function under the null hypothesis is:

$$L_0 = \prod_{j=1}^{n_0+n_1+n_2} \left(\sum_{i=1}^{2} p_{\phi i} N(w_j | \mu_i, \sigma^2) \right) \quad (1)$$

The likelihood function under the unrestricted alternative hypothesis is:

$$L_1 = \left(\prod_{k=1}^{n_0} \left(\sum_{i=1}^{2} p_{0i} N(x_{0k} | \mu_i, \sigma^2) \right) \right) \left(\prod_{m=1}^{n_1} \left(\sum_{i=1}^{2} p_{1i} N(x_{1m} | \mu_i, \sigma^2) \right) \right) \left(\prod_{h=1}^{n_2} \left(\sum_{i=1}^{2} p_{2i} N(x_{2h} | \mu_i, \sigma^2) \right) \right) \quad (2)$$

2.2.2. *Restricted likelihood function*

When there is a biological understanding of the phenotype-genotype relationship, we recommend restricting the mixing proportions of the test to fit the biological model. We demonstrate two possible models, but our general method easily extends to other models. The first model (LRT_{pro}; Table 1) we consider is that the proportion of change between genotypes 0 and 1 is equal to the change between genotypes 1 and 2. Therefore, we can restrict our parameters of interest to $p_{0i}^* = (p_{01}, 1 - p_{01})$, $p_{1i}^* = \big(p_{01}q, 1 - (p_{01}q)\big)$, and $p_{2i}^* = \big(p_{01}q^2\,, 1 - (p_{01}q^2)\big)$. The second restricted model (LRT_{add}; Table 2) that we demonstrate describes an equal difference in proportions between groups 0 and 1 and groups 1 and 2. We can restrict our parameters of interest to $p_{0i}^* = (p_{01}, 1 - p_{01})$, $p_{1i}^* = \big(p_{01} - q, 1 - (p_{01} - q)\big)$, and $p_{2i}^* = \big(p_{01} - 2q\,, 1 - (p_{01} - 2q)\big)$. Therefore, the likelihood function under these restrictions is:

Table 1. LRT_{pro}

Genotype	Component 1 of Mixture Distribution	Component 2 of Mixture Distribution
0	p_{01}	$1 - p_{01}$
1	$p_{01}q$	$1 - (p_{01}q)$
2	$p_{01}q^2$	$1 - (p_{01}q^2)$

Table 2. LRT_{add}

Genotype	Component 1 of Mixture Distribution	Component 2 of Mixture Distribution
0	p_{01}	$1 - p_{01}$
1	$p_{01} - q$	$1 - (p_{01} - q)$
2	$p_{01} - 2q$	$1 - (p_{01} - 2q)$

$$L_2 = \left(\prod_{k=1}^{n_0}\left(\sum_{i=1}^{2} p_{0i}^* N(x_{0k}|\mu_i, \sigma^2)\right)\right)\left(\prod_{m=1}^{n_1}\left(\sum_{i=1}^{2} p_{1i}^* N(x_{1m}|\mu_i, \sigma^2)\right)\right)\left(\prod_{h=1}^{n_2}\left(\sum_{i=1}^{2} p_{2i}^* N(x_{2h}|\mu_i, \sigma^2)\right)\right) \quad (3)$$

2.2.3. *Test statistics*

Because $p_{t2} = 1 - p_{t1}$ for all *t*, we can express each likelihood as a function of the parameters μ_1, μ_2, σ^2, and the mixing proportion(s) associated with the $N(\mu_1, \sigma^2)$ distribution. The resulting likelihood ratio test statistics are given by:

$$LRTS = 2\Big(\max_{p_{01}, p_{11}, p_{21}, \mu_1, \mu_2, \sigma^2} \ln(L_1) - \max_{p_{\phi 1}, \mu_1, \mu_2, \sigma^2} \ln(L_0)\Big) \quad (4)$$

$$LRTS_{res} = 2\Big(\max_{p_{01}, q, \mu_{1,}, \mu_2, \sigma^2} \ln(L_2) - \max_{p_{\phi 1}, \mu_1, \mu_2, \sigma^2} \ln(L_0)\Big) \quad (5)$$

Extending the argument provided by Kim et al. the LRTS under the null hypothesis follows a central chi-squared distribution with the degrees of freedom equal to the difference in parameters of the null and alternative models [5]. Therefore, under the null hypothesis, the LRTS has a central chi-squared distribution with 2 degrees of freedom, and the $LRTS_{res}$ follows a central chi-squared distribution with 1 degree of freedom.

2.3. *Simulation*

Using R software, we simulated 1000 datasets with 10,000 individuals per data set. For each, individual, the genotype for a single SNP was generated by assuming Hardy-Weinberg equilibrium and minor allele frequency of either 0.05, 0.10, or 0.25. Trait values for individuals were simulated from two component Gaussian mixture distributions with centers one unit apart and equal variance of the components σ^2 = 0.5 or 0.75. For the mixing proportions of individuals with genotype 0, we used p_{01} = 0.9 or p_{01} = 0.75. We used two different biological models to simulate. In the proportional model we set q equal to 1, 0.9, or 0.75 so that the other mixing proportions were $p_{11} = p_{01}q$ and $p_{21} = p_{01}q^2$. In the additive model we set q equal to 0.1 or 0.2 so that the mixing proportions were $p_{11} = p_{01} - q$ and $p_{21} = p_{01} - 2q$. Simulations were performed on all combinations of the parameters.

2.4. *Statistical analysis*

To evaluate the performance of these tests in direct comparison to the standard procedure of linear and log-linear models, all tests were run on each simulated SNP and phenotype. Each test produced a *p*-value, test statistic and parameter estimates. Type I error rates and power estimates were calculated by dividing the number of observations less than a significance level (Type I error 0.01, power 0.0001) by the total number of simulations. We used an Expectation Maximization (EM) algorithm to find the global maximums of equations (4) and (5). One hundred random start points (RSP) were used for the null likelihood, and 50 RSP and one start point from the maximum of the null were used in the alternative [5]. The EM algorithm ran until a tolerance of 10^{-5} was reached or until 600 and 300 iterations were performed for the null and alternative models respectively.

2.5. *Real data application*

We analyzed 20315 SNPs on chromosome 11 for 5936 individuals from the Framingham Heart Study using the proposed LRT_{pro} test. We looked exclusively at members in the offspring and generation 3 cohorts, all of whom are of European descent. Detailed descriptions of the sample are available elsewhere [6]–[9]. We looked at the red blood cell fatty acid level ratio of arachidonic acid (AA) to dihomo-gamma-linoleneic acid (DGLA). These fatty acid levels were analyzed by gas chromatography as previously described [6]. The desaturation of AA to DGLA occurs primarily via enzymatic activity in the FADS gene complex on chromosome 11. We will use a Bonferroni correction to control the probability of type I errors at 2.47×10^{-6} (0.05/20315).

3. Results

3.1. *Verifying the null distribution and type I error rate*

To confirm that the null distribution of the unrestricted model is a chi-square distribution with two degrees of freedom and that the null distribution of the restricted model is a chi-square distribution with one degree of freedom, we examined simulations when $q = 1$. In addition to examining the novel tests proposed here (LRT_{pro}, LRT_{add}) we also explored the type I error rates of the linear model, log-linear model, and LRT across these same simulations. As shown in Table 3 the type I error rate was controlled by all tests.

Table 3. Type I Error Estimates

		Nominal Significance Level			
	SD	0.05	0.01	0.001	Kolmogorov-Smirnov test p-value[1]
LRT_{pro}	0.5	0.0497	0.011	0.0012	0.6846
	0.75	0.0515	0.0097	0.0010	0.8832
LRT_{add}	0.5	0.0472	0.0108	0.0012	0.7277
	0.75	0.0495	0.0085	0.0008	0.7091
LRT	0.5	0.0557	0.0108	0.0012	0.2269
	0.75	0.0478	0.0078	0.0013	0.7435
Linear Model	0.5	0.0538	0.0107	0.0007	
	0.75	0.0458	0.0070	0.0005	
Log Linear Model	0.5	0.0523	0.0108	0.0007	
	0.75	0.0460	0.0083	0.0007	

[1]As compared to a chi-square distribution.

3.2. *Power estimates*

There were 48 simulations where the alternative hypothesis was true. As summarized in Table 4 (full detailed results are in Supplemental Table 1), the LRT_{pro} has empirical power equal to or greater than all the other tests in all situations. LRT_{add} was the second most powerful test in all 48 simulations. When comparing a linear model to the unconstrained LRT test directly there were 21 simulations where they had different power. In two-thirds of these cases (14 out of 21), LRT had higher power than the linear model. The log-linear model never had an empirical power higher than any other test.

Table 4 Power Estimates

model	q	maf	p_{01}	Linear Model	Log Linear Model	LRT_{pro}	LRT_{add}	LRT
add	0.1	0.05	0.75	0.343	0.26	0.403	0.39	0.295
			0.9	0.44	0.316	0.631	0.624	0.508
		0.1	0.75	0.824	0.736	0.879	0.871	0.798
			0.9	0.898	0.793	0.967	0.966	0.938
		0.25	0.75	0.999	0.997	0.999	0.999	0.999
			0.9	0.999	0.999	1	1	1
pro	0.9	0.05	0.75	0.12	0.095	0.156	0.153	0.105
			0.9	0.325	0.212	0.478	0.467	0.362
		0.1	0.75	0.388	0.31	0.46	0.451	0.342
			0.9	0.75	0.622	0.891	0.887	0.831
		0.25	0.75	0.904	0.844	0.936	0.932	0.892
			0.9	0.998	0.975	1	1	1

Power estimates for standard deviation of .75 for alpha = 0.0001

The choice of 0.0001 as a cutoff for our power estimates is arbitrary as Figure 2 demonstrates. The LRT_{pro} tends to have a smaller *p*-value than the linear model for all thresholds since almost all of the points are above the gray line.

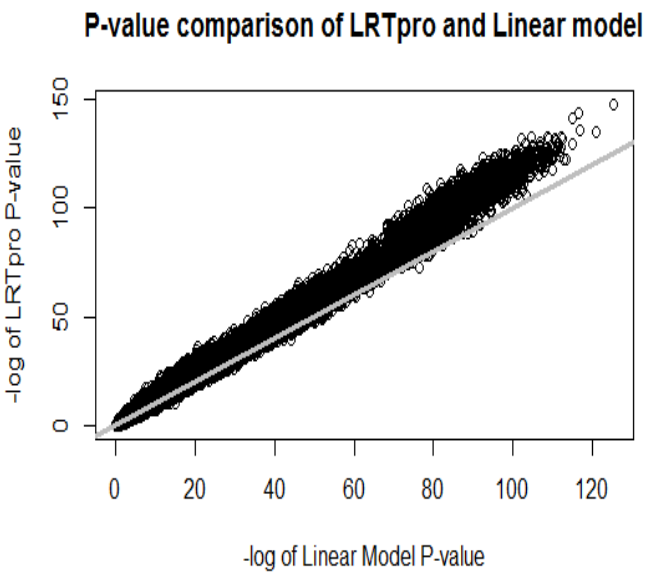


Figure 2. P-value comparison between LRT_{pro} and the linear model.

3.3. *Robustness of model selection*

Since choosing a restriction based on prior knowledge as is done in both LRT_{pro} and LRT_{add} may not be possible in every circumstance, it may not be necessary to choose the exact model. Table 4 shows that LRT_{pro} and LRT_{add} were the most powerful tests even when the other model was simulated. These two restrictions are of similar patterns, but the increase of power is substantial. Therefore, choosing a model at least similar to the true model can increase the power of the test.

3.4. *Parameter estimation*

In order to conduct the LRT, estimates of the underlying parameters of the two-component distribution are obtained. Table 5 illustrates the accuracy and precision of the resulting estimates across a range of simulation settings for the LRT_{pro} approach, with full results for all tests in supplemental tables 2 and 3. In general, LRT_{pro} and LRT_{add} yielded unbiased and accurate estimates across settings. In Table 5, one can see that LRT_{pro} accurately predicted the means of the components both across a wide range of settings and with low variation of the estimate. LRT_{pro} estimated well even when the data was simulated from the additive model. Similar results are obtained when estimating the mixing proportion (see Table 6) and the standard deviation of the components (see supplemental table 4).

Table 5. Estimates of Means for LRT_{pro}

True model	True p_{01}	q	μ_1	Standard deviation of μ_1	μ_2	Standard deviation of μ_2
Add	0.75	0.1	0.0005	0.02936	1.0022	0.0401
	0.9	0.1	-0.0021	0.0240	1.0036	0.0740
	0.75	0.2	-0.0007	0.0240	1.0011	0.0349
	0.9	0.2	-0.0020	0.0206	1.0005	0.0547
Pro	0.75	0.75	0.0005	0.02678	1.0005	0.0356
	0.9	0.75	-0.0014	0.0110	1.0008	0.0518
	0.75	0.9	0.0002	0.0293	1.0030	0.0400
	0.9	0.9	-0.0025	0.02496	1.0028	0.0781

Estimates aggregated across all settings with these parameters and all simulations within each setting, with the true value of $\mu_1 = 0$ and $\mu_2 = 1$.

Table 6 Estimates of Mixing Proportions for LRT_{pro}

model	True p_{01}	q	p_{01}	sd	True p_{11}	p_{11}	sd	True p_{21}	p_{21}	sd
Add	0.75	0.1	0.7509	0.0280	0.65	0.5197	0.1537	0.55	0.5290	0.0625
	0.9	0.1	0.8973	0.0265	0.8	0.6134	0.2786	0.7	0.6059	0.1627
	0.75	0.2	0.7496	0.0247	0.55	0.5097	0.0582	0.35	0.5072	0.1731
	0.9	0.2	0.8985	0.0220	0.7	0.6559	0.1220	0.5	0.5523	0.0744
Pro	0.75	0.75	0.7504	0.0248	0.5625	0.5390	0.0597	0.4219	0.4707	0.1188
	0.9	0.75	0.8983	0.0205	0.6750	0.6627	0.0691	0.5063	0.5139	0.0677
	0.75	0.9	0.7507	0.0284	0.6750	0.5385	0.1754	0.6075	0.5358	0.1037
	0.9	0.9	0.8966	0.0278	0.8100	0.6290	0.2823	0.729	0.6170	0.1814

Estimates aggregated across all settings with these parameters and all simulations within each setting.

3.5. *Real data results*

After analyzing 20321 SNPs on Chromosome 11 in relation to the AA/DGLA ratio, the LRT_{pro} test identified 28 SNPs as significantly associated after applying a Bonferonni multiple testing correction. These 28 SNPs came from 5 different regions on chromosome 11, all of which validated previous GWAS findings. Nineteen significant SNPs are in the well documented [10]–[12]FADS region (bp = 61622896- 61978819). Genes in this region that contain significant SNPs include DAGLA, MYRF, FADS1, FADS2, FADS3, and RAB3IL1 all of which have strong biological basis for desaturation activity [10].

Table 7. Most significant SNPs in each region

rs#	# of SNPs	MAF	Pos	Gene	LRT_{pro} p-value	p_{01}	p_{11}	p_{21}	μ_1	μ_2	σ
rs10751124	1	0.346	85432084	DLG2	$2.50\text{x}10^{-8}$	0.062	0.114	0.162	0.174	0.100	0.023
rs11220658	1	0.350	99618283	CNTN5	$4.52\text{x}10^{-7}$	0.110	0.075	0.051	0.179	0.101	0.024
rs7129015	5	0.198	110772485		$1.86\text{x}10^{-7}$	0.105	0.059	0.034	0.179	0.101	0.024
rs11217753	1	0.167	120181415		$2.94\text{x}10^{-9}$	0.108	0.052	0.025	0.180	0.101	0.024
rs174549	19	0.290	61803910	FADS1	$5.32\text{x}10^{-312}$	0.036	0.183	0.937	0.160	0.097	0.024

As an example interpretation of the results in Table 7, we first note that the significant tests all show similar estimates of the two components of the AA/DGLA ratio (mean of component one between 0.16 and 0.18; mean of component two between 0.097 and 0.101; SD of each component between 0.023 and 0.024). When an individual is genotyped and is the common homozygote at rs174549, they have a 3.6% chance of having their AA/DGLA ratio in the first component. However, if the individual has one less common allele, his chance increases to 18.3%, and with a second copy of the minor allele, it will increase to 93.7%.

4. Discussion

GWAS typically utilize linear models, thus making an assumption about the underlying normality of the data. When data is not normal, a Gaussian mixture distribution may represent a statistically justified and biologically interpretable model of the data. We proposed a constrained likelihood ratio test, which across many simulation settings, was more powerful than the standard linear model and gave accurate parameter estimates. When applied to a real dataset, the method identified biologically relevant SNPs in the well understood FADS region, along with parameter estimates to aid in biological interpretability of the impact of the SNP.

The general LRT framework proposed here shows reasonably good performance compared to the additive linear model, but can be improved upon by further constraining the model and 'saving' a degree of freedom. Our simulations suggest relatively robust performance of the constrained methods (LRT_{pro} and LRT_{add}) to misspecification of the true model though additional simulations across a wider range of misspecifications are needed.

We note that, due to the use of the EM algorithm to generate parameter estimates for use in the LRT, computational time for our proposed methods (3 minutes per test on a single processor with a sample size of 10,000) are much greater than that of the traditional linear model. Nevertheless, with the increasing computational power and the limited number of high minor allele frequency SNPs, it is plausible to run GWAS with this method and is a reasonable option for candidate gene approaches. Further work is necessary to investigate potential areas of computational improvement.

Numerous areas of future work and extension are possible. First, extensions of this work are needed to incorporate covariates and family structure into the method. Standard methods (e.g., first modeling the phenotype by covariates and/or family structure and then modeling the residuals) make normality assumptions and, so, may not be optimal candidates for extension in this Gaussian mixture modeling framework. Imputed data often provides dosages instead of discrete genotypes. Work is needed to extend this framework to allow for dosages in this testing framework. Further applications to genome wide data is necessary to fully understand the impact of this new method. Finally, extensions for multiple-marker testing and relaxing the equal variance assumption are also targets for further exploration.

We have developed a likelihood ratio test that analyzes the differences in mixing proportions between genotypes. The method and null distribution were validated through simulation. There was notable power increase over the more commonly used linear model, especially when we further increased power by restricting the model to incorporate prior biological belief. We have shown that this method is able to accurately predict model parameters. The model was applied to real data, and it replicated many previous findings while also providing more interpretable results. Further work is necessary to apply the model to a wider range of real metabolomics data and to investigate extensions of the model to handle covariates and imputed genotypes.

Supplemental files and R code

All supplemental material can be found at http://homepages.dordt.edu/ntintle/mixture_test.zip.

Acknowledgements

This research was funded by National Institutes of Health (NIH) 2R15HG006915. We thank and acknowledge Hope College and Dordt College for access to the computing clusters.

References

[1] P. M. Visscher *et al.*, "10 Years of GWAS Discovery: Biology, Function, and Translation.," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, Jul. 2017.

[2] N. Tintle, J. W. Newman, and G. C. Shearer, "A novel approach to identify optimal metabotypes of elongase and desaturase activities in prevention of acute coronary syndrome," *Metabolomics*, vol. 11, no. 5, pp. 1327–1337, 2015.

[3] G. C. Shearer, J. V Pottala, J. A. Spertus, and W. S. Harris, "Red blood cell fatty acid patterns and acute coronary syndrome.," *PLoS One*, vol. 4, no. 5, p. e5444, 2009.

[4] D. C. Schwenke, J. P. Foreyt, E. R. 3rd Miller, R. S. Reeves, and M. Z. Vitolins, "Plasma concentrations of trans fatty acids in persons with type 2 diabetes between September 2002 and April 2004.," *Am. J. Clin. Nutr.*, vol. 97, no. 4, pp. 862–871, Apr. 2013.

[5] W. Kim, D. Gordon, J. Sebat, K. Q. Ye, and S. J. Finch, "Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test.," *PLoS One*, vol. 3, no. 10, p. e3475, 2008.

[6] W. S. Harris, J. V Pottala, S. M. Lacey, R. S. Vasan, M. G. Larson, and S. J. Robins, "Clinical correlates and heritability of erythrocyte eicosapentaenoic and docosahexaenoic acid content in the Framingham Heart Study.," *Atherosclerosis*, vol. 225, no. 2, pp. 425–431, Dec. 2012.

[7] B. M. Psaty *et al.*, "Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts.," *Circ. Cardiovasc. Genet.*, vol. 2, no. 1, pp. 73–80, Feb. 2009.

[8] D. R. Govindaraju *et al.*, "Genetics of the Framingham Heart Study Population," *Adv. Genet.*, vol. 62, pp. 33–65, 2008.

[9] L. A. Cupples *et al.*, "The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports.," *BMC Med. Genet.*, vol. 8 Suppl 1, p. S1, Jan. 2007.

[10] N. Tintle *et al.*, "A genome wide assocaition study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart offspring study," *Prostaglandins. Leukot. Essent. Fatty Acids*, vol. 94, pp. 65–72, Mar. 2015.

[11] R. N. Lemaitre *et al.*, "Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium.," *PLoS Genet.*, vol. 7, no. 7, p. e1002193, Jul. 2011.

[12] K. Suhre *et al.*, "Human metabolic individuality in biomedical and pharmaceutical research.," *Nature*, vol. 477, no. 7362, pp. 54–60, Sep. 2011.

Reading Between the Genes: Computational Models to Discover Function from Noncoding DNA

Yves A. Lussier[†], Joanne Berghout, Francesca Vitali, Kenneth S. Ramos
Center for Biomedical Informatics and Biostatistics,
The Center for Applied Genetic and Genomic Medicine,
BIO5 Institute, UA Cancer Center, and Dept of Medicine; University of Arizona
1657 E Helen St, Tucson, AZ 85719, USA
Emails: yves@email.arizona.edu, jberghout@email.arizona.edu, francescavitali@email.arizona.edu, ksramos@email.arizona.edu

Maricel Kann
Dept of Biological Sciences; University of Maryland, Baltimore County,
1000 Hilltop Circle Baltimore, MD 21250 United States
Email: mkann@umbc.edu

Jason H. Moore
Department of Biostatistics, Epidemiology, and Informatics; University of Pennsylvania,
3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: jhmoore@exchange.upenn.edu

Noncoding DNA - once called "junk" has revealed itself to be full of function. Technology development has allowed researchers to gather genome-scale data pointing towards complex regulatory regions, expression and function of noncoding RNA genes, and conserved elements. Variation in these regions has been tied to variation in biological function and human disease. This PSB session tackles the problem of handling, analyzing and interpreting the data relating to variation in and interactions between noncoding regions through computational biology. We feature an invited speaker to how variation in transcription factor coding sequences impacts on sequence preference, along with submitted papers that span graph based methods, integrative analyses, machine learning, and dimension reduction to explore questions of basic biology, cancer, diabetes, and clinical relevance.

Keywords: non-coding DNA, intergenic, LncRNA, microRNA, sncRNA, miRNA, piRNA, LINEs, LINE1, repetitive elements

[†] Work partially supported by This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)

1. Introduction

The majority of the human genome is comprised of noncoding DNA. Estimating the percentage of noncoding DNA that comprises functional elements is somewhat controversial, depending heavily on definitions used, specific assays, and cell type or other biological contexts explored. However, we can all agree that it's not zero. Unlike protein-coding sequences where a biologically active function can be relatively readily assigned through standard experimental techniques, most noncoding sequences largely remain as functional black boxes. Even for those noncoding sequence variants confidently linked to variation in a biological trait (i.e. via GWAS), the mechanism of precisely how they exert an effect on the phenotype often remains unclear. Identifying the potential and most relevant relationships or functions in the absence of an a priori biological hypothesis requires the intersection of big data, computing, and creativity.

Noncoding DNA includes RNA genes (miRNA, lncRNA, piRNA), regulatory regions (transcription factor binding sites, eQTL-associated SNPs, promoters, enhancers, insulators), epigenetic mark associated regions, repetitive elements (LINEs, transposons, Alu elements, telomeres), pseudogenes, and structural elements among others. With this diversity of potential function, largely incomplete annotation, and substantial degree of sequence variation between individuals, defining even common canonical motifs at the resting state is challenging. Recognizing these needs, several international efforts such as ENCODE, GTEX, NIH Epigenomics Roadmap, and the International Epigenome Consortium have been established. These researchers use high throughput technologies and systematic approaches to start developing regulatory maps and begin learning about "genomic grammar", or, the rules that govern meaning as a cell or protein "reads" through DNA. With these international projects along with many other academic laboratories generating vast quantities of genome-scale data from sequence, expression, ChIP-Seq, CHIA-PET, ATAC-Seq and other new technologies [1], there also emerges a need to develop methods for appropriate data handling, integration, and analysis.

Thus, there arises a unique opportunity for computational biologists to identify network and systems properties of noncoding DNA, linking evidence from biochemical assays, genetics, and evolutionary biology with other datasets. These can predict downstream biology, functional convergence, and impacted mechanisms that ultimately lead to disease. In addition to novel basic science insights, understanding the mechanisms perturbed by variation in noncoding DNA can implicate new pathophysiological mediators and unveil new therapeutic targets. The computational tools that have been created by researchers to handle these data have been complex, innovative, and some great work has been done by data-generating scientists and research parasites alike, leading to unexpected discovery and new research questions. As a result, we believe that a PSB session devoted to the topic of non-coding DNA would be timely, interesting and yield some excellent paper submissions.

2. Sessions summary

2.1. *Invited Talk*

We have the privilege to have the participation of a guest speaker, Dr. Martha Bulyk from the Division of Genetics at Harvard, also an Associate Member of the Broad Institute. Dr Bulyk has made a name for herself with work at interface between DNA sequence and protein binding specificity to explore cis regulatory motifs. She invented a high throughput method for detecting transcription factor binding preferences via a novel protein binding microarray, and developed integrated computational approaches to interpret these data, with an eye towards TF binding site clustering, combinatorial co-occurrences, and cross-species conservation.

2.1.1. *Her talk is entitled "Survey of coding variation in human transcription factors reveals prevalent DNA binding changes"*

> Invited talk abstract: Sequencing of exomes and genomes has revealed abundant genetic variation affecting the coding sequences of human transcription factors (TFs), but the consequences of such variation remain largely unexplored. We developed a computational, structure-based approach to evaluate TF coding variants for their impact on DNA-binding activity and used universal protein binding microarrays (PBM) to assay sequence-specific DNA-binding activity across reference and variant alleles found in individuals of diverse ancestries and families with Mendelian diseases. We found variants that affect DNA-binding affinity or specificity and identified thousands of rare alleles likely to alter the DNA-binding activity of human sequence-specific TFs. Altered sequence preferences correlated with changes in genomic TF occupancy (ChIP-Seq peaks) and gene expression of the associated target genes. Our results suggest that most individuals have unique repertoires of TF DNA-binding activities, which may contribute to phenotypic variation.

2.2. *Papers*

2.2.1. *Evaluating relationships between pseudogenes and genes: from pseudogene evolution to their functional potentials*

Johnson et al. developed a novel approach integrating graph analysis, sequence alignment and functional analysis to classify pseudogene-gene relationships. They identified ~15,000 pseudogenes from the Human GENCODE release 24 using cufflinks gffread. Every pseudogene was then aligned to the best consensus sequence of ~3200 gene families using pairwise ClustalW. Finally, they mined the resulting network of pseudogene-gene edges jointly with gene-Gene Ontology terms to impute the former function of poorly characterized pseudogenes. These observations extend on previous work [2], and may lead to new insight on families of pseudogenes and to characterize the potential function of new pseudogenes in cancer.

2.2.2. *Convergent downstream candidate mechanisms of independent intergenic polymorphisms between co-classified diseases implicate epistasis among noncoding elements*

Han et al. characterize convergent downstream candidate mechanisms of distinct intergenic SNPs across distinct diseases within the same clinical classification by conducting an integrative analysis of four networks: disease class to disease annotations, GWAS disease-SNP associations, eQTL SNP-mRNA associations, and Gene Ontology (GO) gene-mechanisms annotations. At FDR≤5%, they prioritize 167 intergenic SNPs, 14 classes, 230 mRNAs, and 134 GO mechanisms. They expand on their previous study [3] and observe that co-classified SNP were more likely to be prioritized in the same mechanisms as compared to those of distinct classes (odds ratio ~3.8). The SNPs prioritized to the same GO mechanisms were also enriched in regions bound to the same/interacting transcription factors and/or interacting in long-range chromatin interactions suggestive of epistasis (odds ratio ~ 2,500). Such network of genetic epistasis associated to candidate biological mechanisms has the potential to reposition medications that target proteins within downstream mechanisms of intergenic SNPs associated to disease risks, the latter generally considered undruggable before this study.

2.2.3. *Pan-cancer analysis of expressed single nucleotide variants in long intergenic non-coding RNAs*

Ching et al. analyzed 6118 primary tumor samples of 12 distinct cancers from TCGA and detected 94,700 somatic mutations in lincRNAs, and 15.5 million germline variants in lincRNAs. They further built machine-learning models to impute sensitive regions to mutations and variants of lincRNAs, suggesting that some nucleotide positions within lincRNA are more likely to gain somatic mutations than other positions. Non-coding but not trivial, the authors extend our understanding of polymorphisms and mutations in lincRNAs, increasing their contributions to "non-coding but not trivial" lincRNAs [4].

2.2.4. *Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 diabetes GWAS*

Manduchi et al. utilized evidence for enhancer-promoter interactions from functional genomics data in order to build biological filters to narrow down the search space for two-way Single Nucleotide Polymorphism (SNP) interactions in Type 2 Diabetes (T2D) Genome Wide Association Studies (GWAS). They confirmed the validity of the method by identifying a statistically significant pairs of type 2 diabetes SNPs in statistical epistasis [5] and replicated in an independent datasets. Their framework, that accounts for epistasis, expands substantially on GWAS analyses as the current paradigm consists on linear additive analyses of GWAS.

References

1. Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease**.** *Nat Biotechnol* 2012, **30:**1095-1106.

2. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JM, Li Y, Menzies A, Mudie L, Ramakrishna M, et al: Processed pseudogenes acquired somatically during cancer development. *Nat Commun* 2014, **5:**3644.
3. Li H, Achour I, Bastarache L, Berghout J, Gardeux V, Li J, Lee Y, Pesce L, Yang X, Ramos KS, et al: Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions**.** *NPJ Genom Med* 2016, **1**.
4. Ching T, Masaki J, Weirather J, Garmire LX: Non-coding yet non-trivial: a review on the computational genomics of lincRNAs**.** *BioData Min* 2015, **8:**44.
5. Moore JH, Williams SM: Epistasis and its implications for personal genetics**.** *Am J Hum Genet* 2009, **85:**309-320.

Pan-cancer analysis of expressed somatic nucleotide variants in long intergenic non-coding RNA

Travers Ching[1,2], Lana X. Garmire[1,2]
[1]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa
Honolulu, HI 96822, USA
[2]Epidemiology Program, University of Hawaii Cancer Center
Honolulu, HI 96813, USA

Long intergenic non-coding RNAs have been shown to play important roles in cancer. However, because lincRNAs are a relatively new class of RNAs compared to protein-coding mRNAs, the mutational landscape of lincRNAs has not been as extensively studied. Here we characterize expressed somatic nucleotide variants within lincRNAs using 12 cancer RNA-Seq datasets in TCGA. We build machine-learning models to discriminate somatic variants from germline variants within lincRNA regions (AUC 0.987). We build another model to differentiate lincRNA somatic mutations from background regions (AUC 0.72) and find several molecular features that are strongly associated with lincRNA mutations, including copy number variation, conservation, substitution type and histone marker features.

1. Introduction

Long intergenic non-coding RNAs (lincRNAs) have been shown to play important roles in many diseases, including cancer. The expression of thousands of lincRNAs are deregulated in cancer, and many lincRNAs have been proposed as biomarkers for tumor tissues and patient prognosis [1]–[3]. There is also strong evidence that lincRNAs may serve as drivers of tumorigenesis, cause drug resistance or cause metastasis [4]–[7]. Mutations in cancer driver genes lead to a series of downstream events, including gene expression changes [8], [9]. However, because lincRNAs are a relatively new class of non-coding RNAs compared to protein coding mRNAs, the mutational landscape of lincRNAs and their impact on gene expression, have not been extensively studied.

While most people use exome-Seq to investigate somatic mutations, the coverage on lincRNA regions is very limited. Furthermore, several previous studies have shown that expressed somatic nucleotide variations (eSNVs) can be robustly called from RNA-Seq data [10]–[12]. Therefore, to interrogate the effects of lincRNA mutations, we used the RNA-Seq data from The Cancer Genome Atlas (TCGA), analyzing 6118 patient samples from 12 cancer datasets.

Due to the fact that most RNA-Seq samples do not contain normal controls, we constructed a Random Forest model on exome-Seq data to differentiate eSNVs and germline variants, and then extrapolated this model to the RNA-Seq eSNVs. Subsequently, we interrogated the features related to eSNVs within lincRNAs. We find several molecular features that are strongly associated with

lincRNA mutations, including copy number variation, conservation, substitution type and histone marker features.

2. Methods

2.1. *TCGA Datasets*

We used 12 cancer datasets from TCGA with a total of 6118 primary tumor samples in this study. These datasets include bladder urothelial carcinoma (BLCA, n=406), breast invasive carcinoma (BRCA, n=1084), head and neck squamous cell carcinoma (HSNC, n=514), kidney renal clear cell carcinoma (KIRC, n=525), liver hepatocellular carcinoma (LIHC, n=364), low grade glioma (LGG, n=513), lung adenocarcinoma (LUAD, n=512), lung squamous cell carcinoma (LUSC, n=498), ovarian serous cystadenocarcinoma (OV, n=300), stomach adenocarcinoma (STAD, n=414), prostate adenocarcinoma (PRAD, n=491) and thyroid carcinoma (THCA, n=497). RNA-Seq fastq files were downloaded using GeneTorrent program from the UCSC Cancer Genomics Hub (https://cghub.ucsc.edu). Additional TCGA samples were downloaded from NCBI Genomic Data Commons Data Portal (https://gdc-portal.nci.nih.gov) using the GDC data transfer tool.

2.2. *Predicting germline and somatic mutations*

The exome sequencing variant calls, including somatic and germline variants, were downloaded for 7 TCGA datasets (BLCA, HNSC, KIRC, LGG, LIHC, LUAD, PRAD and STAD). A Random Forest model was built to classify somatic vs. germline variants, from the exome sequencing data from TCGA. In this model, the class labels were derived as 1 – somatic mutation and 0 – germline mutation, determined by the paired exome-seq data.

The Xgboost package in R was used (version 0.6-0) with 1000 trees. Five features were used in the building of this model: mutation frequency across the entire cohort (frequency), dbsnp (whether an SNV occurred at a position annotated by the dbSNP database), fa.tumor (the estimate allele ratio of the SNV in the tumor exome sample), conservation (PhyloP conservation score from the UCSC genome browser) and transversion (whether the SNV was a transition or transversion mutation). These features were chosen in order to be independent of the subsequent models. The performance was evaluated using 5-fold cross-validation.

2.3. *Expressed somatic nucleotide variations (eSNVs)*

Raw read data were downloaded from UCSC Cancer Genomics Hub in the fastq format. Reads were first aligned to the hg19 genome reference using STAR aligner [13] in two-pass mode. Aligned BAM files were sorted using ReorderSam function in Picard-tools (http://broadinstitute.github.io/picard) and reads were split based on splicing junctions using

SplitNCigarReads function in Genome Analysis Toolkit (GATK)[14]. Reads were then processed through duplicate removal, INDEL realignment and base recalibration, following standard protocols. Variant calling was performed using GATK's Haplotype caller. Data processing was performed on the high performance computing cluster of University of Hawaii. To further reduce potential false positive calls, variants were filtered based on SNV clusters and read strand bias following recommendations from the developers. To identify lincRNA specific eSNVs, variants associated with lincRNAs based on the lncipedia 4.0 reference [6] were used for analysis.

2.4. *Predictive models to classify eSNVs from background nucleotide sites*

We constructed classification models in order to predict eSNVs from germline variants for each cancer type. The class labels for this model were 1 - a eSNVs determined in lincRNA regions from the RNA-Seq data (based on the results of the first model) and 0 - background "negative" eSNVs, i.e., random non-mutated locations on expressed lincRNAs in each RNA-Seq sample. The models were built on balanced datasets. The molecular features in these models include conservation, copy number variation, histone marker features, nucleotide composition features, location on exon junctions and transcription start and end sites. Three algorithms were employed on these datasets: logistic regression with ridge regularization (LR), a fast linear classification algorithm using the glmnet R package (version 2.0-5); a neural network classifier using Tensorflow (version 1.1.0); and Gradient Boosted Trees [15], a fast non-linear tree-based classifier using the xgboost R package (version 0.6-0).

To evaluate each model, the datasets were split into 80% training and 20% testing. AUC was calculated as the performance metric on the testing sets. The Gradient Boosted Trees models were evaluated and the Gain value of each feature was computed, to determine feature importance. In an ensemble forest model (Random Forest or Gradient Boosted Trees), Gain is the average improvement of performance of the model on each tree branch, split by the features in the ensemble forest [15].

3. Results

3.1. *Computational pipeline accurately predicts genetic variation in tumor RNA-Seq samples*

We selected 6118 primary tumor RNA-Seq samples from 12 TCGA datasets and implemented a pipeline for calling mutations from bulk RNA-Seq data described in the methods section (Figure 1). To verify the quality of the results, we compared the variant calls from exome sequencing in paired exome and RNA-Seq sample datasets. On average, 80% of the expressed somatic nucleotide variants (eSNVs) found in RNA-Seq data were also found in the exome sequencing variant calls, within the exome-seq read regions. This high concordance of eSNVs detected by RNA-Seq relative to exome-seq is better than what others showed for the same samples using different analysis platforms (~50%) [16], suggesting that our eSNV calls from the RNA-Seq are reliable.

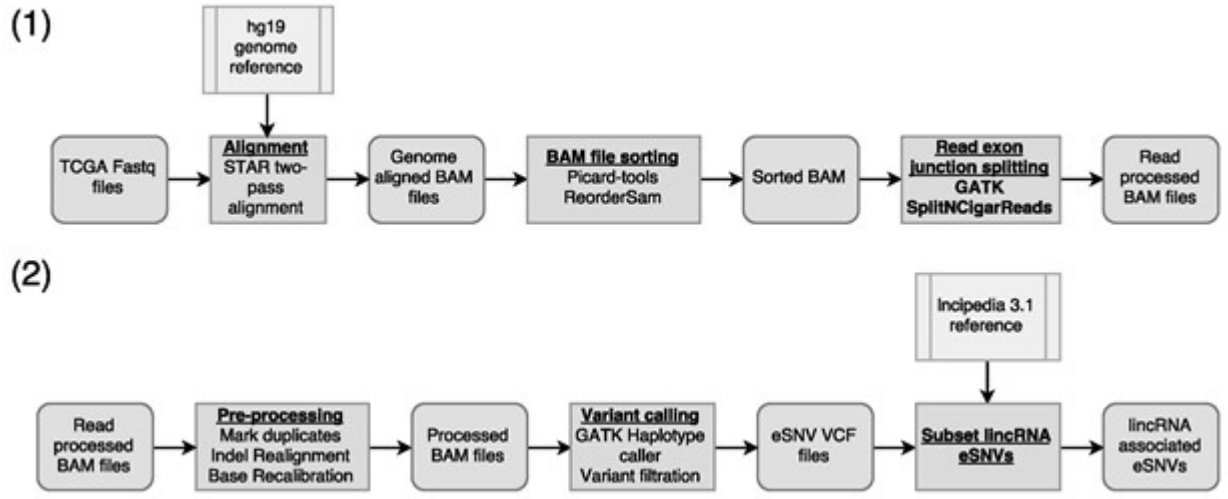


Figure 1. Workflow describing the pipeline for the raw data processing, calling eSNVs and calculating expression data. (1) RNA-Seq alignment and read pre-processing. (2) Re-alignment, re-calibration and lincRNA variant calling.

3.2. *A Random Forest model differentiates somatic and germline mutations*

Because RNA-Seq samples, from which we called eSNVs, usually do not have matched normal samples, directly determining somatic or germline status of a variant through these RNA-Seq samples is not possible. However, since most eSNVs from RNA-Seq overlap with the SNVs detected through exome sequencing in the protein-coding genes, we then aimed to predict the somatic or germline origin of these variants in RNA-Seq using a Random Forest model trained on the exome-seq data. Exome-sequencing data are preferred "gold-standard" training data, as these data had paired normal and tumor samples (and therefore SNVs could be accurately differentiated from germline variants). We built a random forest model classifying the somatic mutations versus the germline mutations based upon five features: frequency (mutation frequency across the samples in a dataset), dbsnp (whether the mutation is documented in the NCBI dbSNP database), FA.tumor (the fraction of the alternate allele in the tumor sample), conservation (PhyloP conservation score) and transversion (whether the mutation is a transversion or a transition mutation). This model had an AUC of 0.988 on the exome sequencing data and an AUC of 0.987 based on RNA-Seq data respectively (Figure 2A). By comparison, the logistic model had slightly lower AUCs of 0.979 and 0.985. We therefore decided to use the results of random forest model for the following sections. Mutation frequency, dbsnp and FA.tumor features have relatively high importance scores relevant to the outcome, with values of

0.485, 0.291, and 0.184 (Figure 2B). Conservation and transversion do not present as important features in this model (Figure 2B).

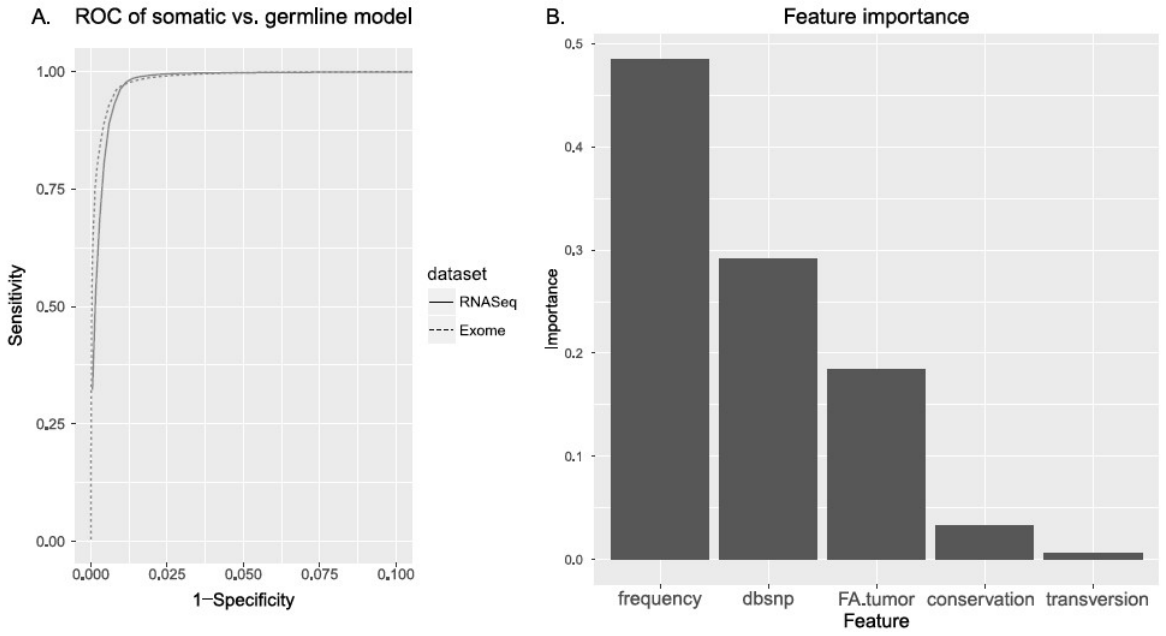


Figure 2. A – Receiver operating characteristic (ROC) curve showing the performance of the random forest models for differentiating somatic and germline mutations in the exome-sequencing data. B – Feature importance based on the random forest model.

Secondly, we then applied this model to the 12 RNA-Seq datasets, and selected eSNVs that are highly confident as either somatic (posterior probability > 0.97) or germline variants (posterior probability < 0.03). Using these thresholds, 1.25 million somatic mutations were detected in protein-coding genes and 94,700 were detected in lincRNAs. For germline variants, 170 million protein coding variants were detected and 15.5 million lincRNA variants were detected. We calculated the density of lincRNAs genome-wide, relative to the lincRNA exon density. There are many regions of enriched lincRNA eSNVs throughout the genome (Figure 3).

There are some regions that have an increased frequency of lincRNA eSNVs. The top four regions included chr2p11.2, chr14q32.33, chr22q11.22 and chr3q29. In particular, chr2p11.2 is known to be heavily associated with breast cancer [17]. Sahin et al. found that copy number imbalances in chr2p11.2 had a significant effect on breast screening and detection. They also found that the imbalance had a significant effect on disease free survival. However, they were not able to

determine any association with protein coding genes. These results suggest that the association of this region with cancer phenotypes could be due to lincRNA mutations.

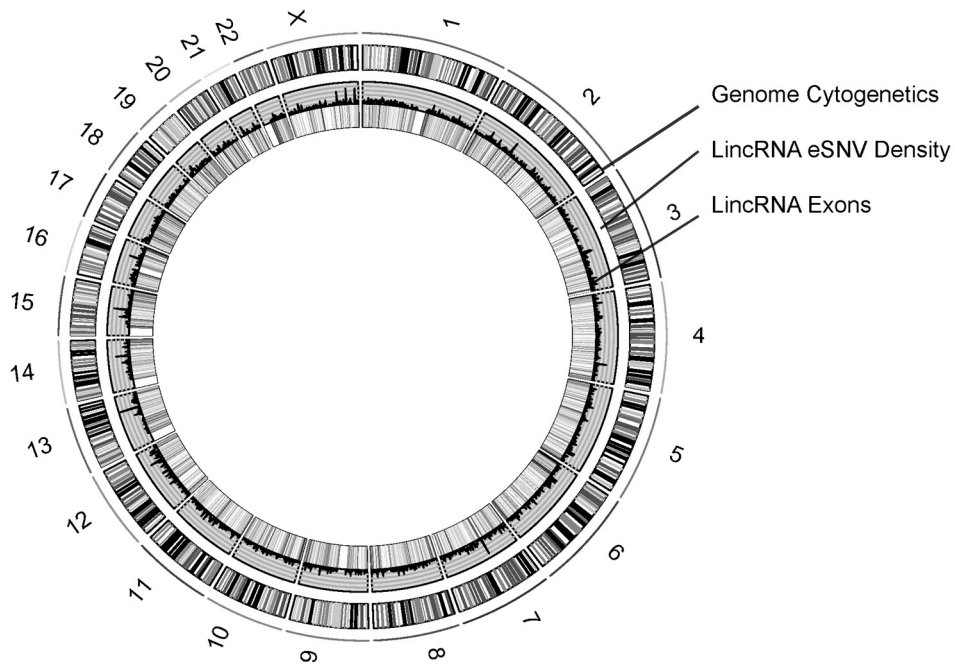


Figure 3. Circos plot of the normalized lincRNA eSNV mutation density. For each cancer type, the number of lincRNA eSNVs was binned using window sizes of 100,000 across the genome. The bin count was then normalized by the lincRNA exon density in the corresponding transcriptomic region. The outer layer shows the human genome cytogenetics. The middle layer shows the eSNV density within lincRNA regions. The inner layer shows the exon density of the lincRNA transcriptome.

3.3. *A machine learning model predicts mutation likelihood on nucleotide positions*

Next, we wanted to determine which nucleotide positions were more likely to have somatic mutations. For each of the 12 TCGA cancer types, we constructed a classification model to predict eSNV likelihood within lincRNAs. Similarly, we also built models using the germline variants using the same features. We applied three machine learning algorithms to each dataset: logistic regression (a linear classifier), a neural networks (a flexible non-linear classifier) and gradient boosted trees (a fast tree-based non-linear classifier). In each dataset, the boosted trees model performed considerably better than the neural network and logistic regression models. The neural network models generally performed better than the logistic regression. Across all 12 TCGA datasets, Boosted Trees had an

average AUC of approximately 0.72 for eSNVs and 0.89 for germline variants (Figure 4). By comparison, the logistic regression models had AUCs of 0.68 and 0.77 for eSNVs and germline variants respectively.

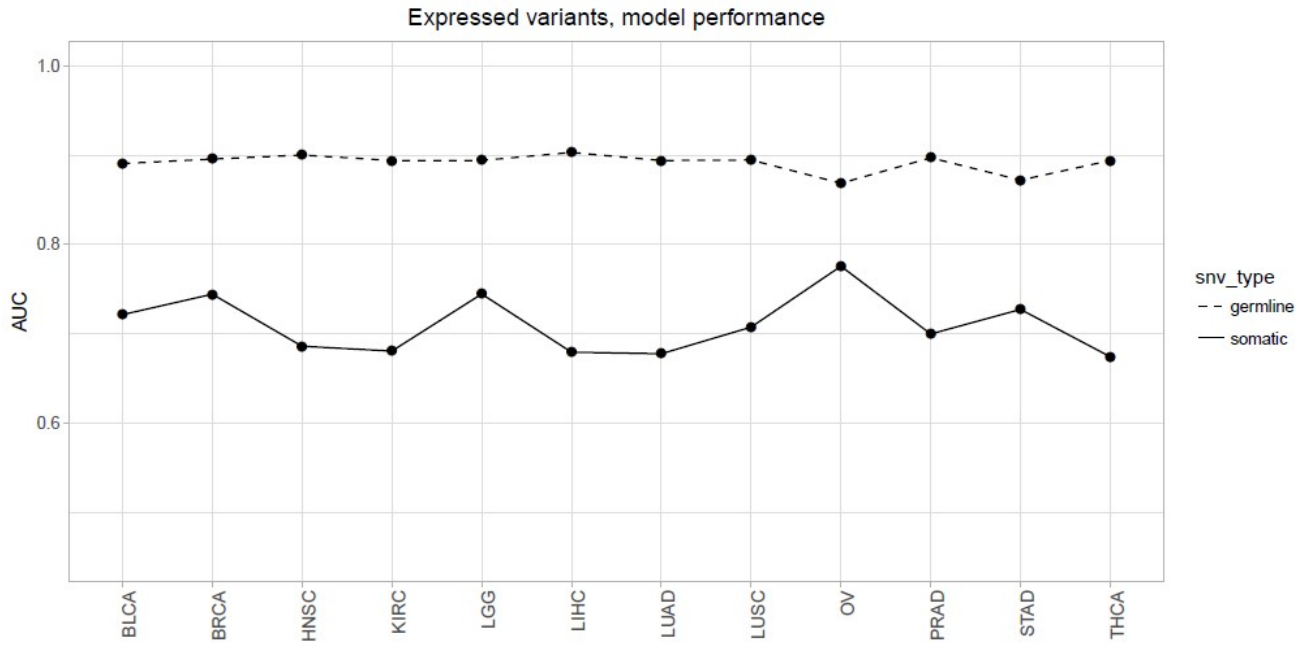


Figure 4. Model performance differentiating somatic eSNVs and germline variants from background in lincRNAs for the 12 TCGA datasets, using the Gradient boosted Trees machine learning algorithm.

3.4. *Molecular features correlated with somatic eSNVs differ from germline variants*

To evaluate the importance of each feature in the two models (somatic vs. germline), we used the Gain measure, which calculates the average increase in performance for each feature in every tree in the Boosted Trees ensemble. For the 12 lincRNA eSNV models, conservation followed by copy number variation (cnv_pos and cnv_promoter) are the most important features (Figure 5A). For the germline variant models, copy number variation does not have a high feature importance score (Figure 5B). For the eSNV models, on average, the third most important feature is tranversion – the type of mutation.

Several histone features show importance in specific datasets (Figure 5A and 5B). We measured histone methylation levels at two locations: the promoter regions of each lincRNA and the position of the eSNV. Promoter methylation signatures are relatively less important than methylation signatures at the eSNV position. For kidney renal cell carcinoma and prostate cancer, H3k04me3 (histone 3 trimethylation signature) position information is the most important histone modification feature.

H3k4me3 and H3k36me3 histone methylation are both important for liver cancer. In addition, nucleotide composition upstream or downstream of the eSNV are not as important as the nucleotide mutation site, with C/G nucleotides being much less likely to be mutated. Exon junction and transcription start and stop sites features (TSS and TES) are among the least important features, suggesting that there was neither enrichment nor depletion of eSNVs on splice junctions and the two ends of lincRNA transcripts.

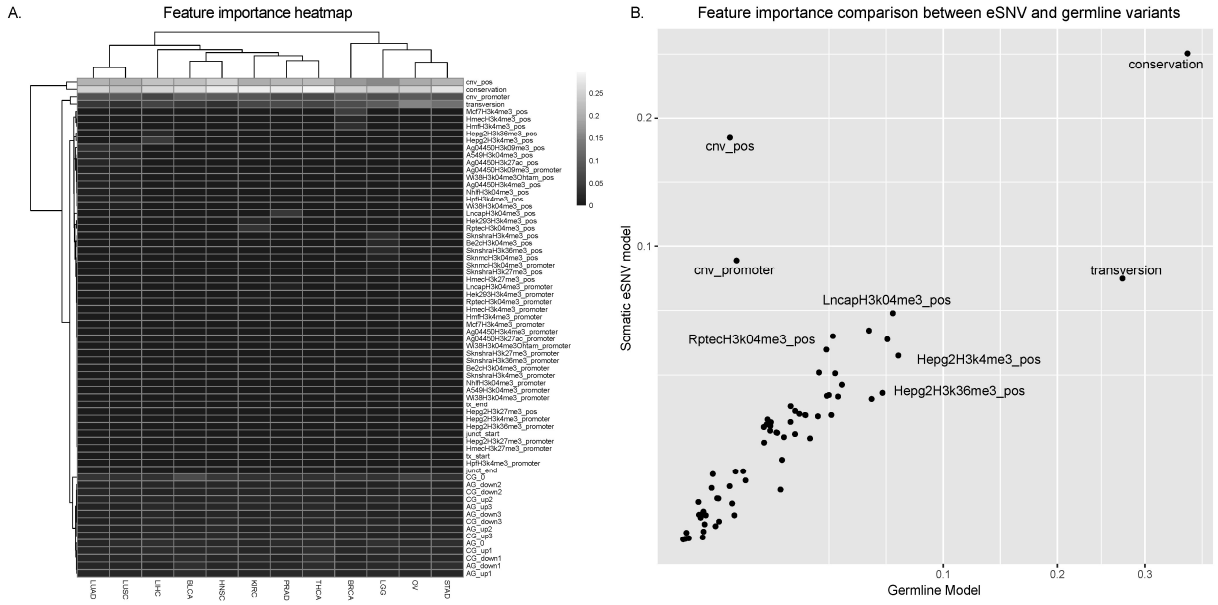


Figure 5. A. Heatmap of feature importance for lincRNA eSNVs. Feature importance was calculated using the Gain measure, which evaluates the average increase in accuracy at a feature's node splitting in each tree. B. Comparison of feature importance between lincRNA eSNVs and germline variants.

4. Discussion

In this study, we show that machine learning models can accurately separate a portion of highly confident eSNVs from germline variants, using exome sequencing data for training. Since paired normal samples are often not available (as in the case of older FFPE DNA samples [18], or RNA-Seq expression samples in the present case) this investigation has wide applications. Based on the random forest model (Figure 2), mutation frequency, dbsnp and allele frequency are important features in predicting which variants are somatic. dbSNP variants are those commonly found in population germlines, and therefore are much less likely to be somatic.

Similarly, it has been noted that variants that have low allele frequency are likely to be cancer mutations, and may even play important roles in cancer development [19]. Thus, variants found to

have 100% allele frequency in the tumor samples are unlikely to be somatic mutations, as normal sample contamination is usually present [20]. Furthermore, even if normal sample contamination were removed, tumor samples often contain multiple populations that may have different alleles and mutational profiles [21]. Therefore, it is unlikely for a somatic mutation to have an allele frequency of 1.

The models predicting eSNVs from the background nucleotide positions showed strong performance (Figure 4), suggesting that some nucleotide positions within lincRNA are more likely to gain somatic mutations than other positions. Comparing the different classification algorithms, the logistic regression performed worse than the non-linear Boosted Trees algorithm, suggesting that that the prediction of lincRNA may be complex and non-linear.

Interestingly, conservation is the most important feature in the lincRNA somatic model. Although conservation scores are determined through evolutionary homology, it has been shown that conservation correlates with somatic mutation hot spots [22]. The germline models for lincRNAs, in contrast, scored conservation slightly higher. This may be expected, as conservation itself is a direct measure of the likelihood of variation through a species' germline lineage.

The second most important feature for most lincRNA somatic models was cnv_pos, followed by cnv_promoter (i.e., copy number variation at the mutation position and promoter, determined by a microarray on a corresponding DNA TCGA sample). Previous studies have found that many somatic gene mutations are significantly correlated with copy number alterations in cancer, including EGFR and KRAS [23]. However, although many genes were found to be correlated, on a global scale, many genes did not reach significance [23]. As may be expected, copy number variation was much more important in the somatic eSNV model, compared to the germline model, as copy number variations themselves are somatic alterations, and should not alter the original germline genomic state.

The next most important feature for the lincRNA somatic model was transversion (whether a mutation was a transversion – 1, or transition mutation – 0). Transition somatic mutations, particularly C>T transitions, are more frequent than transversion somatic mutations [24].

However, for particular tumor types and even specific genes (e.g., p53 somatic mutations), the prevalence of transversions may be higher than transitions [24], [25]. This suggests the type of mutation may potentially be important in determining a mutation's biological importance.

For the datasets with matched tissue cell line histone data, histone features related to the lincRNA sites were determined to have a significant effect on the prediction of eSNVs sites. Previous studies have found that chromatin modifications had a major effect on regional mutation rates in cancer cells [26]. Since histone methylation and acetylation status determines the 3-dimensional conformation and openness of genomic regions, differences in histone modifications between regions may change the exposure of a region to mutagenic forces in a tumor.

While using RNA-Seq to perform mutation calling is an interesting idea to couple SNVs with expression data, false negatives may arise due to the fact that many lincRNAs and transcripts are lowly expressed or not expressed at all in certain tissues or conditions [27]. On the other hand, false

positives may also be introduced as RNA splicing of transcripts could cause additional read misalignment to the genome reference [12].

Additionally, since expression data and eSNVs both come from RNA-Seq and require the presence of expressed transcripts to produce reads for measurement, expression and eSNVs are inherently coupled, at a technical level. A gene that is not expressed will also not have any detected mutations. This suggests that there may be bias towards regions of high read coverage and therefore high expression.

However, within the TCGA RNA-Seq datasets, the majority of eSNVs detected that lie within exome probe boundaries, are also detected in exome-sequencing variant calling from the same patients. Previous studies have found that, from the same patient, the concordance between sequencing platforms and variant calling software to be about 50% [16]. This suggests that the false positives from the eSNV RNA-Seq pipeline are much less of an issue than other technical factors, such as the choice of sequencing platform.

The sparsity of SNPs and SNVs in a genome suggests that individual sites may not be able to be definitively predicted with high certainty. Biologically, this is a result of the stochastic nature of somatic point mutations. However, individual genes, lincRNAs, genomic regions, or possibly individual exons or sections of lincRNAs may be predicted as more or less likely to be mutated, relative to other exons or genes.

5. Conclusion

In this study, we generated two types of models: first, a Random Forest model to differentiate germline and somatic mutations, and second, a Gradient Boosted Trees model that finds lincRNAs nucleotide positions that are more likely to contain mutations. Additionally, we have explored the eSNV landscape and found regions across the genome that have an increase in lincRNA mutations, such as chr2p11.2. This is an important step in finding the biological significance of lincRNAs that are susceptible to somatic mutations in cancer.

6. References

[1] T. Ching *et al.*, “Pan-Cancer Analyses Reveal Long Intergenic Non-Coding RNAs Relevant to Tumor Diagnosis, Subtyping and Prognosis,” *EBioMedicine*, 2016.

[2] T. Ching, J. Masaki, J. Weirather, and L. X. Garmire, “Non-coding yet non-trivial: a review on the computational genomics of lincRNAs,” *BioData Min.*, vol. 8, no. 1, p. 1, 2015.

[3] J. R. Prensner and A. M. Chinnaiyan, “The emergence of lncRNAs in cancer biology,” *Cancer Discov.*, vol. 1, no. 5, pp. 391–407, 2011.

[4] R. Zarate, V. Boni, E. Bandres, and J. Garcia-Foncillas, “MiRNAs and LincRNAs: Could They Be Considered as Biomarkers in Colorectal Cancer?,” *Int. J. Mol. Sci.*, vol. 13, no. 1, pp. 840–865, Jan. 2012.

[5] X. Zhou, J. Chen, and W. Tang, “The molecular mechanism of HOTAIR in tumorigenesis, metastasis, and drug resistance,” *Acta Biochim. Biophys. Sin.*, vol. 46, no. 12, pp. 1011–1015, Dec. 2014.

[6] Y. Yang, H. Li, S. Hou, B. Hu, J. Liu, and J. Wang, "The Noncoding RNA Expression Profile and the Effect of lncRNA AK126698 on Cisplatin Resistance in Non-Small-Cell Lung Cancer Cell," *PLOS ONE*, vol. 8, no. 5, May 2013.
[7] A. Lanzós *et al.*, "Discovery of Cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features," *Sci. Rep.*, vol. 7, p. 41544, 2017.
[8] A. Gonzalez-Perez *et al.*, "IntOGen-mutations identifies cancer drivers across tumor types," *Nat. Methods*, vol. 10, no. 11, pp. 1081–1082, 2013.
[9] D. Tamborero *et al.*, "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Sci. Rep.*, vol. 3, Oct. 2013.
[10] Z. Peng *et al.*, "Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome," *Nat. Biotechnol.*, vol. 30, no. 3, pp. 253–260, 2012.
[11] R. Piskol, G. Ramaswami, and J. B. Li, "Reliable identification of genomic variants from RNA-seq data," *Am. J. Hum. Genet.*, vol. 93, no. 4, pp. 641–651, 2013.
[12] X. Tang *et al.*, "The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data," *Nucleic Acids Res.*, p. gku1005, 2014.
[13] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.
[14] A. McKenna *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, 2010.
[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[16] J. O'Rawe *et al.*, "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing," *Genome Med.*, vol. 5, no. 3, p. 1, 2013.
[17] A. A. Sahin, M. E. Edgerton, J. L. Murray, and M. Bondy, "Copy Number Imbalances between Screen- and Symptom-Detected Breast Cancers and Impact on Disease-Free Survival," 2011.
[18] F. Meric-Bernstam *et al.*, "A Decision Support Framework for Genomically Informed Investigational Cancer Therapy," *J. Natl. Cancer Inst.*, vol. 107, no. 7, p. djv098, Jul. 2015.
[19] M. Costello *et al.*, "Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation," *Nucleic Acids Res.*, p. gks1443, 2013.
[20] K. Cibulskis *et al.*, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nat. Biotechnol.*, vol. 31, no. 3, pp. 213–219, 2013.
[21] G. H. Heppner, D. L. Dexter, T. DeNucci, F. R. Miller, and P. Calabresi, "Heterogeneity in drug sensitivity among tumor cell subpopulations of a single mammary tumor," *Cancer Res.*, vol. 38, no. 11 Part 1, pp. 3758–3763, 1978.
[22] R. Walker *et al.*, "Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features," *Oncogene*, vol. 18, no. 1, pp. 211–218, 1999.
[23] L. Ding *et al.*, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, 2008.
[24] C. Kandoth *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, 2013.
[25] M. Hollstein, D. Sidransky, B. Vogelstein, and C. C. Harris, "p53 mutations in human cancers," *Science*, vol. 253, no. 5015, pp. 49–54, 1991.
[26] B. Schuster-Böckler and B. Lehner, "Chromatin organization is a major influence on regional mutation rates in human cancer cells," *nature*, vol. 488, no. 7412, p. 504, 2012.

[27] M. N. Cabili *et al.*, “Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses,” *Genes Dev.*, vol. 25, no. 18, pp. 1915–1927, 2011.

Convergent downstream candidate mechanisms of independent intergenic polymorphisms between co-classified diseases implicate epistasis among noncoding elements [§]

Jiali Han[†], Jianrong Li[†] and Ikbel Achour[†,&]

Center for Biomedical Informatics and Biostatistics (CB2) and Departments of Medicine and of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA
Email: jialih@email.arizona.edu, jianrong@email.arizona.edu, ikachour@gmail.com

Lorenzo Pesce and Ian Foster

Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, IL 60637, USA
Email: lpesce@cs.uchicago.edu, foster@cs.uchicago.edu

Haiquan Li[*] and Yves A. Lussier[*]

CB2, BIO5 Institute, UACC, and Dept of Medicine, The University of Arizona, Tucson, AZ 85721, USA
Email: haiquan@email.arizona.edu, yves@email.arizona.edu

Eighty percent of DNA outside protein coding regions was shown biochemically functional by the ENCODE project, enabling studies of their interactions. Studies have since explored how convergent downstream mechanisms arise from independent genetic risks of one complex disease. However, the cross-talk and epistasis between intergenic risks associated with distinct complex diseases have not been comprehensively characterized. Our recent integrative genomic analysis unveiled downstream biological effectors of *disease-specific* polymorphisms buried in intergenic regions, and we then validated their genetic synergy and antagonism in distinct GWAS. We extend this approach to characterize convergent downstream candidate mechanisms of distinct intergenic SNPs *across distinct* diseases *within* the same clinical classification. We construct a multipartite network consisting of 467 diseases organized in 15 classes, 2,358 disease-associated SNPs, 6,301 SNP-associated mRNAs by eQTL, and mRNA annotations to 4,538 Gene Ontology mechanisms. Functional similarity between two SNPs (similar SNP pairs) is imputed using a nested information theoretic distance model for which p-values are assigned by conservative scale-free permutation of network edges without replacement (node degrees constant). At FDR≤5%, we prioritized 3,870 intergenic SNP pairs associated, among which 755 are associated with distinct diseases sharing the same disease class, implicating 167 intergenic SNPs, 14 classes, 230 mRNAs, and 134 GO terms. Co-classified SNP pairs were more likely to be prioritized as compared to those of distinct classes confirming a noncoding genetic underpinning to clinical classification (odds ratio ~3.8; $p \leq 10^{-25}$). The prioritized pairs were also enriched in regions bound to the same/interacting transcription factors and/or interacting in long-range chromatin interactions suggestive of epistasis (odds ratio ~ 2,500; $p \leq 10^{-25}$). This prioritized network implicates complex epistasis between intergenic polymorphisms of co-classified diseases and offers a roadmap for a novel therapeutic paradigm: repositioning medications that target proteins within downstream mechanisms of intergenic disease-associated SNPs. Supplementary information and software: http://lussiergroup.org/publications/disease_class

Keywords: SNP; Intergenic; Noncoding; Disease class; Biological similarity; Enrichment.

[§] This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)
[&] now employed at AstraZeneca MedImmune

1. Introduction

Human diseases can be classified via multiple criteria: cell type, tissue, organ, system, topological body region, pathophysiological, epidemiological characteristics, and etiological causes. Thus, in clinical classification of diseases, genetic disorders have conventionally relegated to a subset of the classification pertaining to its etiology. The advent of genomic assays now offers the opportunity to utilize unbiasedly a broad number of molecules of life to redefine the architecture of clinical classifications.

For example, cancers pertaining to distinct organ and cell types have been shown to share common somatic mutations [1] or transcriptomes and sometimes respond to the same therapy in spite of their distinct conventional classification, suggesting a new systems oncology etiology to cancer pathophysiology. We have previously shown that the miRNome of tumors classify the primary cancers by organ of origin as expected, while their paired metastases remarkably classify according to their progression (oligometastatic vs. polymetastatic) regardless of the primary site and metastatic site [2]. Recently, Genome-Wide Association Studies (**GWAS**) have implicated the same polymorphisms to distinct diseases of the same clinical class (e.g., cardiovascular system). Many distinct autoimmune diseases are found to have the same polymorphisms relating to the major histocompatibility complex region of chromosome 6, along with some other chromosome regions involving signaling in immune response (e.g., cytokine, interleukin, and interferon) [3, 4]. These same polymorphisms have also been associated with distinct traits of the metabolic syndrome [5].

In addition to studying each disease class separately, studies have also been conducted at a system level to unveil mechanisms that link individual diseases to a disease class. A disease class is likely to be driven by common genes and even common biological sub-networks, thus rendering a cluster structure or modularity in the biological network that separates it from other classes [6]. The modularity for disease classes has been observed in various types of molecular networks based on their risks identified in shared intragenic regions, including disease-gene networks [7-10], drug-target networks [11], transcription factor networks [6, 12], and protein-protein interaction networks [13]. Ohn broadened the similarity between diseases by looking into correlated polymorphisms by GWAS p-values [14]. In addition, two studies leveraged trans- Expression Quantitative Trait Loci (**eQTL**) analyses studies respectively limited to the immune systems and node-degree properties [15, 16]. On the other hand, traditional genetic-interaction studies such as PLINK [17] and BOOST [18], as well as recent integrative functional studies on non-coding disease variants [19, 20] such as GWAS3D [21] and CEPID [22] may also provide insight into how distinct diseases of the same disease class co-classify together. In spite of the genetic, genomic, and biological network studies generally conducted for specific disease classes, the biological mechanisms of the majority of disease-associated intergenic polymorphisms remain obscure as well as their contribution to explaining these risks at the disease class level.

We recently reported that downstream functional effects of distinct intergenic Single Nucleotide Polymorphisms (**SNP pairs**) associated with the same complex disease are likely to converge at some levels of biology such as sharing downstream transcripts or regulating functionally similar biological pathways or processes [23]. Our collaborators, Moore and Denny research groups, confirmed genetic synergy or antagonism between the top prioritized convergent intergenic SNP-

pairs in a GWAS of Alzheimer's and a Phenome-Wide Association Study (PheWAS) of rheumatoid arthritis [23]. However, this study did not address the convergent mechanism of SNP pairs between distinct diseases associated with the same clinical classification (**co-classified**).

Here, the downstream functional similarity between two SNPs (**similar SNP pairs**) is imputed using a multiscale information theoretic distance model for which p-values are assigned by conservative permutation resampling of network edges without replacement (node degrees constant). We hypothesized that we could extend this approach to identify downstream mechanisms of **intergenic SNPs with distinct co-classified diseases**, by integrating the classification information of the NHGRI diseases/traits and reanalyzing the results, to infer the noncoding genetic architecture of disease classes, which has implications for drug repositioning and mitigation of risks for multiple diseases within the same class.

2. Methods

2.1. *Main Datasets*

We surveyed Lead SNPs (SNPs investigated in GWAS) from two datasets, the National Human Genome Research Institute (**NHGRI**) GWAS catalog [24] and the eQTL association dataset named SNP and Copy Number Variant Annotation (SCAN) database [25]. The NHGRI GWAS catalog provides a comprehensive resource by systematically cataloging and summarizing the key characteristics of reproducible trait/disease-associated SNPs from currently published GWAS [24]. The NHGRI GWAS catalog comprises 7,236 associations between 574 diseases/traits and 6,432 distinct SNPs (6/7/2012). The SCAN database contains 4,189,682 eQTL associations between 833,004 distinct SNPs and 11,860 mRNA at $P \leq 10^{-4}$ from lymphoblastic cell lines. The integration of these two datasets yields 2,358 Lead SNPs in common (1,092 intergenic SNPs), along with their traits/diseases and mRNA information. The 574 NHGRI diseases/traits were classified into 15 organ & clinical systems disease classes according to Maurano et al. [6] along with curation (Suppl. Tab. 1).

A pairwise analysis was conducted on all possible combinations of two Lead SNPs inherited in distinct haplotypes (pairs of SNPs in strong linkage disequilibrium (**LD**) were removed from our study). The HapMap CEU LD dataset [26] was used to determine LD level and the exclusion criterion of $r^2 \geq 0.8$. Since our major interest is in intergenic variants (i.e., located between genes), the pairs in which both SNPs are intragenic (i.e., located within genes) were also excluded. The definition of "intergenic" and "intragenic" are derived from dbSNP (Build 138 on 2/21/2014) [27], which considers a SNP in a gene region to be intragenic if it is within 2kb upstream (5' side) or 0.5 kb downstream (3' side) of that gene. ~2.8 million pairwise combinations were derived from these Lead SNPs with $r^2<0.8$, associated with 467 diseases, 6,301 mRNAs, 1,635 molecular functions (**MF**), and 2,903 biological processes (**BP**). Among them, 1,977,927 pairs contain at least one intergenic SNP (named as intergenic Lead SNP pairs): 595,053 intergenic-intragenic and 1,382,874 intergenic-intergenic. 800,438 pairs are intragenic-intragenic. Among the intergenic Lead SNP pairs, 211,808 are associated with same disease classes (i.e., each SNP in one pair is associated with a specific disease class) while 1,766,119 are associated with distinct ones.

2.2. *Calculation of SNP similarity*

The prioritization process was applied to the intergenic Lead SNP pairs based on their convergence of eQTL-associated biological mechanisms. Three approaches were exploited to determine such shared (convergent) candidate mechanisms: (1) eQTL-associated mRNA overlap, (2) molecular function (MF) similarity of eQTL-associated mRNA, and (3) biological process (BP) similarity. We extracted MFs and BPs of each mRNA associated with a SNP from gene ontology (GO) annotations [28, 29] to calculate the similarity of a SNP pair [23] (**Table 1 & Figure 1**).

Table 1. Biological similarity calculations between two SNPs using nested Information Theoretic Similarity (ITS)

Nested calculations (3 steps)
1. Calculate the Information Theoretic Similarity (ITS) between two GO terms (GO_{ITS}) associated with the two SNPs through mRNAs using Lin's method [30, 31].
2. Based on GO_{ITS}, calculate the information theoretic similarity between two distinct mRNAs ($mRNA_{ITS}$), each associated with a SNP within a SNP Pair, using a modified Tao's approach [31-33].
3. Determine the semantic biological similarity between two SNPs (SNP_{ITS}) within a SNP pair using the $mRNA_{ITS}$ of pairs of mRNAs associated with the two SNPs respectively, using Li's nested ITS approach we recently published [23]. The SNP_{ITS} values range from 0 to 1, with 0 corresponding to no similar downstream effects and 1 corresponding to identical downstream effects (e.g., either the same mRNAs or distinct mRNAs with the equivalent GO terms). The similarity measurement between SNPs can capture relationships between SNPs including the ones without any common mRNAs in their eQTL associations.

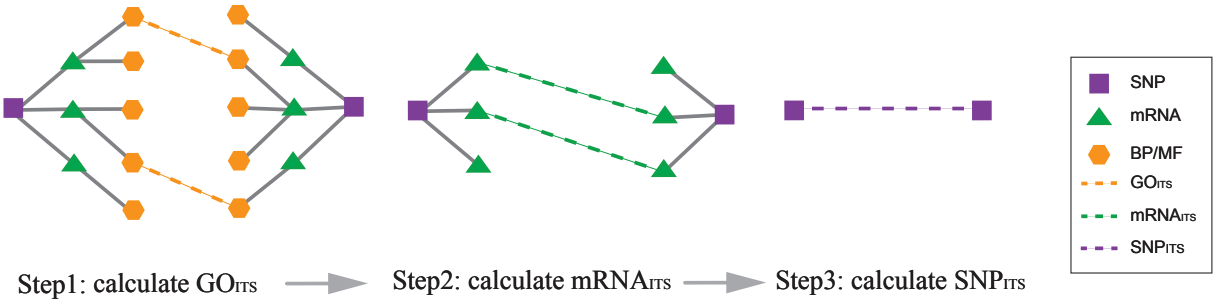


Fig. 1. Nested Information theoretic calculations. The similarity between SNP pairs is calculated by three nested steps subsequently (I) similarity between two gene ontology terms (GO_{ITS}), (II) similarity between two genes ($mRNA_{ITS}$) using GO term similarities, and (III) similarity between two SNPs (SNP_{ITS}) using mRNA similarities.

2.3. *Network permutation to establish the p-values for observed mRNA overlap and ITS scores between two SNPs*

To determine the statistical significance of imputed biologically convergent mechanisms of SNP pairs, permutation of the eQTL network was conducted for mRNA overlap, molecular function similarity, and biological process, separately. We also included the eQTL associations of SNPs not known to be associated with any diseases to create a null distribution of SNP mRNA overlap (**statistical mRNA overlap**) and ITS. When examining the significance of each of the three mechanisms, we controlled the original node degree (ND) of each specific SNP and each specific mRNA. Specifically, we kept the number of mRNAs associating with one SNP the same, or vice versa, during the resampling of the bipartite eQTL network (shuffling the associations between

SNPs and mRNAs). Deep permutations at 100,000 times were conducted on the Argonne Lab Beagle supercomputer to reach a sufficient power (20 million core hours). P-values were derived from the imputed results of the observed eQTL network and the set of permuted networks. False Discovery Rate (**FDR**) was used to adjust for multiplicity, and the SNP pairs with FDR<0.05 are termed prioritized Lead SNP pairs.

For MF and BP similarity calculations, a similar permutation procedure was conducted as done for mRNA overlap, except that SNPs and mRNAs without corresponding GO annotations were removed and only those with BP or MF associations remained in the bipartite network for resampling. We further investigated the significance of overlapped GO terms from the SNP-GO-SNP triplets for every pair of SNPs based on the same set of permutations and prioritized the overlapped terms between pairs of SNPs with a FDR<0.05. The whole procedure of permutations was conducted multiple times for different eQTL association cutoffs ranging from $P \leq 10^{-4}$ to $P \leq 10^{-6}$ and at three levels of node degrees: ND≥1, ND≥3, and ND≥5.

Through such stringent scale-free network controls, not only will the SNP pairs associated with same mRNAs be prioritized, but also the pairs in which two SNPs are associated with distinct mRNAs, if biological similarity exists.

2.4. *Internal Validation: enrichment studies of co-classified intergenic SNP pairs among prioritized pairs*

To demonstrate whether the shared biological mechanisms of intergenic Lead SNP pairs are relevant to the underlying biology of disease classes, we assessed whether they are more likely to be found related to the same disease class than those across distinct classes. One-tailed Fisher's Exact Test (**FET**) was applied for the enrichment study, and odds ratios of significant mRNA overlapping, MF, and BP similarities for SNP pairs associated with the same disease classes were calculated by FET at multiple eQTL p-value cutoffs and three levels of node degrees.

2.5. *External Validation: ENCODE regulatory elements and chromatin interaction enrichment of co-classified prioritized intergenic SNP pairs*

The potential mechanisms at play for the prioritized SNP pairs were also investigated. We evaluated whether regulatory mechanisms were more likely to occur in prioritized intergenic SNP pairs associated with the same disease class as compared to their counterparts (distinct classes or insignificant). We integrated Encyclopedia of DNA Elements (**ENCODE**) data [19] of Lead SNPs and conducted Fisher's Exact Test to assess the enrichment of molecular regulations within prioritized SNP pairs of the same disease classes. Three possible shared regulatory mechanisms are assessed for pairs of SNPs located in distinct regions, including (1) binding with same transcription factor (via ChIP-seq), (2) binding with distinct transcription factors (via **ChIP-seq**) connecting through protein-protein interaction (**PPI**), and (3) within the anchor regions of long-range chromatin interactions (via **ChIA-PET** [34]). We compared the enrichment of regulatory mechanisms with two conventional methods, which prioritized SNP pairs by (1) any intergenic Lead SNP pairs and (2) intergenic Lead SNP pairs with at least one mRNA overlap (**non-statistical mRNA overlap**) in eQTL associations, respectively. To avoid loss of information when calculating regulatory functions

between Loci in ENCODE, every Lead SNP was extended to its strongly associated LD SNPs based on the RegulomeDB database [35] (inheritable haplotype).

3. Results and Discussion

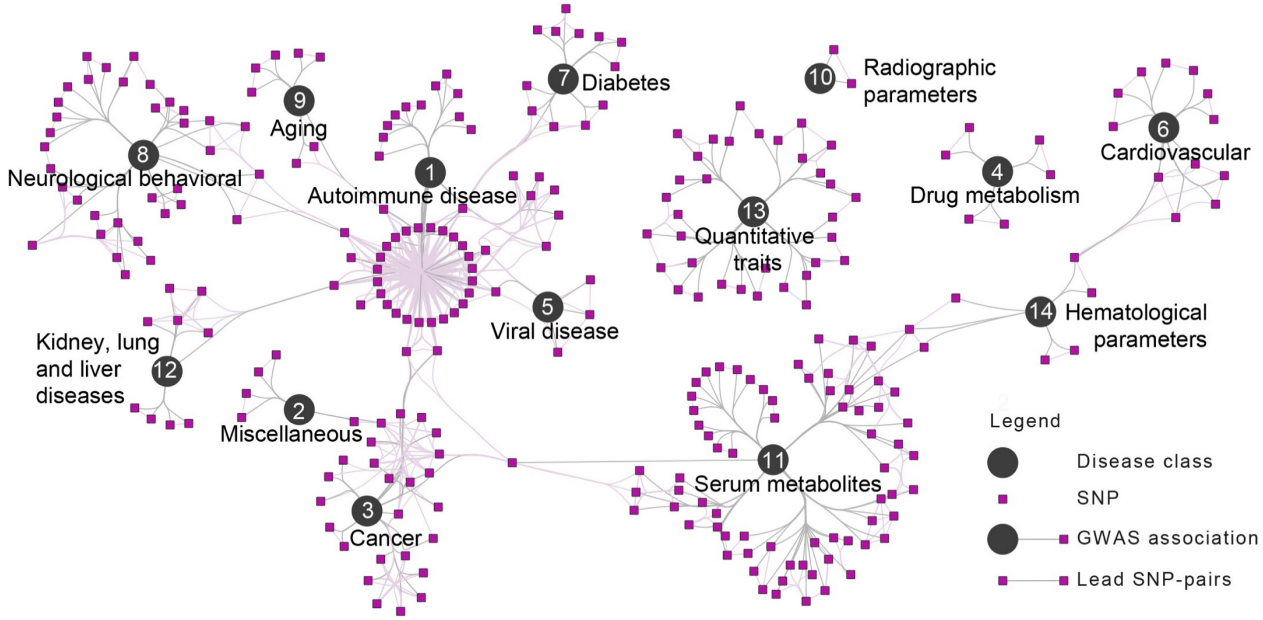


Fig. 2. The network of 755 prioritized intergenic SNP pairs within disease class at FDR<0.05. 80 SNP pairs are within the same disease (previously published), 675 are within the same disease class but across distinct diseases (new). 3,115 SNP pairs prioritized cross-class are not shown. 19 SNPs were associated with two distinct diseases in distinct classes by GWAS and shown.

3.1. *Overall results and visualization*

Prioritization of convergent downstream mechanisms of SNPs required extensive conservative scale-free permutation resampling of network edges (node degrees constant), shown substantially more conservative than conventional theoretical statistics or similarity-scores cutoffs (Suppl. Fig. 1). We prioritized 3,870 intergenic Lead SNP pairs (1,378 intergenic-intergenic; 2,492 intergenic-intragenic) at FDR<0.05 that share at least one of the three imputed biological mechanisms, of which 755 pairs are found within the same disease class (280 intergenic-intergenic pairs; 475 intergenic-intragenic; 80 were associated with the same diseases). Without additional prioritization, the network relates these 755 pairs with as many as 1,683 mRNAs and 2,060 GO terms. After convergent mechanism prioritization, these SNP pairs implicate 14 disease classes, 277 Lead SNPs (167 intergenic, 98 noncoding intragenic, 12 protein-coding), 230 mRNAs, and 134 GO

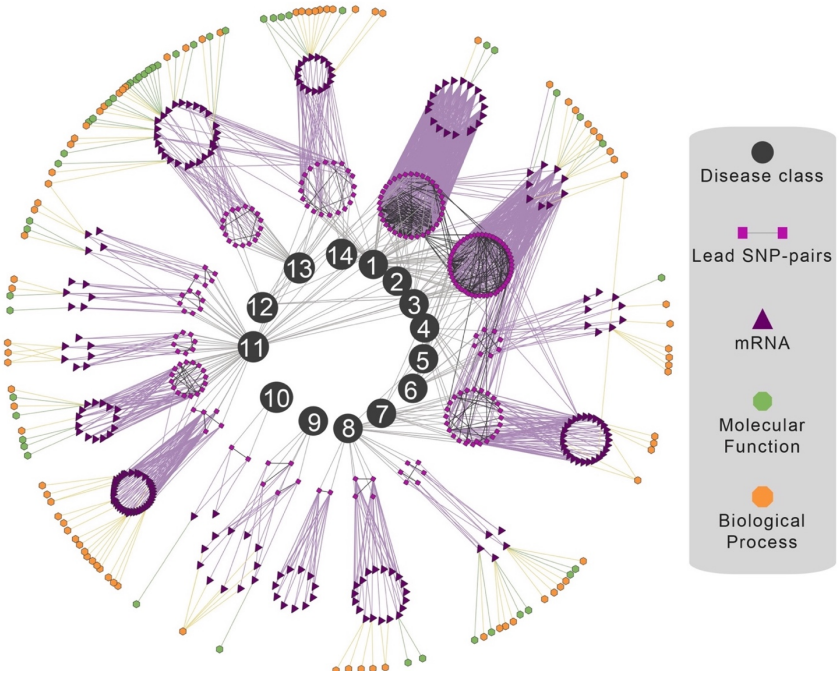


Fig. 3. The subset of the prioritized network of disease class mechanisms containing 230 mRNAs shared between 428 SNP pairs and their associated GO mechanisms (48 GO-MFs, and 86 GO-BPs). Biological modularity of shared groups of mRNAs is associated with distinct SNPs themselves associated with distinct co-classified diseases. Not shown are the biomodules where 327 SNP-pairs are associated by distinct mRNAs to distinct but similar pathways (Methods 2.2). Names of classes are defined in Fig. 2.

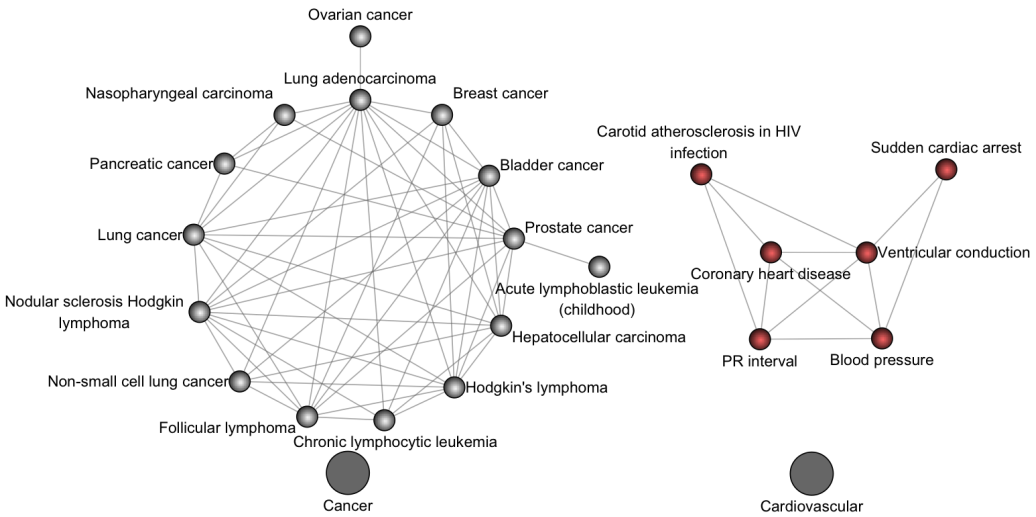


Figure 4. Details of implicated co-classified diseases through SNP pair similarity confirming shared genetic underpinning and biological mechanisms. Two classes, cancers (Fig. 2-3 #3) and cardiovascular disease (Fig.2-3; #6), shown. Disease-pairs are related by at least one out of 755 prioritized pairs of Lead SNPs, each associated with a disease in the pair respectively. Previous studies have shown somatic mutations and transcriptomes can reclassify cancers molecularly. Here a new property is presented: common mechanisms of noncoding intergenic regions.

mechanisms. A simplified network shows only the 755 prioritized intergenic Lead SNP pairs and their related disease classes, leaving out the mRNAs and GO-terms for simplicity (**Fig. 2**). 14 of the 15 studied disease classes harbor convergent biological processes and molecular functions perturbed by a set of intergenic SNPs with similar downstream effects, presenting an apparent modularity for each class. We further show a sub-network of prioritized biological mechanisms for the prioritized SNP pairs associated with the same classes in **Fig. 3**. The convergent connections among intergenic SNPs of distinct diseases within the same disease class suggest the investigation of an unusual form of pleiotropy: distinct intergenic risks of co-classified disease sharing common downstream mechanisms that could affect the same target transcripts that may relate to the emergence of both diseases in the same pathophysiological classification (e.g., **Fig. 4** showing the detail of co-classified diseases associated through SNP pair similarity in Fig. 2, only cancer and cardiovascular system shown).

3.2. *Enrichment of shared biological mechanisms in prioritized intergenic SNP pairs of distinct co-classified diseases (Methods 2.4)*

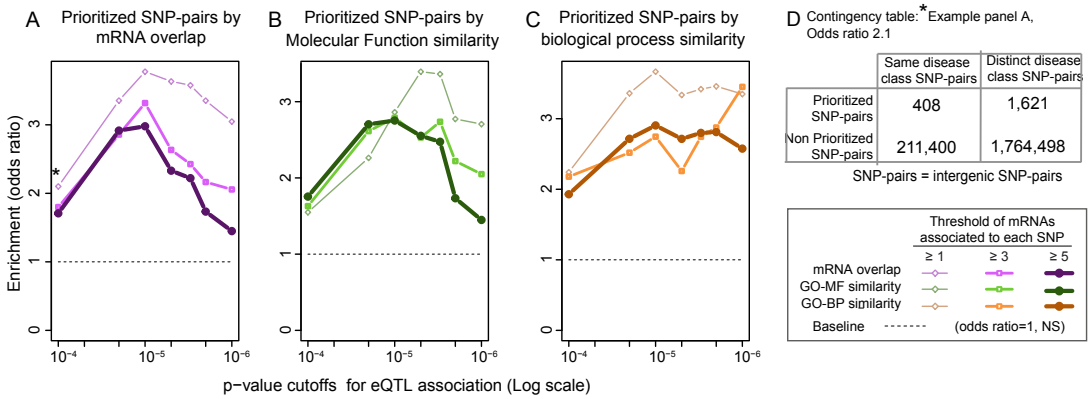


Fig. 5. Enrichment of shared biological mechanisms among 755 intergenic Lead SNP pairs associated with the same disease classes (Method 2.1, LD cutoff $r^2<0.8$), remains similar with more stringent LD cutoff ($r^2<0.01$, not shown) and also remains the same when excluding the previously published 80 SNP pairs associated with the same diseases (results not shown). The subset of 280 prioritized SNP pairs comprising only intergenic-intergenic pairs also remains significant (Suppl. Fig. 2).

We investigated whether intergenic Lead SNP pairs, with each SNP associated with two distinct co-classified diseases, were more likely to share a biological mechanism (prioritized) than SNP pairs associated with distinct diseases classified in distinct pathophysiological classes. Enrichment analyses were performed for the 755 prioritized SNP pairs associated with same classes among 3,870 prioritized intergenic Lead SNP pairs at different eQTL p-value cutoffs ($10^{-6} \leq$ eQTL p-value$\leq 10^{-4}$; 100,000 permutation resampling, SNP pair FDR<0.05) and different node degrees SNP node degree (count of mRNAs associated with that SNP at the eQTL p-value cutoff). As shown in **Fig. 5,** odds ratios (**ORs**) range from 1.4 to 3.8 (x-axis: $5.1\times10^{-6}\leq$p-value≤0.02), 1.4 to 3.4 ($6.5\times10^{-26}\leq P\leq2.1\times10^{-2}$), and 1.9 to 3.7 ($8.3\times10^{-4}\leq P\leq2.2\times10^{-7}$) for mRNA overlapping, MF similarity, and BP similarity, respectively. This internal validation supports the hypothesis that biological

mechanisms are more likely to be shared within a class of diseases and may define in part a common pathophysiology of otherwise distinct diseases.

3.3. *Enrichment of ENCODE regulatory elements and chromatin interaction in prioritized intergenic SNP pairs of distinct co-classified diseases (Methods 2.5)*

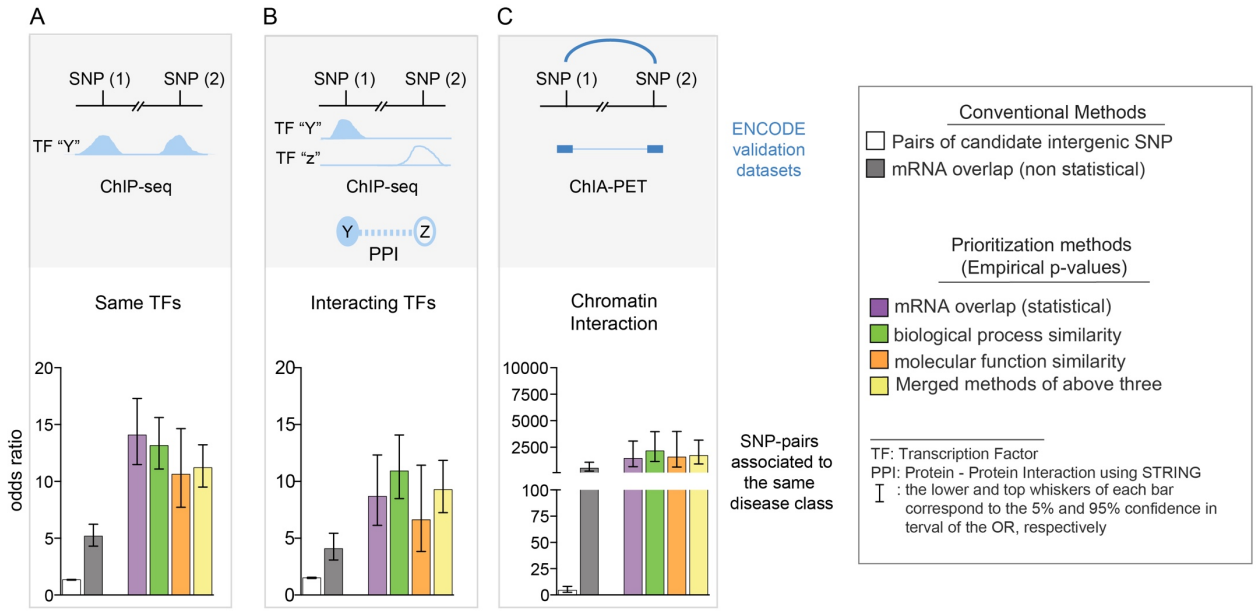


Fig. 6. Enrichment of common ENCODE-derived regulatory mechanisms in genomic regions of the prioritized intergenic Lead SNP pairs for disease classes. More stringent LD cutoff (R^2<0.01) yielded similar results (not shown).

ENCODE data provides an opportunity to question if convergent candidate mechanisms of prioritized SNP pairs of co-classified diseases imputed by eQTL associations may be attributed to common regulatory elements (e.g., transcriptional factors) or long-range chromatin interactions. If so, this could be suggestive of possible epistasis between disease risks of distinct co-classified diseases, in other words, a disease class epistasis. We identified substantial enrichment in three types of regulatory elements: shared transcription factor (**Fig. 6 panel A**), interacting transcription factors (**Fig. 6 panel B**), and long-range chromatin interactions in the region of the SNPs in the pair (**Fig. 6 panel C**). However, the effect size (odds ratio) of enrichment of regulatory elements in SNP pairs associated with distinct co-classified diseases shown in the figure is about 30 percent smaller than that of our previously published enrichment of SNP pairs associated with the *same* disease (not shown [23]). Taken together, these results indicate that common regulatory mechanisms of intergenic SNPs strongly underpin the pathogenesis of a disease and to a moderate degree some mechanisms are also shared by distinct, yet pathophysiologically co-classified diseases.

4. Limitations and future studies

First, we only reported eQTLs derived from LCL cell lines. Studies on 44 tissues in the GTEx project are ongoing and will be reported elsewhere. SNPs with marginal p-values [36] will also be investigated

using the proposed method to unveil their pairwise synergy. Second, gene ontology annotations are biased by human interest. Even though the biases were controlled partially by the scale-free persisted permutations, some biases may still exist and induce false positive results. Alternative unbiased approaches may be worth incorporating in the future such as the information-theoretic framework to address the accuracy of the GO annotation [37-39]. Third, the permutations on large eQTL networks are expensive; we are working on more efficient implementations and strategies. Fourth, the validation in a GWAS of epistasis between convergent intergenic SNPs associated with distinct co-classified diseases is not possible retrospectively as clinical phenotypes are generally obtainable for only one disease in a GWAS. A prospective study for the validation is cost-prohibitive; we are thus planning a collaboration with eMERGE researchers to conduct a PheWAS. Finally, the SNPs prioritized in this study are statistically associated with but not necessarily functionally causal to a disease (or co-classified diseases) thus other polymorphisms inherited in the same loci must be considered. Of note, our approach incorporated this calculation through the Linkage Disequilibrium parameter (**Methods section 2.5**). Also, further systematic investigation on the relationship between functional synergy and genetic interaction of SNPs prone to same or co-classified diseases will provide insight into the mechanisms of disease classes.

Beyond the modularity within classes, related disease classes are obviously also interconnected through shared genes and gene ontology annotations in Fig. 3. This study focused on intergenic SNPs prioritized across distinct diseases of the same class, leaving out thousands of SNP pairs prioritized across classes. Indeed, cross-class biomodularity merits its own publication and additional analyses due to its complexity.

5. Conclusion

Using the quantified measurement of SNP biological similarity we recently developed, we identified 755 intergenic SNP pairs associated by convergent eQTL function to distinct, yet pathophysiologically co-classified diseases. We found that these independently inherited (LD $r^2<0.01$) intergenic SNP pairs were more likely to be enriched in (i) shared transcription factors, (ii) interacting transcription factors, and (iii) long-range chromatin interactions. A common genetic architecture of the pathophysiology of co-classified diseases is unsurprising; however, a common noncoding intergenic architecture for clinical classification harbors many new questions. For example, is epistasis occurring between distinct disease risks, and if so, can some disease risks protect against other diseases through antagonism of long-range chromatin interactions implicating noncoding intergenic regions? Additionally, can we implicate new drug targets or reposition drugs through the shared intergenic interactions between distinct co-classified diseases? While the prioritized intergenic SNP pairs associated with each disease class reassuringly recapitulates the pathophysiological classification of disease of complex inheritance, does this implicate that complex diseases are fundamentally distinct from Mendelian ones through these noncoding interactions? Indeed, GWAS identified about half the variants in intergenic regions. However, the array platforms are seeded biasedly with half the probes in intergenic regions (selection bias). This proposes that more than 80% of the complex-disease associated variants could be located in intergenic regions, suggesting that if the heritability gap is attributable to genetic interactions, the majority of these

would occur with intergenic noncoding regions. On the other hand, our study aligns further intergenic genetic signal with that of the central dogma of molecular biology, as we provide for each prioritized SNP pair falsifiable hypotheses of convergent mechanisms implicating coding regions (eQTL mRNAs).

This prioritized network implies complex epistasis between intergenic polymorphisms of co-classified diseases and offers a roadmap for a novel therapeutic paradigm: repositioning medications that target proteins within downstream mechanisms of intergenic disease-associated SNPs.

6. Acknowledgements

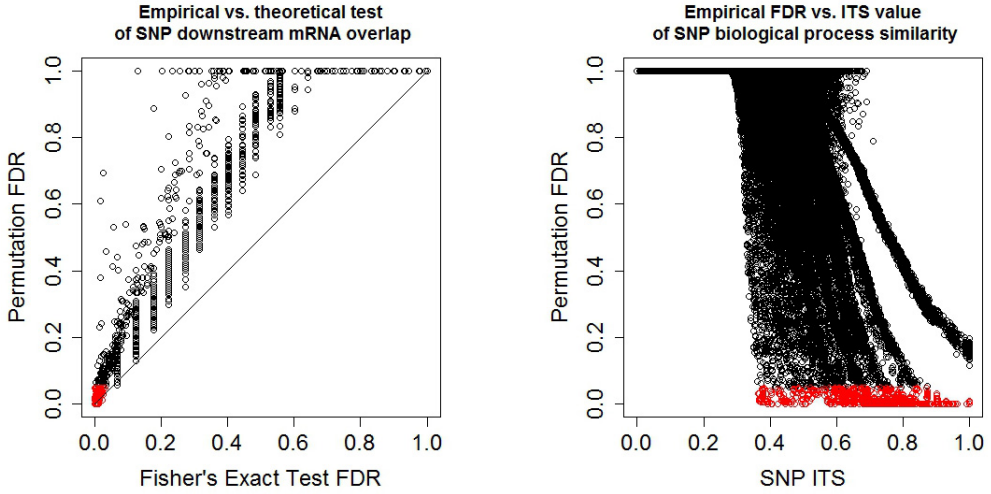
We thank Drs. Nancy Cox and Eric R. Gamazon for sharing their eQTL associations, Dr. Roger Luo for curating the disease classification, and Dr. Colleen Kenost for proofreading the manuscript.

References

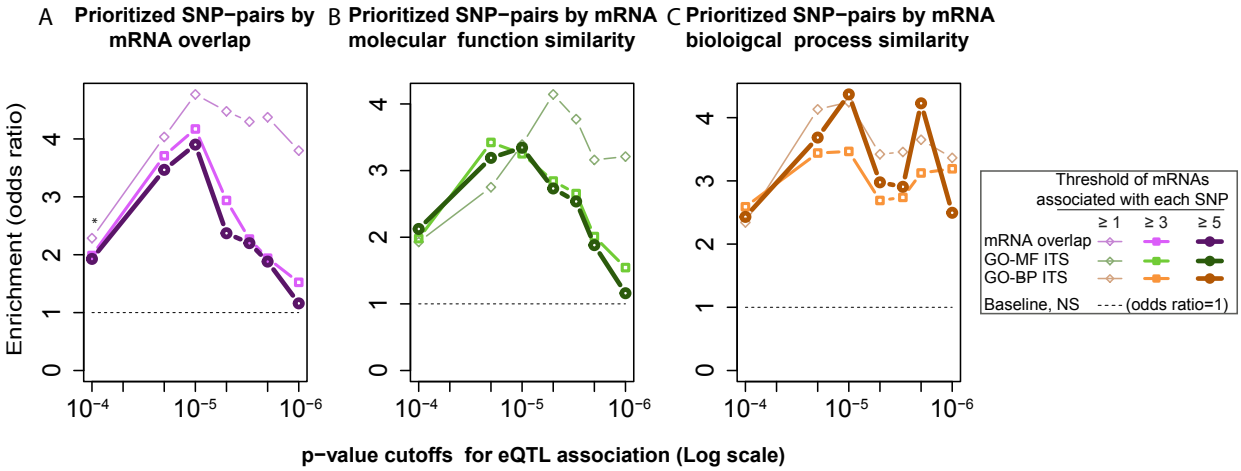
1. K. Inaki and E. T. Liu, *Trends in Genetics*, 2012, **28**, 550-559.
2. Y. A. Lussier, H. R. Xing, et al., *PloS one*, 2011, **6**, e28650.
3. M. F. Seldin, *Journal of autoimmunity*, 2015, **64**, 1-12.
4. L. A. Zenewicz, C. Abraham, et al., *Cell*, 2010, **140**, 791-797.
5. S. Vattikuti, J. Guo and C. C. Chow, *PLoS genetics*, 2012, **8**, e1002637.
6. M. T. Maurano, R. Humbert, et al., *Science*, 2012, **337**, 1190-1195.
7. B. Bulik-Sullivan, H. K. Finucane, et al., *Nature genetics*, 2015, **47**, 1236-1241.
8. K.-I. Goh, M. E. Cusick, et al., *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685-8690.
9. X. Jiang, B. Liu, et al., *FEBS letters*, 2008, **582**, 2549-2554.
10. Y. Lee, J. Li, et al., *Summit on Translational Bioinformatics*, 2010, 31.
11. M. A. Yildirim, K.-I. Goh, et al., *Nature biotechnology*, 2007, **25**, 1119.
12. K. J. Karczewski, J. T. Dudley, et al., *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 9607-9612.
13. L. Sam, Y. Liu, et al., *Pacific Symposium on Biocomputing* 2007, 76.
14. J. H. Ohn, *J. Am. Med. Inform. Assoc.*, 2017, ocx026.
15. R. S. Fehrmann, R. C. Jansen, et al., *PLoS genetics*, 2011, **7**, e1002197.
16. H. Li, N. Pouladi, et al., *Journal of biomedical informatics*, 2015, **58**, 226-234.
17. S. Purcell, B. Neale, et al., *The American Journal of Human Genetics*, 2007, **81**, 559-575.
18. X. Wan, C. Yang, et al., *The American Journal of Human Genetics*, 2010, **87**, 325-340.
19. ENCODE Project Consortium, *Nature*, 2012, **489**, 57.
20. Y. Lee, E. R. Gamazon, et al., *PLoS genetics*, 2012, **8**, e1002998.
21. M. J. Li, L. Y. Wang, et al., *Nucleic acids research*, 2013, **41**, W150-W158.
22. M. J. Li, M. Li, et al., *Genome biology*, 2017, **18**, 52.
23. H. Li, I. Achour, et al., *NPJ genomic medicine*, 2016, **1**, 16006.
24. D. Welter, J. MacArthur, et al., *Nucleic acids research*, 2013, **42**, D1001-D1006.
25. E. R. Gamazon, W. Zhang, et al., *Bioinformatics*, 2009, **26**, 259-262.
26. R. A. Gibbs, J. W. Belmont, et al., 2003.
27. S. T. Sherry, M.-H. Ward, et al., *Nucleic acids research*, 2001, **29**, 308-311.
28. M. Ashburner, C. A. Ball, et al., *Nature genetics*, 2000, **25**, 25.
29. Gene Ontology Consortium, *Nucleic acids research*, 2010, **38**, D331-D335.
30. D. Lin, *International Confernece on Machine Learning*, 1998, 296-304.

31. Y. Tao, L. Sam, et al., *Bioinformatics*, 2007, **23**, i529-i538.
32. K. Regan, K. Wang, et al., *J. Am. Med. Inform. Assoc.*, 2012, **19**, 306-316.
33. H. Li, Y. Lee, et al., *J. Am. Med. Inform. Assoc.* , 2012, **19**, 295-305.
34. S. Djebali, C. A. Davis, et al., *Nature*, 2012, **489**, 101.
35. A. P. Boyle, E. L. Hong, et al., *Genome research*, 2012, **22**, 1790-1797.
36. Y. Lee, H. Li, et al., *J. Am. Med. Inform. Assoc.*, 2013, **20**, 619-629.
37. G. Alterovitz, M. Xiang, et al., *Nucleic acids research*, 2006, **35**, D322-D327.
38. J. L. Chen, Y. Liu, et al., *BMC bioinformatics*, 2007, **8**, S7.
39. W. T. Clark and P. Radivojac, *Bioinformatics*, 2013, **29**, i53-i61.

Supplementary Figures



Suppl. Fig. 1. Permutation-based empirical statistics is more conservative than Fisher's Exact Test when assessing mRNA overlap (left panel) and semantic similarity of biological processes (right panel) of SNP pairs. Prioritized SNP pairs are shown in red.



Suppl. Fig. 2. Enrichment of shared mechanisms among the subset of intergenic-intergenic Lead SNP pairs associated with distinct diseases of the same disease classes. Compared with Fig. 5 where intergenic-intragenic pairs were included with the same LD cutoff r2<0.8, the enrichment is higher for exclusively intergenic pairs.

Network analysis of pseudogene-gene relationships: from pseudogene evolution to their functional potentials

Travis S Johnson, MS
Dept. Biomedical Informatics, Ohio State University,
5000 HITS, 410 W. 10th St. Indianapolis, Indiana, 46202
Travis.Johnson@osumc.edu

Sihong Li
Dept. Biomedical Informatics, Ohio State University,
250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
li.6001@buckeyemail.osu.edu

Jonathan R Kho
Dept. Computational Science and Engineering, Georgia Institute of Technology
Klaus Advanced Computing Building, 266 Ferst Dr. Atlanta, Georgia, 30332
jkho@gatech.edu

Kun Huang, PhD
Dept. Hematology Oncology, Indiana University,
335 Regenstrief Institute, 1101 W 10th St. Indianapolis, Indiana, 46202
kunhuang@iu.edu

Yan Zhang, PhD*
Dept. Biomedical Informatics, Ohio State University,
310-B Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
Yan.Zhang@osumc.edu

Pseudogenes are fossil relatives of genes. Pseudogenes have long been thought of as "junk DNAs", since they do not code proteins in normal tissues. Although most of the human pseudogenes do not have noticeable functions, ~20% of them exhibit transcriptional activity. There has been evidence showing that some pseudogenes adopted functions as lncRNAs and work as regulators of gene expression. Furthermore, pseudogenes can even be "reactivated" in some conditions, such as cancer initiation. Some pseudogenes are transcribed in specific cancer types, and some are even translated into proteins as observed in several cancer cell lines. All the above have shown that pseudogenes could have functional roles or potentials in the genome. Evaluating the relationships between pseudogenes and their gene counterparts could help us reveal the evolutionary path of pseudogenes and associate pseudogenes with functional potentials. It also provides an insight into the regulatory networks involving pseudogenes with transcriptional and even translational activities.

In this study, we develop a novel approach integrating graph analysis, sequence alignment and functional analysis to evaluate pseudogene-gene relationships, and apply it to human gene homologs and pseudogenes. We generated a comprehensive set of 445 pseudogene-gene (PGG) families from the original 3,281 gene families (13.56%). Of these 438 (98.4% PGG, 13.3% total) were non-trivial (containing more than one pseudogene). Each PGG family contains multiple genes and pseudogenes with high sequence similarity. For each family, we generate a sequence alignment network and phylogenetic trees recapitulating the evolutionary paths. We find evidence supporting the evolution history of olfactory family (both genes and pseudogenes) in human, which also supports the validity of our analysis method. Next, we evaluate these networks in respect to the gene ontology from which we identify functions enriched in these pseudogene-gene families and infer functional impact of pseudogenes involved in the networks. This demonstrates the application of our PGG network database in the study of pseudogene function in disease context.

Keywords: Pseudogene-gene (PGG) relationship; Network analysis; Pseudogene function; PGG network database.

* To whom correspondence should be addressed.

1 Introduction

Pseudogenes have long been deemed "relics of evolution", because they are homologous to protein-coding genes but lack protein products[1]. Recently this nonfunctional label has started to be revised. Although most of the human pseudogenes do not have noticeable functions, ~20% of them exhibit transcriptional activity[2]. Recent studies have shown that pseudogenes can modulate gene expression and thus may influence signaling pathways in cancer[3]. Acquired somatic mutations can create pseudogenes in cancer development[4]. Evidence of pseudogene transcription and translation have been observed in cancer cell lines[5]. Besides, transcriptomics analysis has shown that transcribed pseudogenes are differentially expressed in specific cancer subtypes and could potentially be used as both prognostic and diagnostic biomarkers[6]. Another area of interest is the role of pseudogene transcripts in regulating gene expression. Some pseudogenes generate RNA products that can competitively bind to microRNAs thus regulating the expression of their homologous gene counterparts (i.e. ceRNAs)[7-9]. They can also be oncomodulatory, such as pseudogene PTENP1 regulating the PTEN tumor suppressor gene. PTENP1 locus lost in the genome could lead to tumorigenesis[10,11]. Pseudogenes might also represent a genetic diversity reservoir[12] and play a role in new gene generation[3]. Because of all this, understanding the relationships of pseudogenes and gene counterparts on a systems biology level is important in not only understanding evolution but also understanding diseases like cancer. However, the role(s) of pseudogenes are still not fully understood.

Efforts have been made to explore the roles of pseudogenes. A prominent example was Pseudogene.org, which compiled information on pseudogenes of various species. This database annotated pseudogenes and compared pseudogenes with their parent genes[13]. We attempt to complement this knowledge by focusing only on human and by comparing all pseudogenes to all gene families.

Processed pseudogenes and duplicated pseudogenes are two major types of pseudogenes. They are derived from functional genes through (retro-)duplication followed by accumulation of loss-of-function mutations[5,14]. Conventionally, a pseudogene is often paired with a homologous gene counterpart referred to as a "parent gene"[14,15]. This understanding of pseudogenes, though informative, does not encompass the entirety of genome-wide relationships, where multiple genes and pseudogenes can be homologous. Thus we instead compare pseudogenes with homologous gene families, such that pseudogenes are not only considered as a descendent of a single gene, but rather a relative of a group of homologous genes and pseudogenes. In this study, we develop a novel approach to generate the comprehensive set of pseudogene-gene families in human, and further characterize the networks and study the potential functions of pseudogenes and their associated networks in more detail using sequence alignment, network analysis, and functional annotation.

2 Materials and Methods

The simplified workflow is shown in Figure 1.

2.1 *Generating gene homolog families*

We constructed the gene homolog families in which all the members are homologous genes. Specifically, all gene homolog pairs in human genome GRCh38 were downloaded from Ensembl BioMart. These gene homolog pairs were combined to generate a network of all homology relationships in the genome. In this network, each node represents a gene, and each edge indicates the existence of homology between a pair of genes. This basic structure of genes/pseudogenes as nodes and homology as edges was used throughout this project. Because not all genes share a common homolog, the full GRCh38 gene homology network is not a connected graph. Thus we performed an initial separation of the gene homolog network into connected subgraphs, denoted as *gene families* throughout this paper (Figure 1). In total we generated 3,281 gene families.

The human genome GRCh38 annotation was downloaded from Human GENCODE release 24[16]. Full length gene sequences were extracted using the *gff2sequence* tool[17].

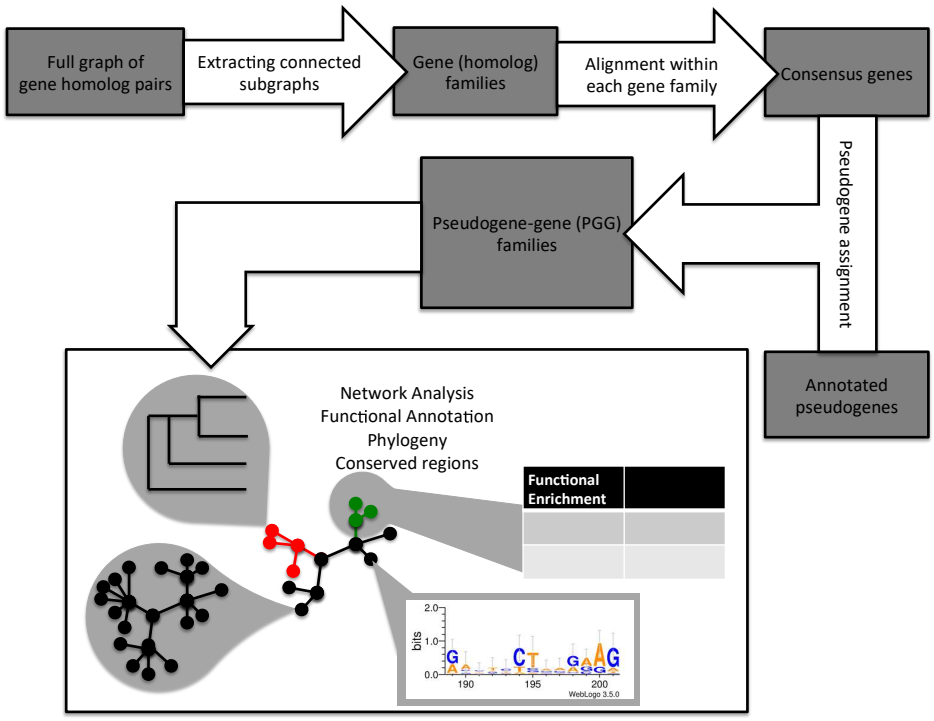


Figure 1. Simplified workflow from raw annotation through processing and analysis of the resulting pseudogene-gene (PGG) families.

2.2 *Mapping pseudogenes to gene homolog families*

We extracted pseudogene sequences annotated in Human GENCODE release 24 using cufflinks *gffread*[18]. In total we obtained 14,548 pseudogenes. We assigned these pseudogenes to homologous gene families, by aligning every pseudogene sequence to a consensus sequence from each gene family. Consensus sequences were used to reduce the computational complexity and reduce bias associated to gene family size. The consensus sequence representing a gene family was selected from the set of all sequences within a gene family, by performing a multiple sequence alignment and selecting the gene with the highest sum of all pairwise alignment scores. To limit errors associated with long runtime, only sequences with less than 10,000 bp were retained for this alignment step. After a consensus gene was selected from each gene family, all pseudogenes were aligned to that consensus sequence using pairwise ClustalW[19], which was installed on the high-performance computing cluster and has the ability to perform pairwise and multi-sequence alignment. The DataCutter framework[20] was used to parallelize the alignments producing a 14,548 by 3,281 matrix of alignment scores between each pseudogene and every consensus sequence for the gene families. Each pseudogene was then subsequently assigned to the gene family with the highest alignment score. Not all gene families have pseudogenes assigned to them. Those gene families with assigned pseudogenes, i.e. pseudogene-gene (PGG) families, were examined further (Figure 1).

2.3 *Network analysis of individual PGG families*

The resulting PGG families were used to generate PGG networks based on the pairwise local alignment scores between all members of the family. Local alignment was used so that shorter sequences could still have high alignment scores when they match to a short segment of a larger sequence. The pairwise alignments were performed with a GPU parallelized local alignment tool CUDA-align[21] in order to boost alignment performance for this large-scale computing. Using the resultant within-PGG family alignment matrix, a minimum spanning tree (MST) was generated for each PGG family. The alignment matrix for each PGG family consists of a complete network where all pseudogenes/genes within that family were

nodes and the pairwise alignment scores edges. The MST was calculated from the alignment matrix producing a network in which bottlenecks had high sequence similarity to other nodes.

One facet of interest was identifying bottleneck nodes (gene or pseudogene) in the PGG family networks. As a measure of importance, betweenness centrality (BC) was calculated for all genes and pseudogenes contained within each network. A node with high non-zero BC is more likely to be a bottleneck in the network. (The smaller the proportion of zero-BC nodes is in a network, the more bottlenecks there are in the network.) Thus we record the proportion of zero-BC nodes in pseudogenes and genes respectively, and compare the number of bottlenecks in pseudogenes and genes. The distribution of BC for pseudogenes and genes were also plotted to evaluate the importance of pseudogene and gene bottlenecks.

2.4 *Functional annotation of PGG families*

Next we evaluated the functional enrichment of genes contained within pseudogene-gene (PGG) families (i.e. gene families that were assigned at least one pseudogene). A list of all genes (excluding the assigned pseudogenes) contained within PGG families were extracted and submitted to the DAVID Functional Annotation Tool[22,23]. DAVID functional annotation clusters (at High stringency) were evaluated for over-represented annotations.

2.5 *Identifying phylogenetic relationships and conserved regions*

For each PGG family multiple sequence alignment (MSA) was performed with MUSCLE aligner[24]. Phylogenetic trees were created based on these alignments using FastTree[25]. The resulting MSAs and trees were used as input for the PhastCons[26] program to identify conserved regions within each PGG family. We used the two-step approach outlined in their user manual in which the first step trains the Hidden Markov Model (HMM) transition model and the second identifies conserved regions (CRs). We then evaluated gene families (with no aligned pseudogenes) and PGG families (with at least one aligned pseudogene) for differences in their likelihood of containing conserved regions using a Fisher's exact test. The test was conducted such that the two rows in the contingency table consisted of whether a gene family contained a pseudogene (row 1) or not (row 2). The columns consisted of whether a conserved region was identified in a gene family by PhastCons (column 1) or not (column 2).

2.6 *Identifying GO networks associated with PGG families*

The PGG networks identified (in-house PGG family IDs: 1149,1152,1235) were individually used to generate GO term networks using the BiNGO tool[27] in cytoscape[28,29]. GO term annotations and hierarchy are used to generate a network from the members of each PGG family. Within each PGG family we use these GO networks to view the possible functional impact of the pseudogenes assigned to each PGG family. Each pseudogene that is contained within the PGG families could have an effect on the functions detected by the BiNGO tool. Specifically, functional impact (functional roles) of the pseudogenes of interest are interpretable from the proximity of pseudogenes to their gene counterparts in the PGG networks.

3 Results

3.1 *Generating gene homolog families*

In total, 3,281 exclusive subgraphs were generated by separating all connected subgraphs in the full GRCh38 gene homolog graph. These subgraphs represent 3,281 gene families that varied greatly in size with most having relatively few genes. The larger gene families had important structural features that included hub and bottleneck genes that connected to the entire family through homology. These gene family networks could take different forms containing a single or multiple hubs and bottlenecks (Figure 2A-C). These network structures indicate that genes in the same family can vary greatly in sequence, and help us understand how new genes arose and evolved through sequence changes.

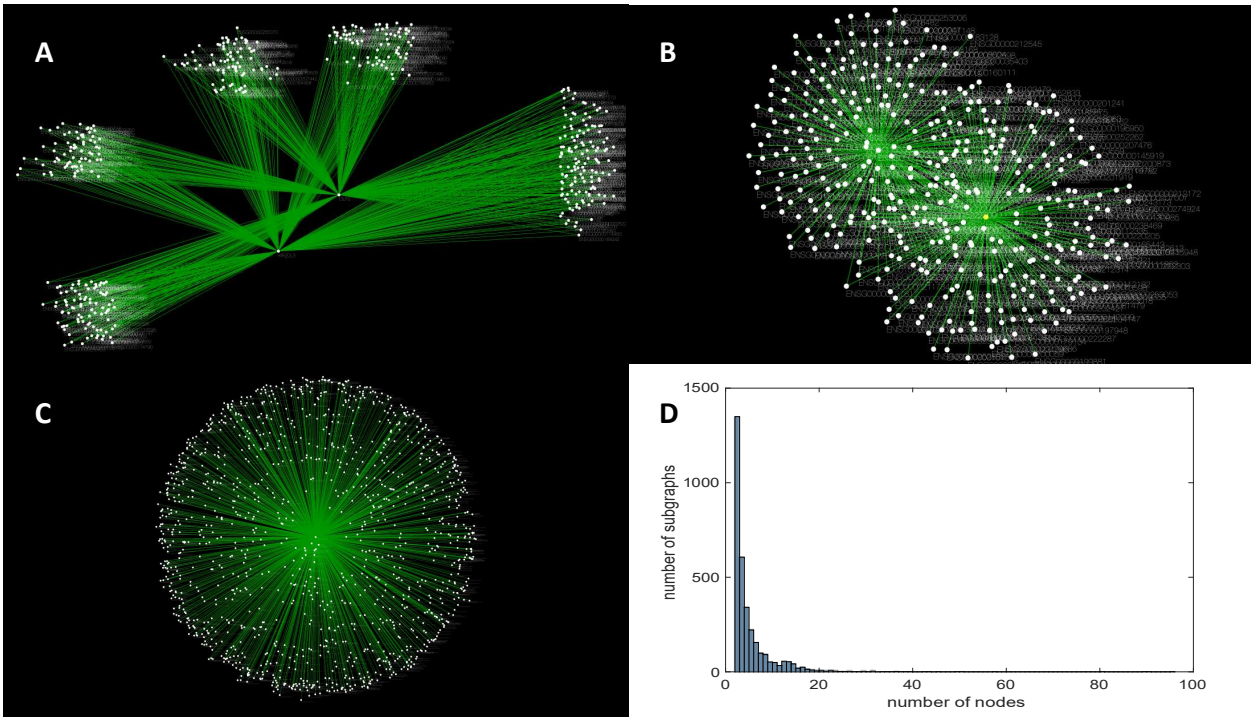


Figure 2. Subgraphs separated from gene homolog network. A: Gene family 6, B: Gene family 18, C: Gene family 32, D: Histogram of gene family sizes. Outliers were removed past 100 nodes so that the distribution of the common (smaller) sizes could be seen.

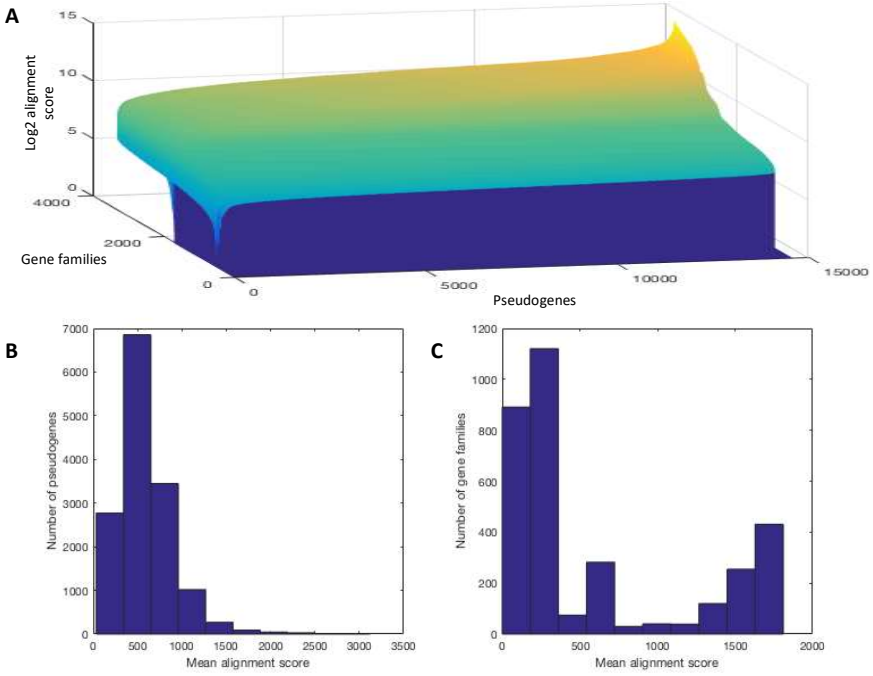


Figure 3. A) The distribution of pairwise pseudogene to consensus gene alignment scores for all pseudogenes and consensus genes (color also signifies Log2 alignment scores, lighter is higher and darker is lower). B) Distribution of mean pseudogene alignment score. C) Distribution of mean gene family alignment score.

3.2 *Mapping pseudogenes to gene homolog families*

Through mapping pseudogenes to gene homolog families, we generated the comprehensive set of pseudogene-gene (PGG) families. Pseudogenes are relatives of genes and other pseudogenes in the same PGG family. The alignment scores between the pseudogenes and the consensus genes representing gene homolog families varied greatly between different pairs (i.e. alignments between pseudogenes and consensus genes) (Figure 3). Some had high conservation of sequences thus sequences are closely aligned with high scores. While some others had negative alignment scores indicating no relationship in sequence – the alignment incurred many more penalties than matches (these pairs would not be combined into PGG families). Pseudogenes were assigned to gene families with the highest alignment score for that pseudogene across all gene families. Thus each pseudogene was assigned to one unique gene family, and each gene family could accept multiple pseudogenes.

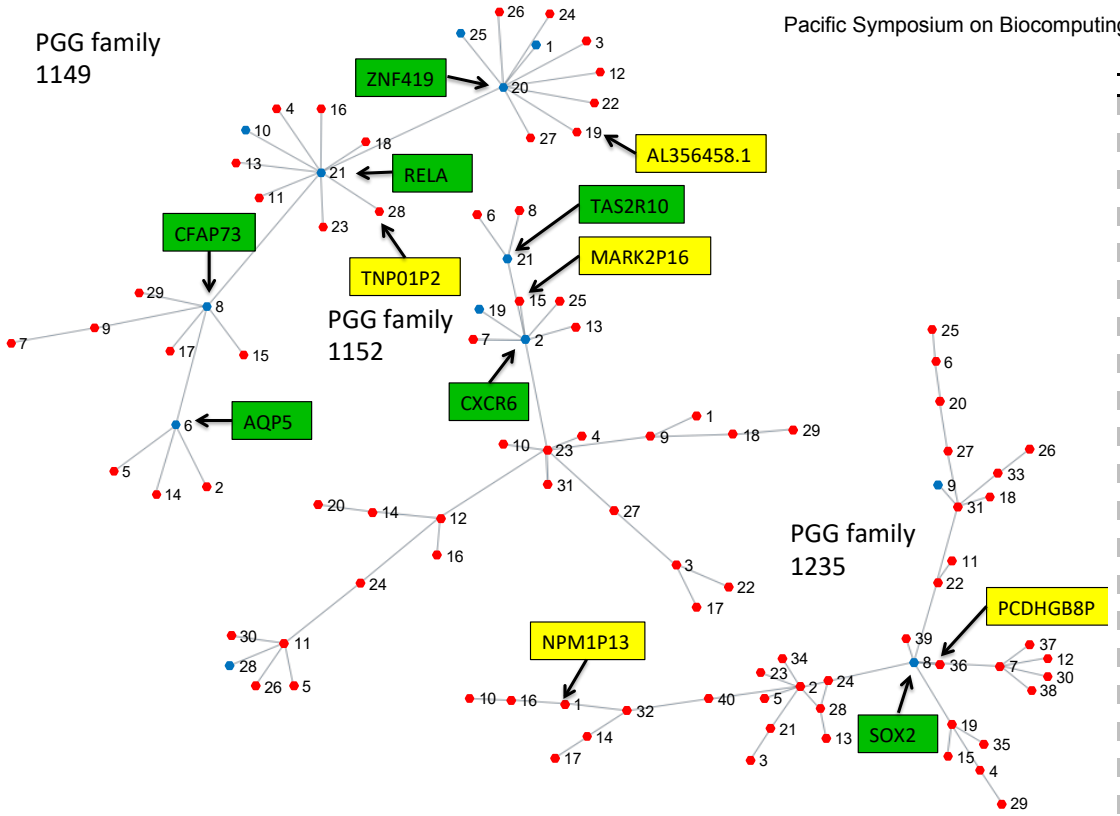
3.3 *Network analysis of individual PGG families*

PGG networks were transformed into MSTs and graphed to view the relationship structure between the genes and pseudogenes in PGG families. The MSTs also highlight the bottlenecks. Three PGG families were selected as examples due to the large number of possible PGG families all of which could not be displayed. We also provide data matrices to generate all PGG family MSTs in Supplementary Materials which can be downloaded from GitHub (https://github.com/yanzhanglab/PGG_DB). Figure 4 shows the MSTs for three PGG families of interest. PGG family 1152 is of interest because it contains multiple genes that are related to chemokine receptors, olfactory receptors, and taste receptors. In this family a large portion of the nodes are pseudogenes, which is consistent with the knowledge of olfactory family reduction in primates with the greatest reduction of olfactory genes in hominoids (e.g. mice have more olfactory genes than primates of which humans have the least)[30,31]. There are even some examples of chemokine receptors becoming pseudogenes in Humans from other primates. An example is CCR5 which has a known pseudogene polymorphism in human that is known for reduced risk of HIV infection in exposed individuals[32,33]. PGG family 1149 is of interest because it contains the proto-oncogene RELA which has been implicated in pseudogene regulatory activity[34,35]. Another family of interest is 1235 that contains SOX2, which has been identified as a member of ceRNA networks[36,37].

The betweenness centrality (BC) of the PGG networks (Figure 5) showed that there were a higher proportion of genes with non-zero BC (17.29%) than pseudogenes (13.82%) with an odds ratio of 1.304 95% CI (1.056, 1.610) and p-value of 0.017. Higher non-zero BC implicates higher importance of the node in the network. After removing nodes with zero BC, it was found that the BC across all genes and pseudogenes was skewed higher in genes (Kolmogorov-Smirnov p-value = 2.856×10^{-8}). These observations implicates that more genes than pseudogenes work as bottlenecks in the networks. Both gene and pseudogene BC followed exponential distributions with λ= 0.081 95% CI (0.075, 0.088) and λ=0.217 95% CI (0.199,0.235) respectively (Figure 5A-B).

3.4 *Functional annotation of PGG families*

Functional annotation of all genes contained in PGG families showed that there was enrichment in olfactory receptor and sensory receptor terms (Table 1). This supports the validity of our method since the most enriched function (Annotation Cluster 1) recapitulated the high number of known pseudogenes related to olfactory and other senses in human (Figure 5). Cluster 3 is also of interest due to the increasing evidence of the regulatory role of pseudogenes. DNA binding could be indicative of some forms of RNA regulation, which is further supported by the over-represented GO terms (GO:0045893:positive regulation of transcription, DNA-dependent and GO:0045944~positive regulation of transcription from RNA polymerase II promoter). Also, there is a growing body of evidence that proteins do not exclusively bind to DNA or RNA[38] and a presence of over-represented ribonucleotide binding GO terms in our DAVID functional enrichment which could be indicative of ceRNA networks in which pseudogenes compete with genes for regulatory binding elements.



Node	PGG family 1149	PGG family 1152	PGG family 1235
1	ENSG00000251763.1	ENST00000447582.1	ENST00000452663.1
2	ENST00000503899.1	ENSG00000172215.5	ENST00000517788.1
3	ENST00000440335.1	ENST00000592975.1	ENST00000570323.1
4	ENST00000432727.1	ENST00000436067.1	ENST00000613638.1
5	ENST00000458502.2	ENST00000457236.1	ENST00000444036.1
6	ENSG00000161798.6	ENST00000523272.1	ENST00000608772.1
7	ENST00000413087.1	ENST00000549485.1	ENST00000436977.1
8	ENSG00000186710.11	ENST00000529746.2	ENSG00000181449.3
9	ENST00000573526.1	ENST00000588007.1	ENSG00000276746.1
10	ENSG00000201749.1	ENST00000461060.1	ENST00000593607.1
11	ENST00000479876.2	ENST00000503971.1	ENST00000447105.1
12	ENST00000511924.1	ENST00000527904.1	ENST00000414751.1
13	ENST00000566323.1	ENST00000612510.1	ENST00000456160.1
14	ENST00000416620.2	ENST00000545069.1	ENST00000435568.1
15	ENST00000510918.5	ENST00000605714.1	ENST00000467018.3
16	ENST00000394467.3	ENST00000511142.1	ENST00000406097.2
17	ENST00000603274.1	ENST00000401540.2	ENST00000431137.1
18	ENST00000411655.1	ENST00000415286.1	ENST00000549314.1
19	ENST00000604764.1	ENSG00000253031.1	ENST00000466609.1
20	ENSG00000105136.19	ENST00000447665.2	ENST00000415181.1
21	ENSG00000173039.18	ENSG00000121318.2	ENST00000510009.1
22	ENST00000582254.1	ENST00000459978.1	ENST00000417699.1
23	ENST00000619819.1	ENST00000445455.1	ENST00000524675.1
24	ENST00000519099.1	ENST00000407015.1	ENST00000605279.1
25	ENSG00000281516.1	ENST00000510646.1	ENST00000532761.1
26	ENST00000372011.4	ENST00000507189.2	ENST00000414395.1
27	ENST00000446719.1	ENST00000421347.2	ENST00000422734.1
28	ENST00000426249.1	ENSG00000200597.1	ENST00000406017.3
29	ENST00000396820.2	ENST00000616818.1	ENST00000391729.1
30		ENST00000558683.2	ENST00000605404.1
31		ENST00000559181.3	ENST00000478870.1
32			ENST00000589292.1
33			ENST00000450305.2
34			ENST00000436499.1
35			ENST00000415292.1
36			ENST00000507007.2
37			ENST00000499125.3
38			ENST00000485242.1
39			ENST00000429392.1
40			ENST00000366220.2

Figure 4. Minimum spanning trees of PGG families 1149, 1152 and 1235 with pseudogenes (red nodes) and genes (blue nodes). Genes of interest (with GO annotation or bottleneck node) are highlighted in green and pseudogenes of interest (with possible functional relationship to gene family) are highlighted in yellow.

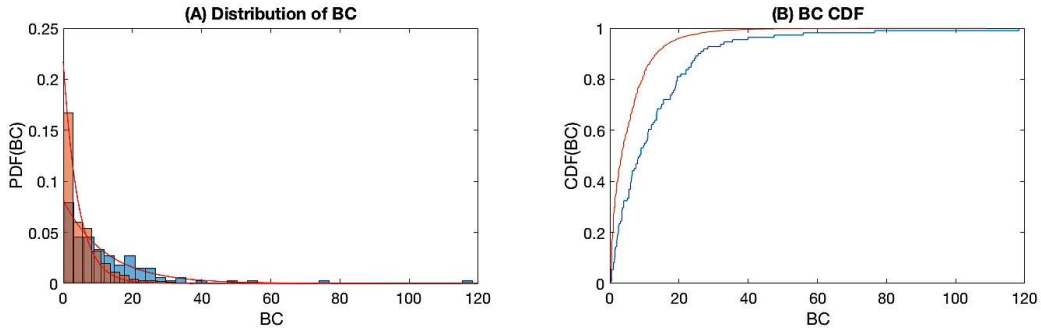


Figure 5. BC of PGG families. A) BC distributions and associated exponential empirical PDF for pseudogenes (red) and genes (blue). B) Empirical CDF for pseudogenes (red) and genes (blue).

We extracted olfactory genes from the first functional annotation cluster using the full DAVID clustering table and labeled PGG networks as Olfactory if they contained at least one of the extracted olfactory genes (27 Olfactory and 418 Not Olfactory). Based on this stratification we found that olfactory related PGG families were more likely to contain both gene and pseudogene bottlenecks – olfactory related PGG families were more likely to have both genes and pseudogenes with non-zero BC (OR = 3.641 95% CI: (1.650,8.035), p-value = 0.002).

Table 1. DAVID functional annotation of the genes contained within PGG families (High stringency).

Category	Term	P-value	Fold Enrichment	FDR
Annotation Cluster 1	Enrichment Score: 12.91			
SP_PIR_KEYWORDS	olfaction	2.89E-16	7.33	4.33E-13
GOTERM_BP_FAT	GO:0007608~sensory perception of smell	4.38E-15	6.45	6.98E-12
INTERPRO	IPR000725:Olfactory receptor	5.28E-15	6.52	7.15E-12
Annotation Cluster 2	Enrichment Score: 12.36			
SP_PIR_KEYWORDS	g-protein coupled receptor	1.69E-19	5.45	2.20E-16
INTERPRO	IPR017452:GPCR, rhodopsin-like superfamily	1.15E-18	5.44	1.54E-15
INTERPRO	IPR000276:7TM GPCR, rhodopsin-like	1.20E-18	5.43	1.61E-15
Annotation Cluster 3	Enrichment Score: 2.64			
SMART	SM00389:HOX	8.08E-04	4.46	0.83
SP_PIR_KEYWORDS	Homeobox	2.42E-03	3.85	3.11
UP_SEQ_FEATURE	DNA-binding region:Homeobox	2.56E-03	4.33	3.61

3.5 *Identifying phylogenetic relationships and conserved regions*

The phylogenetic tree (PGG family 1149) shows that one lineage consists purely of genes while the other consists of both genes and pseudogenes (Figure 6A). Not surprisingly these same genes are the bottlenecks in the MST graph (Figure 4), which could be indicative of the pseudogenes being generated from the bottleneck genes. Another result of note was the lack of conserved regions (CRs) found in the majority of PGG families. PGG family 1149 had an identified CR but PGG families 1152 and 1235 did not have identified CRs. We tested whether this was related to the containment of pseudogenes within PGG families and found that PGG families that have assigned pseudogenes are more likely to contain CRs with an odds ratio of 13.79 95% CI (10.70, 17.78) and p-value 3.169×10^{-95}.

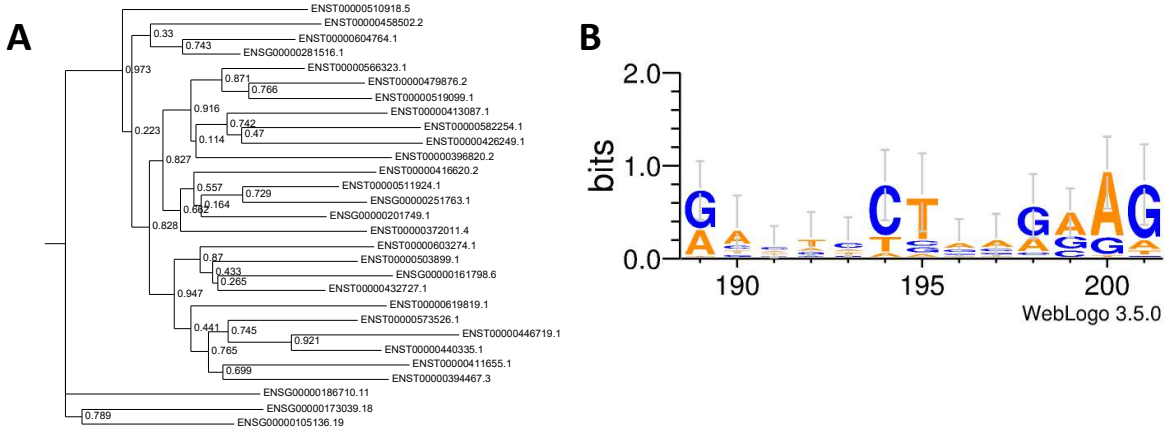


Figure 6. Phylogenetic trees and CRs, A) phylogenetic tree for PGG family 1149, B) CR for PGG family 1149 identified by PhastCons. IDs starting with ENSG constitute genes and ENST constitute pseudogenes.

3.6 *Functional analysis of pseudogenes within PGG networks*

Using BiNGO, we can evaluate the GO terms in each of the PGG networks. From PGG families 1149, 1152, 1235 we produced the following GO term networks (Figure 7), which can be used to evaluate the pseudogene functions included in the networks. PGG family 1149 had significant terms related to neurogeneration (Figure 7A). Pseudogene AL356458.1 is contained in PGG family 1149 and has copy number variations in oral carcinogenesis[39]. Pseudogene TNPO1P2 is also in PGG family 1149 and has implications in neurodegeneration in the frontotemporal lobe[40]. PGG family 1152 included multiple significant sensory related GO terms (Figure 7B). The pseudogene MARK2P16 is present in PGG family 1152 and its related gene MARK2 is needed for the migration of postnatal neuroblasts in the olfactory bulb[41]. PGG family 1235 included both NP1P13 and PCDHGB8P pseudogenes. PCDHGB8P is a protocadherin pseudogene with high sequence homology to other protocaderins such as PCDHGB3 and PCDHGB4 that have implications in multiple forms of cancer. The PGG family 1235 GO network includes significant proliferation terms (Figure 7C). PCDHGB3 and PCDHGB4 have implications in various cancer including lymphoma[42] and PCDHBG4 has implications in metastatic breast cancer[43]. Analysis of Wilms' tumors has shown frequent hypermethelated down-regulation of protocaderins (including PCDHGB4) in the tumor samples. NPM1P13 is implicated in a neurodevelopmental disorder, Saethre-Chotzen syndrome[44]. Significant GO terms in PGG family 1235 related to neurological development are also present (Figure 7C). PGG families 1149, 1152, and 1235 all have interesting functional relationships. These functional relationships (Figure 7) share documented functional similarities to either pseudogene members of the PGG family or the pseudogene members gene counterparts.

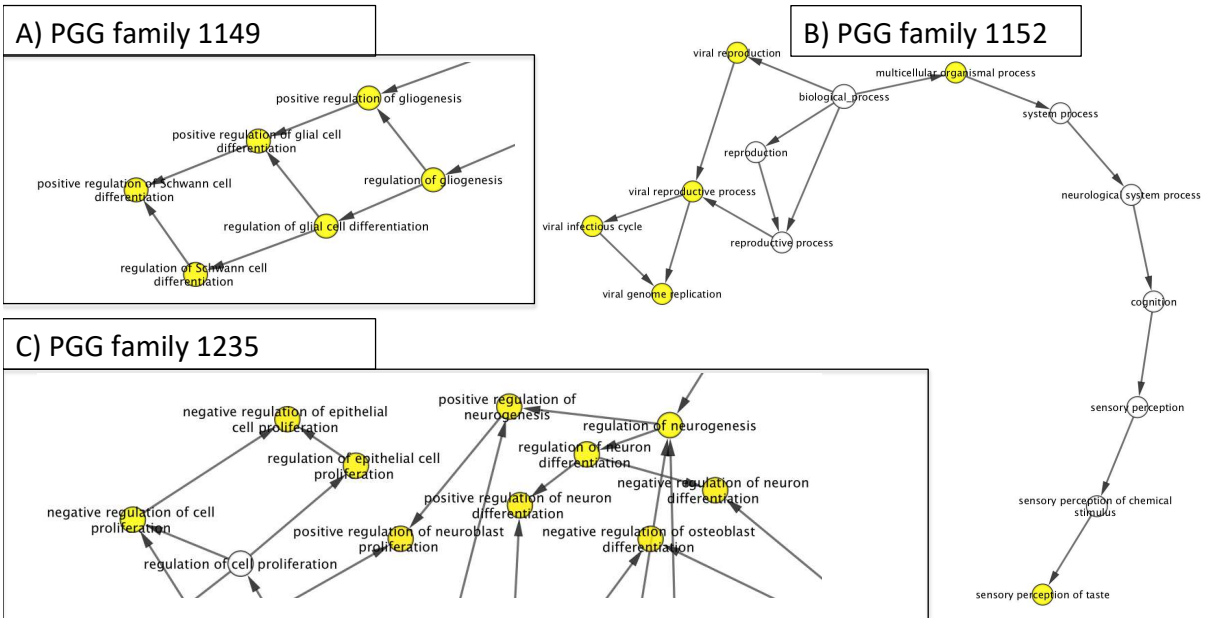


Figure 7. GO term networks from example PGG families. Each node denotes a GO term. Yellow nodes designate GO terms that are significant at p-value= 0.05. A) is a partial GO term BiNGO network for PGG family 1149 highlighting cell differentiation and glial cell development. B) is the full GO term BiNGO network from for PGG family 1152. C) is a partial BiNGO GO term network for PGG family 1235 highlighting cell proliferation GO terms and neuron development.

4 Discussion

4.1 *Insights*

Aside from the evolutionary relationships of pseudogenes to the genome we also find proto-oncogenes in PGG families containing pseudogenes. In PGG family 1149 we observed a network inclusive of a possible proto-oncogene RELA and small regulatory RNAs. RELA over-expression in mice was shown

to delay the appearance of tumors and reduce proliferation *in vitro*[45]. The mapping of large numbers of pseudogenes to the RELA homolog family supports a possible regulatory relationship. Within these pseudogenes we find TNPO1P2 and AL356458.1 with some evidence in the literature of possible relationships to the BiNGO functional networks generated from the PGG families. Aside from RELA, another PGG family 1235 contained SOX2, a gene that has been implicated in ceRNA networks[7]. Within this network we identify PCDHGB8P and NPM1P13 that have literature supporting potential functions related to those identified in the BiNGO functional network. These findings support the hypothesis that pseudogenes may play a regulatory role in the genome, and the networks of interest are worth further investigation. Within the olfactory related PGG network 1152 we find olfactory related function in both the BiNGO functional network and literature supporting olfactory function for MARK2 the gene precursor to MARK2P16 pseudogene.

Genes and pseudogenes with a high non-zero BC are important to the structure of the PGG network. Directly it means these genes/pseudogenes constitute bridges between more dissimilar sequences within a PGG network. Biologically this could imply that that a gene/pseudogene likely contains key mutation signatures that triggers the change of function (e.g. silencing an ancient gene, or re-activating a pseudogene) or contributes to the gene/pseudogene family expansion (i.e. generating large number of descendants in the PGG network). The distribution of non-zero BC for genes and pseudogenes were also altered where genes tend to have higher BC values. Evolutionarily this could imply that genes are more likely to be bottlenecks in PGG families and are more important bottlenecks than pseudogenes.

We also find that gene families that were assigned pseudogenes were more likely to contain CRs. This could be examined further to evaluate what function this conservation could have. This difference in enrichment level of CRs may be related to family size, or biased by the sequence similarity threshold defined by computational tools used for identifying pseudogenes.

Of the gene families with many aligned pseudogenes, there was enriched annotation of olfactory receptor genes, as shown in DAVID results. This is in congruence with previous findings that the olfactory receptor family in humans has large numbers of pseudogenes[46,47].

Another important note is that we generate PGG families through alignment of pseudogenes to gene families. Especially in the case of processed pseudogenes this unbiased approach should be taken into account when examining the potential of competing endogenous RNA. Since the sequences by definition must closely related to the assigned gene family, many of these pseudogenes could be candidates for ceRNA networks and evaluated further.

In the following work we will integrate all of the aspects of this project into an online query tool, which will return PGG families and functional prediction for specific novel pseudogene sequences of interest. The functional enrichment included in existing GO annotation tools (e.g. BiNGO) do not take into account the proximity of pseudogenes and genes in the same network. In the future we will try to overlay these weights in our GO inference method to improve the functional predictions given by our tool. Aside from the user interface and refined functional prediction, the tool itself due to its network structure could easily be integrated with other experimental methodologies (e.g. ChiP-Seq and Competition-ChiP) paired by gene/pseudogene IDs. Our goal is to use our current database as a baseline functional prediction for pseudogenes that can be easily augmented by the improved methodologies both currently available and developed in the future. This makes our database immensely scalable as new and improved features are added to aid in functional prediction of pseudogenes.

4.2 *Limitations*

One limitation affecting this approach includes the ambiguous definition of pseudogenes in available annotations. There are DNA segments annotated as genes but do not code proteins. There are also gene-like DNAs generating regulatory RNAs and are not annotated as pseudogenes. This ambiguity in annotation could introduce noise to the alignment steps.

In our current approach, we treat different pseudogene biotypes (processed, duplicated, or unitary pseudogenes) uniformly. The full gene sequences were used to make this approach more computationally tractable. Because there were intronic regions in gene homologs, whereas processed pseudogenes do not

contain introns, the alignments of gene families to processed pseudogenes are not as accurate as aligning to duplicated or unitary pseudogenes. However, the effects of using full length gene sequences is mitigated by the use of local alignment. In our future studies, we will fine-tune our approach to treat different pseudogene biotypes respectively.

A large portion of the genes within the PGG families did not have associated functional annotation within DAVID, which implies annotation bias where not all genes are equally well-studied. Because the number of genes in some networks was small, some families had few genes to study functional enrichment from.

5 Conclusion

In this study, we investigate the functional relationships between pseudogene and gene homolog families, by integrating graph analysis, sequence alignment and functional analysis, and generate the comprehensive set of pseudogene-gene families in human. By studying the network structure of these pseudogene-gene networks, we find that there is an over-representation of olfaction related PGG families, differential BC between genes and pseudogenes, and structural patterns that can be used to differentiate PGG networks. These patterns in network structure also can be used to differentiate different classes of networks. Olfactory PGG families were associated with a network structure in which both genes and pseudogenes had bottleneck qualities (measurable BC). Similarly we view these networks as important tools in predicting function for pseudogenes, similar to previous methods that infer gene ontologies for under-annotated genes[48]. We use our PGG families to associate GO terms to under-documented pseudogenes and describe the utility of this database to query new pseudogene sequences to infer functional potential. In summary, here we have proposed a novel, comprehensive, and scalable evaluation of pseudogenes at the gene homolog level and showed that network structure can be related to functionality.

We also propose in future work a refined pipeline that i) treat different pseudogene biotypes respectively, ii) preprocesses the gene annotation prior to analysis to reduce ambiguity, iii) can identify possible ceRNA networks algorithmically, iv) provide a search utility to query novel pseudogene sequences against our network database to predict pseudogene function and v) use network weights to more accurately associate known and novel pseudogene sequences to GO terms within the assigned PGG family. Aside from these immediate improvements that are under development currently we also plan to make this database scalable in the different types of data that can be integrated (e.g. ChiP-Seq and Competition-ChiP) via pairing nodes via gene names or interactions.

6 Acknowledgements

The work is supported by NIH-NLM MIDAs Training Fellowship (4T15LM011270-05) to Travis Johnson, and The Ohio State University Startup Funds to Yan Zhang. The authors also thank the Ohio Supercomputer Center for providing computing resources.

References

1. Jacq C, Miller JR, Brownlee GG. A pseudogene structure in 5S DNA of Xenopus laevis. *Cell.* 1977;12(1):109-120.
2. Zhang ZZ, Deyou. Pseudogene evolution in the Human Genome. *eLS.* 2014.
3. Poliseno L. Pseudogenes: newly discovered players in human cancer. *Sci Signal.* 2012;5(242):re5.
4. Cooke SL, Shlien A, Marshall J, et al. Processed pseudogenes acquired somatically during cancer development. *Nat Commun.* 2014;5:3644.
5. Sisu C, Pei B, Leng J, et al. Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A.* 2014;111(37):13361-13366.
6. Poliseno L, Marranci A, Pandolfi PP. Pseudogenes in Human Cancer. *Front Med (Lausanne).* 2015;2:68.
7. Cheng DL, Xiang YY, Ji LJ, Lu XJ. Competing endogenous RNA interplay in cancer: mechanism, methodology, and perspectives. *Tumour Biol.* 2015;36(2):479-488.
8. Poliseno L, Pandolfi PP. PTEN ceRNA networks in human cancer. *Methods.* 2015;77-78:41-50.
9. Sanchez-Mejias A, Tay Y. Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics. *J Hematol Oncol.* 2015;8:30.
10. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature.* 2010;465(7301):1033-1038.
11. Tay Y, Kats L, Salmena L, et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell.* 2011;147(2):344-357.

12. Zhang Y, Li S, Abyzov A, Gerstein MB. Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput Biol.* 2017;13(6):e1005567.
13. Karro JE, Yan Y, Zheng D, et al. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* 2007;35(Database issue):D55-60.
14. Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 2003;13(12):2541-2558.
15. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DR. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA.* 2011;17(5):792-798.
16. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22(9):1760-1774.
17. Camiolo S, Porceddu A. gff2sequence, a new user friendly tool for the generation of genomic sequences. *BioData Min.* 2013;6(1):15.
18. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;7(3):562-578.
19. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007;23(21):2947-2948.
20. Beynon Michael D. KT, Catalyurek Umit, Chang Chialin, Sussman Alan, Saltz Joel. Distributed processing of very large datasets with DataCutter. *Parallel Computing.* 2001;27(11):1457-1478.
21. Chirag Jain SK. Fine-grained GPU parallelization of pairwise local sequence alignment. 21st International Conference on High Performance Computing (HiPC; 2014.
22. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44-57.
23. Huang DW, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 2007;35(Web Server issue):W169-175.
24. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792-1797.
25. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26(7):1641-1650.
26. Siepel A, Haussler D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol.* 2004;11(2-3):413-428.
27. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics.* 2005;21(16):3448-3449.
28. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.
29. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics.* 2011;27(3):431-432.
30. Gilad Y, Przeworski M, Lancet D. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol.* 2004;2(1):E5.
31. Rouquier S, Blancher A, Giorgi D. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc Natl Acad Sci U S A.* 2000;97(6):2870-2874.
32. Dean M, Carrington M, Winkler C, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science.* 1996;273(5283):1856-1862.
33. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.* 2010;11(3):R26.
34. Porter KA, Duffy EB, Nyland P, Atianand MK, Sharifi H, Harton JA. The CLRX.1/NOD24 (NLRP2P) pseudogene codes a functional negative regulator of NF-kappaB, pyrin-only protein 4. *Genes Immun.* 2014;15(6):392-403.
35. Rapicavoli NA, Qu K, Zhang J, Mikhail M, Laberge RM, Chang HY. A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. *Elife.* 2013;2:e00762.
36. Arancio W, Carina V, Pizzolanti G, et al. Anaplastic Thyroid Carcinoma: A ceRNA Analysis Pointed to a Crosstalk between SOX2, TP53, and microRNA Biogenesis. *Int J Endocrinol.* 2015;2015:439370.
37. Xu J, Feng L, Han Z, et al. Extensive ceRNA-ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues. *Nucleic Acids Res.* 2016.
38. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol.* 2014;15(11):749-760.
39. Vincent-Chong VK, Salahshourifar I, Razali R, Anwar A, Zain RB. Immortalization of epithelial cells in oral carcinogenesis as revealed by genome-wide array comparative genomic hybridization: A meta-analysis. *Head Neck.* 2016;38 Suppl 1:E783-797.
40. Troakes C, Hortobagyi T, Vance C, Al-Sarraj S, Rogelj B, Shaw CE. Transportin 1 colocalization with Fused in Sarcoma (FUS) inclusions is not characteristic for amyotrophic lateral sclerosis-FUS confirming disrupted nuclear import of mutant FUS and distinguishing it from frontotemporal lobar degeneration with FUS inclusions. *Neuropathol Appl Neurobiol.* 2013;39(5):553-561.
41. Mejia-Gervacio S, Murray K, Sapir T, Belvindrah R, Reiner O, Lledo PM. MARK2/Par-1 guides the directionality of neuroblasts migrating to the olfactory bulb. *Mol Cell Neurosci.* 2012;49(2):97-103.
42. Takata K, Tanino M, Ennishi D, et al. Duodenal follicular lymphoma: comprehensive gene expression analysis with insights into pathogenesis. *Cancer Sci.* 2014;105(5):608-615.
43. Shima J, Delaney J, Umesh A, et al. Disruption of protocadherin function and correlation with metastasis and cancer progression in TCGA patients. *Journal of Clinical Oncology.* 2012;30(30_suppl):70-70.
44. Shimbo H, Oyoshi T, Kurosawa K. Contiguous gene deletion neighboring TWIST1 identified in a patient with Saethre-Chotzen syndrome associated with neurodevelopmental delay: Possible contribution of HDAC9. *Congenit Anom (Kyoto).* 2017.
45. Ricca A, Biroccio A, Trisciuoglio D, Cippitelli M, Zupi G, Del Bufalo D. relA over-expression reduces tumorigenicity and activates apoptosis in human cancer cells. *Br J Cancer.* 2001;85(12):1914-1921.
46. Niimura Y. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. *Curr Genomics.* 2012;13(2):103-114.
47. Prieto-Godino LL, Rytz R, Bargeton B, et al. Olfactory receptor pseudo-pseudogenes. *Nature.* 2016.
48. Dutkowski J, Kramer M, Surma MA, et al. A gene ontology inferred from molecular networks. *Nat Biotechnol.* 2013;31(1):38-45.

Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 Diabetes GWAS

Elisabetta Manduchi

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA and Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: manduchi@pennmedicine.upenn.edu

Alessandra Chesi

Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: chesia@email.chop.edu

Molly A. Hall

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: hallma@mail.med.upenn.edu

Struan F.A. Grant

Division of Human Genetics and Endocrinology, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard, Philadelphia, PA 19104, USA
Email: grants@email.chop.edu

Jason H. Moore*

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA
Email: jhmoore@exchange.upenn.edu

We utilized evidence for enhancer-promoter interactions from functional genomics data in order to build biological filters to narrow down the search space for two-way Single Nucleotide Polymorphism (SNP) interactions in Type 2 Diabetes (T2D) Genome Wide Association Studies (GWAS). This has led us to the identification of a reproducible statistically significant SNP pair associated with T2D. As more functional genomics data are being generated that can help identify

* Correspondence

potentially interacting enhancer-promoter pairs in larger collection of tissues/cells, this approach has implications for investigation of epistasis from GWAS in general.

Keywords: epistasis; GWAS; type 2 diabetes; enhancers; genetic encoding.

1. Introduction

In the context of GWAS, epistasis refers to SNP interactions associated with a particular complex trait. There are different schools of thought regarding the role of epistasis in human genetics. The majority of human GWAS to date have focused on detecting main effects, i.e. single SNP associations. This is naturally the first aspect to explore, and many would argue that this is the most relevant since the vast majority of SNP contributions to a given trait are largely additive.[1] However, a different opinion is that epistasis is a non-negligible component of human genetic architecture, possibly accounting for the predicted 'missing heritability',[2-3] based on the observations that biological systems are regulated by complex biomolecular networks and epistasis plays an important role in model organisms.[4]

Exploring epistasis in a typical GWAS is complicated by several factors. One major challenge arises from the large search space and subsequent limited computational and statistical power. Genes and regulatory elements typically form complex networks, thus epistatic interactions could well involve several SNPs. But, even if one wanted to focus on two-way interactions on genotyped SNPs, a typical GWAS involves at least half a million such markers, so the number of possible pairs is greater than 10^{11}. Another difficulty is that different SNPs relevant to a complex trait may have different mechanisms of action so the choice of genetic encoding affects interaction detection (Hall *et al.*, in preparation). Given all of this, it is therefore uncommon to find statistically significant epistatic interactions in a GWAS data set. Moreover, it is even less common for such observations to be reproduced in a replication data set.

In this work we illustrate an example where the combination of suitable biological filters and a data-driven weighted encoding approach has led us to a statistically significant pairwise interaction in a T2D discovery data set which we were able to go on to replicate in an independent data set. Statistical epistasis is different from biological epistasis[5-7] so establishing whether or not this pair corresponds to an actual biological mechanism associated with T2D will require additional experimentation. However we are using this example to highlight a possible avenue of epistasis investigation which exploits the increasing availability of functional genomics data sets aimed at exploring regulatory and physical interactions among genomic features, such as ChIP-Seq or high-throughput chromatin capture data sets. Our approach is illustrated in the next section.

2. Filters and Encodings

Figure 1 outlines our workflow, which involves two main components:

2.1. *Defining biological filters based on functional genomics data*

Due to the large search space, the first step in an epistasis analysis is to reduce the number of models (i.e. candidate interacting SNP sets) to analyze. To this end, both computational and biological filter approaches have been previously proposed in the literature.

Computational approaches include methods such as ReliefF and its derivatives,[8-10] MDR,[11] and "greedy" approaches which first identify SNPs with main effects (significant or marginally significant) in a GWAS and then use models involving only these SNPs.[12-13] The latter is certainly a reasonable approach; however it will miss potential interactions which involve SNPs with no main effects (i.e. it will miss what is referred to as 'pure and strict epistasis'[14]).

Biological filters may exploit biological annotations (derived from curation of low or high-throughput experiments) or analyses of functional genomics data sets to reduce the search space. For example, knowledge about the biological relevance of the Ras/MAPK pathway to Autism Spectrum Disorders has been used to limit the search space of SNP pairs analyzed for interactions to those where one of the SNPs is in a Ras/MAPK pathway gene.[15] In this work we too use biological filters, but of a different type as described below.

We sought to exploit the increasing availability of functional genomics data sets elucidating genomic features with likely regulatory functions in different tissues and cell lines. This was motivated by the recognized importance of regulatory networks in genomic studies.[16-17] The regulation of gene expression is complex, but a fundamental component lies in enhancers, i.e. non-coding regions in the genome which may affect the expression of distal genes through chromatin looping. We reasoned that natural candidates for two-way interactions are SNP pairs from interacting enhancer-promoter regions in tissues or cell lines relevant to the trait being studied. Based on this, we have selected appropriate interacting regions from the EnhancerAtlas.[18] This resource provides collections of enhancer-gene interactions for several tissues and cell lines, derived from the integration of almost 4000 high throughput experimental data sets from resources including the UCSC genome browser,[19] NCBI GEO,[20] Cistrome database,[21] ENCODE project data portal,[22] Epigenome Roadmap data portal[23] and eRNA.[24]

2.2. *Using weighted encoding*

When we consider a single SNP, we want to encode the biological action in the way that it likely functions, so if a genotype has no alternate alleles, the risk would be 0 and if it has two such alleles the risk would be 1. For a heterozygous genotype, coding it as recessive assumes it has no risk (equal to homozygous referent), coding it as dominant assumes it yields full risk (the same as two alternate alleles), and coding it as additive is right in between. Yet in biology, a heterozygous genotype may act anywhere in this range from recessive to dominant. This has been heavily discussed in the literature for single SNP associations and the consensus has been that additive encoding will capture the largest amount of genetic effects.[25]

PLATO software[26] (https://ritchielab.psu.edu/files/RL_software/plato-manual-2.1.pdf) allows for different choices of encoding. One of them is a data-driven approach to compute an appropriate SNP-specific encoding weight for the heterozygous genotype. In order to describe the latter, we first need to define what is meant by 'codominant' encoding. As described in the

PLATO software manual, in this encoding each marker uses two variables as a dummy encoding; the "Het" variable is 1 only when the marker is heterozygous, and the "Hom" variable is 1 only when the marker is homozygous alternate. In weighted encoding, for each marker, the result from a univariate model (with appropriate covariates) is used to determine an encoding from marker state to the set {0, x, 1}, where x is chosen such that the model with the encoded allele is identical to the codominant model. Data-driven weighted encoding was tested on simulated data sets spanning a comprehensive array of underlying interactions of genetic models, concluding that it had a better performance than the other encodings based on a combination of power and type I error (Hall *et al.*, in preparation).

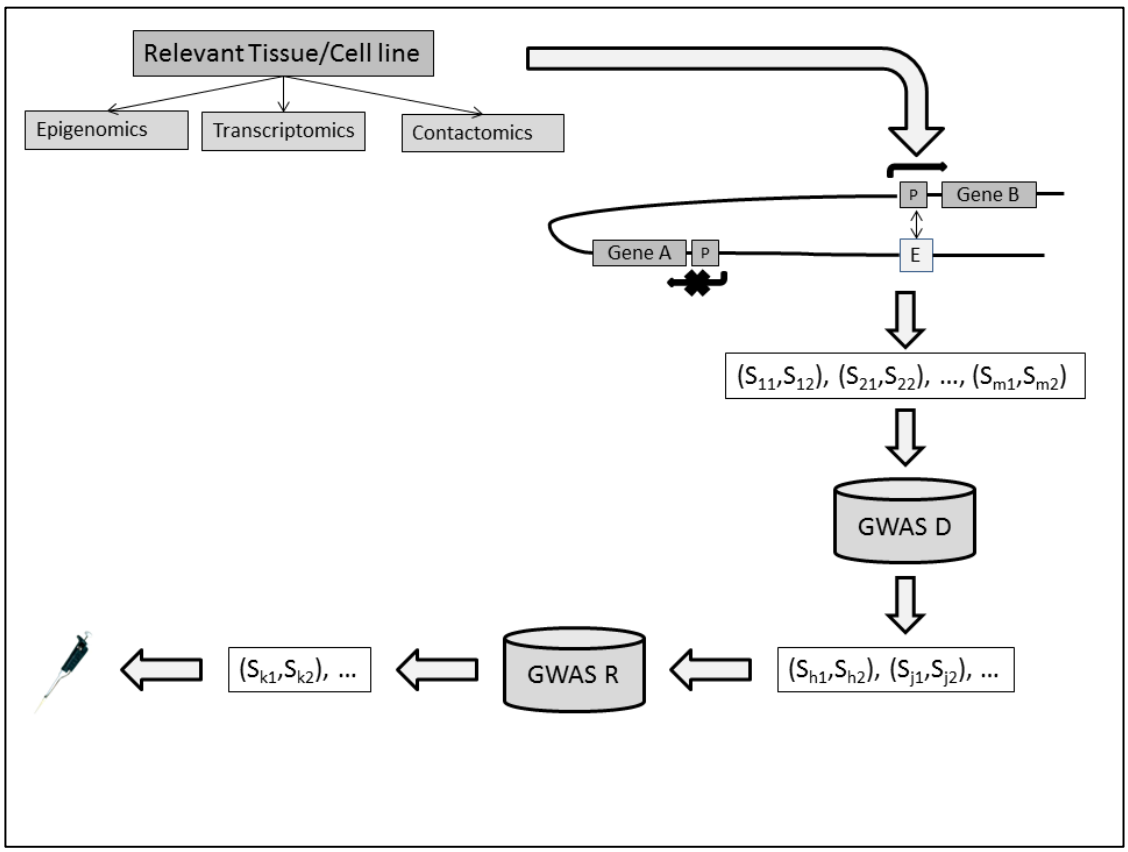


Figure 1. Functional genomics data sets of different kinds, for tissues or cells relevant to the trait of interest, are used to identify putative active and interacting enhancer-promoter pairs. Pairs of SNPs harbored in these interacting regions are extracted and analyzed for epistasis in a discovery GWAS. Significant pairs from this analysis are then examined in one or more replication GWAS to identify candidates for subsequent follow-up work.

3. Methods

3.1. *GWAS data sets*

In this work we utilized three T2D GWAS data sets. As discovery data set, we used GWAS data generated by the Welcome Trust Case Control Consortium (WTCCC),[27] precisely derived from the data sets EGAD00000000009 and EGAD00000000021.

As replication data sets we used two GWAS studies from the database of Genotypes and Phenotypes (dbGaP).[28]

1. The GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study (NHS)/Health Professionals Follow-up Study (HPFS)). Data were downloaded from the dbGaP web site, under phs000091.v2.p1 **(**https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1).
2. The Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Data were downloaded from the dbGaP web site, under phs000867.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000867.v1.p1)**.**

3.2. *Data pre-processing*

3.2.1. *WTCCC data set*

Quality Control (QC) followed standard recommendations.[29] PLINK 1.9 (https://www.cog-genomics.org/plink2/) was used to filter subjects based on ambiguous gender assignment and missing call rate (threshold=95%) and duplicated individuals where removed using a PI_HAT threshold of 0.8. Then SNPs were filtered based on missing call rate (threshold=95%), Hardy-Weinberg Equilibrium tests (HWE, p=0.00001) and Minor Allele Frequency (MAF, threshold=0.01). Ambiguous SNPs (A/T or G/C) were removed. After QC we had data for 4916 subjects (1960 cases and 2956 controls) over 341,531 SNPs. These data were used to compute the first 10 Principal Components (PC) using PLINK v1.9, after Linkage Disequilibrium (LD) pruning.

Phasing was performed with SHAPEIT[30] and imputation with IMPUTE2[31] using 1000 Genomes phase 1 version 3 (http://www.internationalgenome.org) as reference panel. For imputed SNPs with information score >0.70, genotype was assigned according to the best probability if this was >0.90.

3.2.2. *GENEVA data set*

QC was first separately applied to each of the two panels (NHS and HPFS) and followed standard recommendations.[29] In each case PLINK 1.9 was used to filter subjects based on ambiguous gender assignment and missing call rate (threshold=95%). Then SNPs were filtered based on missing call rate (threshold=95%), differential missing call rates between cases and controls (p=0.00001) and MAF (threshold=0.05). The resulting data from the two panels were then merged and SNP filtering was applied again as above, finally subjects were filtered again based on missing call rate. This resulted in a QC-ed data set for 5485 subjects (2524 cases and 2961

controls) over 656,226 SNPs. The first 10 PCs were computed as for WTCCC, after LD pruning, with PLINK v1.9.

We did not impute this data set as it was only used to verify two SNP pairs resulting from the analyses in the discovery data set. If a SNP in a pair was not genotyped we used a proxy obtained from HaploReg,[32] selecting the proxy as a genotyped SNP having the highest r^2 with the SNP in our pair. For each pair to verify, we extracted the data corresponding to that pair and filtered individuals with missing genotypes on those two SNPs.

3.2.3. *FUSION data set*

QC was performed similarly to the GENEVA data set, yielding a data set of 1706 subjects (919 cases and 787 controls), with over 301,195 SNPs. PCs were also computed as above.

3.3. *Candidate pairs selection*

Among the cell lines and tissues for which enhancer-gene interactions were available in EnhancerAtlas at the time of these analyses, HCT116 and pancreas were the most relevant to T2D. For each of these two biological sources we proceeded as follows to identify candidate SNP pairs for interaction analyses.

The files from EnhancerAtlas link enhancers to ENSEMBL (ensemblgenomes.org) transcripts. We identified the regions spanning from 1000bp upstream to 500bp downstream of the Transcription Start Site (TSS) of each transcript as its promoter region. We extracted all SNPs in the enhancer and resulting promoter regions using Biofilter[33-34] (https://ritchielab.psu.edu/software/biofilter-download-1). Of these SNPs, we retained those which were either genotyped or imputed with an information score >0.70 and a probability >0.90 in the discovery data set.

We then performed LD pruning separately within each enhancer and within each promoter using a 0.8 threshold for r^2. For each enhancer-gene interaction from EnhancerAtlas we then paired up each resulting SNP in the enhancer with each resulting SNP in the corresponding promoter and removed all pairs where the two SNPs had an $r^2>0.6$. After this processing we had 11,395 pairs for HCT116 and 1,220 for pancreas.

3.4. *Two-way interaction analyses*

We analyzed separately the HCT116 and the pancreas pair collections for interactions using our discovery data set. To this end we utilized the PLATO software mentioned above, with data-driven weighted encoding and logistic regression, adjusting for gender and the first 10 PCs. This adjustment was applied after assessing the association between these covariates and the phenotype in the discovery data set. PLATO computes p-values for interactions based on the Likelihood Ratio Test (LRT) between the full model ($\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 +$ covariates) and the reduced model ($\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +$ covariates), where π is the probability of Y=1 for

a binary outcome Y and X_i is the encoded genotype at SNP_i i= 1,2. PLATO outputs unadjusted p-values, as well as Bonferroni and FDR multiple-testing adjusted p-values.

For the GENEVA replication analyses we only tested the pairs detected as significant in the discovery data set. Again we used PLATO to run the interaction analyses and we adjusted for the 10 PCs, but needed not to adjust for gender. This was determined after assessing the association between gender and the first 10 PCs with the phenotype in this data set.

For the FUSION replication analyses we tested the pair that was detected as significant in the discovery data set and the GENEVA data set, proceeding as above but adjusting for gender and the first 3 PCs. This was again determined after assessing the association between gender and the first 10 PCs with the phenotype in this data set.

4. Results

The PLATO analysis on the SNP pairs derived from the pancreas regulation data did not yield any significant pair. The PLATO analysis on the pairs derived from HCT116 yielded two pairs with FDR <10%: the pair (rs1474445, rs78495961) on chr7 (p-value=1.00394e-05; FDR adjusted p-value=0 .082295) and the pair (rs8008440, rs12882535) on chr14 (p-value= 1.44441e-05; FDR adjusted p-value=0 .082295).

We followed up only these two pairs in the GENEVA replication data set. We used rs7777433 as a proxy for rs78495961 (r^2=0.92) and rs8007341 as a proxy for rs12882535 (r^2=0.99), since we did not have genotypes for these two SNPs in GENEVA. The first pair (rs1474445, rs7777433) did not have a significant p-value in GENEVA (p=0.682361), whereas the second pair (rs8008440, rs8007341) was significant (p-value= 0.0181457; Bonferroni adjusted p-value <0.04).

Finally we followed up in the FUSION data set the pair (rs8008440, rs12882535) replicated in the GENEVA data set. The FUSION p-value was higher in this case, but remained borderline significant (p=0.155562). The combined p-value across the three data sets using the 'sumz' method in the 'metap' R package (https://cran.r-project.org/web/packages/metap/), with weights proportional to the square root of the sample sizes was 4.199818e-06, which is significant at the 0.05 level after Bonferroni correction by the number of HCT116 SNP pairs (11,395). The regression coefficients of the full model for each data set are reported in Table 1.

The two SNPs forming the pair are not in LD (r^2<0.01). One SNP (rs12882535) is in the promoter of the *OR6S1* gene while the other (rs8008440) is in a predicted EnhancerAtlas HCT116 enhancer (for *OR6S1*) located within an intron of *ANG*.

Table 1. Regression coefficients in the full model for each of the three data sets. The un-adjusted LRT p value (full model versus reduced model) and sumz combined p value are also reported. For the GENEVA data set rs8007341 was used as a proxy for SNP1.

	SNP1	SNP2	β_SNP1	β_SNP2	β_SNP1×SNP2	LRT p	comb p
WTCCC	rs12882535	rs8008440	0.053	-0.923	-7.672	1.44E-05	4.20E-06
GENEVA			0.001	0.465	-1.152	0.018	
FUSION			0.131	100.158	-199.465	0.156	

5. Discussion

In this work we have utilized functional genomics-based filters to identify candidate SNP pairs to be analyzed for epistatic interactions associated with T2D, based on GWAS data. This led to the two significant pairs (FDR<0.1) in our discovery data set. We followed up these two pairs in a replication data set and one of them (rs8008440, rs12882535) remained significant. It is typically difficult to replicate epistasis results at the SNP level, so this was encouraging. We also looked at this pair in a third data set. Here the pair did not quite reach significance, but its p-value remained relatively low. This third data set had half the number of subjects as the other two and the MAFs of the two SNPs were somewhat different (0.47 and 0.08 in this data set versus 0.45 and 0.12 in the previous two). This may explain the increase in p-value, especially considering that detection of statistical epistasis is very sensitive to MAF.[35] Overall, the meta-analysis adjusted combined p-value across the three data sets is significant. In both the GENEVA and FUSION data sets the regression coefficients of the full model correspond to an antagonistic type of interaction. However, considering that we used data-driven weighted encodings, comparisons of these coefficients across models are hard to interpret, especially when interaction terms are involved.

The pair we identified consists of a SNP in the promoter of *OR6S1* and a SNP in a putative enhancer for this gene based on HCT116 data. *OR6S1* belongs to the family of G-Protein-Coupled Receptors (GPCRs). Besides the EnhancerAtlas evidence supporting its expression in HCT116, expression of *OR6S1* was detected both in colon and pancreas in various microarray and RNAseq surveys reported by GeneCards (www.genecards.org). The presence and role of taste and olfactory receptors in the gut has been discussed in recent papers,[36-38] which indicate the importance of these chemosensors in detecting luminal contents and inducing the modulation of systemic metabolism, including glucose homeostasis. Indeed some GPCRs have recently received attention as new therapeutic targets for type 2 diabetes.[39]

HCT116 is a human cell line derived from colonic carcinoma, which has been used in other studies on T2D[40] and, together with pancreas, was the most relevant for T2D among the sources for which EnhancerAtlas provided enhancer-gene interaction data at the time of our download. We did not detect significant interactions among SNP pairs derived from the pancreas data.

The T2D statistical epistasis association that we found is consistent with the recent literature support for the relevance of GPCRs to T2D. Nevertheless, ascertaining whether or not this pair represents actual biological epistasis requires follow-up experiments. This work should be taken as a proof of concept. One of the limitations of our study was that albeit whole pancreas and HCT116 are relevant tissues/cells for T2D, they are not ideal. Having enhancer-promoter data from biosources such as the pancreatic beta-cell and the enteroendocrine L cell would improve the chance to detect epistatic pairs associated with T2D. As more data become available for these particular biosources, both in the public domain and in our labs, we can generate richer and more specific sets of candidate pairs to explore. The types of relevant data include epigenetic information (e.g. open chromatin, enhancer and promoter histone marks) and gene expression data. We are working on a generalization of our workflow which incorporates relevant

epigenomics data as well as high-throughput chromatin conformation capture data (contactomics) to identify additional tissue-specific active and interacting genomic regions from which to select SNP pairs (Manduchi E., Grant S., Moore J., in preparation) .

Our intent is to exploit the availability of different types of omics data sets and to incorporate state of the art analysis methods into a reasonable workflow to explore a typically difficult-to-detect phenomenon such epistasis. Based on our results we believe that this type of workflow is promising and besides significantly reducing the search space it could yield results that are much easier to interpret. Since it is applicable to any GWAS, as we build on data sets which are more extensive in terms of type of functional genomics data and trait-specific tissues, it has the potential to unveil interactions associated to many traits, with a higher likelihood of being reproducible and biologically meaningful.

In this work we were interested in assessing reproducibility and hence utilized a Discovery/Replication data set paradigm. An alternative way to take advantage of independent GWAS for a trait is to combine them in a meta-analysis framework, which can help increase power. We are investigating this approach also in the context of exploring results at different resolutions, besides SNP-pair level (Manduchi E., Grant S., Moore J., in preparation).

6. Availability of Data and Materials

The GWAS data sets analyzed in study are available, upon application, at: https://www.wtccc.org.uk/ (WTCCC, use the data set IDs provided in the main text), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000091.v2.p1 (GENEVA), and https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000867.v1.p1 (FUSION).

7. Funding

Funding for this work was provided by NIH grants LM010098, DK112217, ES013508, the CHOP's Spatial and Functional Genomics Initiative, and the Daniel B. Burke Endowed Chair for Diabetes Research.

8. Authors' Contributions

EM, JHM, and SFAG conceived the study, identified appropriate data sets, and obtained the necessary access. EM designed the workflow, performed the analyses, and prepared the initial manuscript draft. AC performed QC and imputation on the WTCCC GWAS data set. MAH helped with PLATO settings and weighted encoding. All authors contributed to and approved the final version of the manuscript. SFAG and JHM supervised this work.

9. Acknowledgments

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award

076113, 085475 and 090355. The Consortium and/or Individual Investigators bear no responsibility for the further analysis or interpretation of these data, over and above that published by the Consortium.

The authors thank P.Schmitt, S. Dudek and C. Calafut for their assistance with the Biofilter and PLATO software installation and administration.

References

1. W.G. Hill, M.E. Goddard, P.M. Visscher, *PLoS Genet*. **4(2)**, e1000008 (2008).
2. E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, J.H. Nadeau, *Nat. Rev. Genet.* **11**, 446 (2010).
3. O. Zuk, E. Hechter, S.R. Sunyaev, E.S. Lander, *PNAS USA* **109**, 1193 (2012).
4. T.F.C. Mackay, *Nat. Rev. Genet*. **15(1)**, 22 (2014).
5. J.H. Moore, S.M. Williams, *Bioessays* **27(6**), 637 (2005).
6. J.H. Moore, S.M. Williams, *The American Journal of Human Genetics* **85**, 309 (2009).
7. P.C. Phillips PC, *Nat Rev Genet.* **9(11)**, 855 (2008).
8. M. Robnik-Šikonja, I. Kononenko, *Mach. Learn*. **53**, 23 (2003).
9. J.H. Moore, B.C. White, in E. Marchiori, J.H. Moore, J.C. Rajapakse (eds), *Evolutionary computation, machine learning and data mining in bioinformatics*, Springer, Berlin, pp.166 (2007).
10. J.H. Moore, *Methods Mol Biol.* **1253**, 315 (2015).
11. M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, *Am. J. Hum. Genet.* **69**, 138 (2001).
12. L. Qi, R.M. van Dam, F.W. Asselbergs, F.B. Hu, *Diabetic Medicine* **24**, 1187 (2007).
13. S.S. Verma, J.N. Cooke Bailey, A. Lucas, Y. Bradford, J.G. Linnemann, M.A. Hauser, L.R. Pasquale, P.L. Peissig, M.H. Brilliant, C.A. McCarty, J.L. Haines, J.L. Wiggs, T.R. Vrabec, G. Tromp, M.D. Ritchie, eMERGE Network, NEIGHBOR Consortium, *PLOS Genetics* **12(9)**, e1006186 (2016).
14. R.J. Urbanowicz, A.L.S. Granizo-Mackenzie, J. Kiralis, J.H. Moore, *BioData Min.* **7**, 8 (2014).
15. I. Mitra, A. Lavillareuix, E. Yeh, M. Traglia, K. Tsang, C.E. Bearden, K.A. Rauen, L.A. Weiss, *PLOS Genetics* **13(1)**, e1006516 (2017).
16. R. Cowper-Sal Iari, M.D. Cole, M.R. Karagas, M. Lupien, J.H. Moore, *Wiley Interdiscip Rev Syst Biol Med.* **3(5)**, 513 (2011).
17. E.A. Boyle, Y.I. Li, J.K. Pritchard, *Cell* **169(7)**, 1177 (2017).
18. T. Gao, B. He, S. Liu, H. Zhu, K. Tan, J. Qian, *Bioinformatics* **32(23)**, 3543 (2016).
19. W. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, *Genome Res.* **12(6)**, 996 (2002).
20. T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L.

Robertson, N. Serova, S. Davis, A. Soboleva, *Nucleic Acids Res.* **41(Database issue)**, D991 (2013).
21. B. Qin, M. Zhou, Y. Ge, L. Taing, T. Liu, Q. Wang, S. Wang, J. Chen, L. Shen, X. Duan, S. Hu, W. Li, H. Long, Y. Zhang, X.S. Liu, *Bioinformatics* **28(10)**, 1411 (2012).
22. ENCODE Project Consortium, *Nature* **489(7414)**, 57 (2012).
23. Roadmap Epigenomics Consortium, et al., *Nature* **518(7539)**, 317 (2015).
24. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F.O. Bagger, M. Jorgensen, P.R. Andersen, N. Bertin, O. Rackham, A.M. Burroughs, J.K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C.J. Mungall, T.F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C.O. Daub, P. Heutink, D.A. Hume, T.H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller; FANTOM Consortium, A.R. Forrest, P. Carninci, M. Rehli, A. Sandelin, *Nature* **507(7493)**, 455 (2014).
25. G.M. Clarke, C.A. Anderson, F.H. Pettersson, L.R. Cardon, A.P. Morris, K.T. Zondervan, *Nat Protoc.* **6(2)**, 121 (2011)
26. M.A. Hall, J. Wallace, A. Lucas, D. Kim, A.O. Basile, S.S. Verma, C.A. McCarty, M.H. Brilliant, P.L. Peissig, T.E. Kitchner, A. Verma, S. Pendergrass, S. Dudek, J.H. Moore, M.D. Ritchie, *Nature Communications* **in press**, (2017).
27. Welcome Trust Case Control Consortium, *Nature* **447(7145)**, 661 (2007)
28. K.A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z.Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, M. Feolo, *Nucleic Acids Res.* **42(Database issue)**, D975 (2014).
29. C.A. Anderson, F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris, K.T. Zondervan, *Nat. Protoc.* **5(9)**, 1564 (2010).
30. O. Delaneau, J. Marchini, J.F. Zagury, *Nat. Methods* **9(2)**, 179 (2012).
31. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, *Nature Genetics* **39**, 906 (2007).
32. L.D. Ward, M. Kellis, *Nucleic Acids Res.* **40(D1)**, D930 (2011).
33. W.S. Bush, S.M. Dudek, M.D. Ritchie, *Pac. Symp. Biocomput*, 368 (2009).
34. S.A. Pendergrass, A. Frase, J. Wallace, D. Wolfe, N. Katiyar, C. Moore, M.D. Ritchie, *BioData Min.* **6(1)**, 25 (2013).
35. C.S. Greene, N.M. Penrod, S.M. Williams, J.H. Moore JH, *PLoS One*. 2009 **4(6)**, e5639 (2009).
36. F. Reiman, G. Tolhurst, F.M. Gribble, *Cell Metabolism* **15(4)**, 421 (2012).
37. C. Sternini, L. Anselmi, E. Rozengurt, *Curr Opin Endocrinol Diabetes Obes.* **15(1)**, 73 (2008).
38. I. Kaji, S. Karaki, A. Kuwahara, *Curr. Pharm. Des.* **20(16)**, 2766 (2014).
39. F. Reimann, F. Gribble, *Diabetologia* **59**, 229 (2016).
40. Q. Xia, A. Chesi, E. Manduchi, B.T. Johnston, S. Lu, M.E. Leonard, U.W. Parlin, E.F. Rappaport, P. Huang, A.D. Wells, G.A. Blobel, M.E. Johnson, S.F. Grant, *Diabetologia* **59(11)**, 2360 (2016).

Advances in Text Mining and Visualization for Precision Medicine

Graciela Gonzalez-Hernandez†, Abeed Sarker, Karen O'Connor† and Casey Greene
Perelman School of Medicine, University of Pennsylvania
Philadelphia, Pennsylvania 19104, USA
Email: gragon@pennmedicine.upenn.edu

Hongfang Liu
Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine
Rochester, Minnesota 55902, USA
Email: liu.hongfang@mayo.edu

According to the National Institutes of Health (NIH), precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." Although the text mining community has explored this realm for some years, the official endorsement and funding launched in 2015 with the Precision Medicine Initiative are beginning to bear fruit. This session sought to elicit participation of researchers with strong background in text mining and/or visualization who are actively collaborating with bench scientists and clinicians for the deployment of integrative approaches in precision medicine that could impact scientific discovery and advance the vision of precision medicine as a universal, accessible approach at the point of care.

Keywords: Text mining; natural language processing; precision medicine; personalized medicine; visualization; biomedicine.

1. Introduction

According to the National Institutes of Health (NIH), precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." Announced in 2015, the Precision Medicine Initiative (PMI) seeks to promote research at the intersection of lifestyle, environment, and genetics to produce new knowledge for more effective ways to prolong health and treat disease [1].

Information and knowledge that could be instrumental to advances in precision medicine are buried in a vast, ever-increasing, and diverse range of data sources in structured and unstructured format: patient medical records (EMRs), standardized clinical data (such as what is required by Medicare), administrative data –from hospitals, insurance companies, and pharmacies-, patient surveys and self-reported comments from individual patients[2], the published literature, clinical trials, and research data deposited in public collections such GenBank[3] or the Gene Expression

† Work partially supported by the National Institute of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under grant number R01AI117011. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Omnibus (GEO) database[4], and many curated databases of interactions and pathways, to name just a few. Some recent text mining approaches related to precision medicine include automatically extracting and normalizing variant mentions in biomedical literature to reference variants in a curated database, thus allowing for analysis and novel discoveries.[5] Other relevant effort used associative text mining analysis of the free narratives of EMR to develop a system to identify previously unrecognized disease-associated factors.[6] Recent visualization advances have focused, for example, on genomic cancer data to help improve clinical decisions for precision oncology. [7–9] This session highlights original research and invited presentations on novel text mining, natural language processing (NLP), and visual analytics approaches at the intersection of lifestyle, environment, and genetics that enable further understanding of disease processes and effective treatment for individuals and cohorts that share specific characteristics.

2. Session Summary

The session includes two keynote talks by leaders in the field in biomedical data visualization and text mining, Jason Moore and Sophia Ananiadou. There are four full-length papers competitively selected for inclusion amongst the varied high-quality submissions exploring problems associated with the annotation of gene data sets, visualization of electronic health records and gene interaction to facilitate precision medicine, and concept normalization in clinical text. We selected contributions that are applicable to big genomic and text based data from multiple sources.

2.1. *Keynote: Visualization for Precision Medicine*

The first invited talk focusing on visualization is given by Jason Moore, Ph.D, Director of the Institute of Biomedical Informatics (IBI) and Senior Associate Dean for Informatics at the University of Pennsylvania's Perelman School of Medicine. Dr. Moore's work spans from artificial intelligence, data science, visualization and complex adaptive systems to systems biology, precision medicine and human genetics. Dr Moore's work relevant to this session is ample and varied. We outline here three of his most relevant papers as a quick reference:

- ViSEN, a methodology and software for visualization of statistical epistasis networks [10]. Epistasis, defined as the non-linear interaction effect among multiple genetic factors, has been recognized as a key component in understanding the underlying genetic basis of complex human diseases and phenotypic traits. ViSEN allows the analysis and visualization of two and three-way epistatic interactions. This visualized information could be very helpful to infer the underlying genetic architecture of complex diseases and to generate plausible hypotheses for further biological validations. ViSEN is freely available at https://sourceforge.net/projects/visen/.

- In PSB 2011, Dr Moore introduced a 3D visualization methodology and freely-available software package for facilitating the exploration and analysis of high-dimensional human microbiome data [11]. Powered by commercial video game development engines, the approach

provides an interactive medium in the form of a 3D heat map for exploration of microbial species and their relative abundance in different patients.

- Pioneering visualization of biological interpretation of gene expression microarray results, Dr Moore presented EVA (Exploratory Visual Analysis) [12] as a flexible combination of statistics and biological annotation to provide a visual interface for the interpretation of microarray analyses of gene expression in the most commonly occurring class of brain tumors, glioma.

Dr Moore's keynote is focused on big data processing and visualization techniques for precision medicine, and provides an insight about this rapidly emerging field. A world awash in big data presents significant computational challenges for identifying meaningful and actionable patterns. Visualization methods and technology are advancing at a rapid pace and have the potential to enable a deeper understanding of big data and derivative research results. This will require an active effort to adopt new visualization methods and to integrate them with computational analysis methods such as machine learning and natural language processing.

2.2. *Keynote: Text Mining for Precision Medicine*

The second keynote, focusing on text mining, is given by Sophia Ananiadou, PhD, director of the National Centre for Text Mining (NaCTeM) and Professor in the School of Computer Science at the University of Manchester. She has led the development of the numerous text mining tools and services currently used in NaCTeM with the aim to provide scalable text mining services: information extraction, intelligent searching, association mining, etc. She has received the IBM UIMA innovation award 3 consecutive times and is also a Daiwa award winner. Dr Ananiadou's publications relevant to this session span back at least a decade. We highlight three as a quick reference:

- In a recent publication[13], Dr Ananiadou presents a novel method that improves identification of textual uncertainty for extracted events and explores how it can be used as an additional measure of confidence for biomedical models. They use a hybrid approach that combines rule induction and machine learning with subjective logic theory to combine multiple uncertainty values extracted from different sources for the same interaction. The approach makes considerable improvements over previously published work. They evaluate their proposed system on pathways related to two different leukemia and melanoma cancer research.

- With the continuously rising need to understand the etiology of diseases as well as the demand for their informed diagnosis and personalized treatment, the curation of disease-relevant information from medical and clinical documents has become an indispensable scientific activity. Dr Anaiadou offers Argo (http://argo.nactem.ac.uk), a generic text mining workbench that can help in semi-automatic annotation of literature, including

annotation to standard terminologies, such as the UMLS. Argo's flexibility is put to the test with the semi-automatic curation of chronic obstructive pulmonary disease (COPD) phenotypes in this publication[14].

- To create, verify and maintain pathway models, curators must discover and assess knowledge distributed over vast biological literature. Dr Ananiadou explores methods for associating pathway model reactions with relevant publications[15]. The approach extracts the reactions directly from the models and then turns them into queries for three text mining-based MEDLINE literature search systems. These queries are executed, and the resulting documents are combined and ranked according to their relevance to the reactions of interest. An online demonstration of PathText 2 and the annotated corpus are available for research purposes at http://www.nactem.ac.uk/pathtext2/.

Dr Ananiadou's keynote focuses on text mining techniques to assist in the annotation and discovery of biological pathways. Pathway models are valuable resources that help us to understand the various mechanisms underpinning complex biological processes. Their curation is typically carried out through manual inspection of the scientific literature, a knowledge-intensive and laborious task. Text mining methods are used to automate model reconstruction by increasing the speed and reliability of discovery and extracting evidence from the literature. Complex information from the literature is automatically extracted and then mapped to reactions in existing pathway models. Information from the literature (events) can act as corroborative evidence of the validity of these reactions in a model or help to extend it. In addition, by contextualizing the textual evidence (extracting uncertainty, negation), we can provide additional confidence measures for linking and ranking information from the literature for model curation and ultimately better experimental design.

2.3. *Full-length Papers*

In *VisAGE: Integrating External Knowledge into Electronic Medical Record Visualization,* **Huang et al.** present a method that visualizes electronic medical records (EMRs) in a low dimensional space. Their work addresses a common issue with EMRs—that they are often fragmented and so visualization techniques often place unrelated patients close together in the visualized space. By integrating knowledge from external data sources, the system attempts to enrich EMR databases to solve this issue. This approach could aid clinicians in diagnosing and treating patients with conditions that are often misdiagnosed because they either have a collection of non-specific symptoms or are overshadowed by more prevalent conditions. The evaluations presented by the authors suggest that the method produces effective clustering of patients suffering from Parkinson's disease.

In *GeneDive: A Gene Interaction Search and Visualization Tool to Facilitate Precision Medicine,* **Previde et al.** address the problem of information overload that is faced by users of automatically mined, text-based gene interaction data by proposing a web-based tool that performs information retrieval, filtering and visualization tool. The tool, GeneDive, attempts to bring some of the best of the breed, adopting functionalities of text mining tools in the biomedical domain into a single platform. Inspired by the work of Literome [16]. GeneDive leverages Cytoscape [17], a software

package popularly used for visualization of biomolecular interactions, and DeepDive [18], a text mining tool for extracting gene interactions from literature, to provide a web-based retrieval, filtering and visualization tool for large volumes of interaction data. GeneDive offers various features and modalities that guide users through the search process to efficiently reach the information of their interest. The tool is time-efficient and it can process millions of interactions within seconds. The authors also discuss that in the future, the tool can be seamlessly extended to other interaction types such as gene-drug and gene-disease. The tool will also be made publicly available at: http://www.genedive.net.

In *Annotating Gene Sets by Mining Large Literature Collections with Protein Networks*, **Wang et al.** propose a natural language processing system that infers common functions for a gene set via the automated mining of scientific literature for relevant phrases. The system creates a heterogeneous network that connects genes with lexical concepts from the literature and combines these connections with protein interactions. The method works by performing a random walk over a heterogeneous network of phrases and genes. The authors argue that this approach presents two major advantages over previous text mining methods: (i) it integrates semantic information derived from the literature with biological information derived from experimental and interactome data, and (ii) the visualization technique reduces redundant information and visual complexity by utilizing a novel mechanism to organize functional annotations using a data structure called 'Hierarchical Concept Ontology'. The authors evaluate their method's ability to recover GO term names from the literature, applying the method to CLiXO gene sets [19], and identify a number of cancer-related terms. Evaluations of the method show substantial improvement in predicting manually curated annotations compared to a baseline text mining approach. The returned phrases remain relatively broad; however, the GO evaluation results are promising and the method takes an interesting step in the efforts to explain a gene set from literature.

In *Improving Precision in Concept Normalization*, **Boguslav et al.** propose a strategy for improving precision in medical text concept normalization by utilizing an existing high-performance biomedical concept recognition pipeline and a manually annotated corpus. The authors argue that precision is more important for health-related tasks, such as patient-centered decision support, since decisions based on false positives can be detrimental to patients' health. Although one counter-argument could be that computational system outputs are not directly used to make decisions but are vetted by human experts, and thus the role of such systems is to decrease the burden on the human agent. Thus, recall might indeed be important, but it is definitely a worthy endeavor to work toward precision gains if the loss in recall is small or can be addressed in the future, and hence the work by **Boguslav et al.** is a welcome direction. The normalization method primarily relies on a set of pre- and post-processing techniques that enable the use of a pre-existing corpus to perform the actual normalization task. The approach shows statistically significant improvements in precision over an existing baseline system for eight datasets, at the expense of recall.

3. Discussion

Text mining and visualization methods for biomedical data such as those presented in this session enable unprecedented use of data from diverse sources that can inform clinical decisions, and have come to be accepted as a necessary tool in advancing precision medicine. Visualizing such voluminous and heterogeneous data is a significant challenge, and tackling it in a way that can enable clinicians and researchers to advance precision medicine requires not only computational and logic acumen, but also creative visualization and attention to cognitive processes.

Visualization approaches and text mining techniques for information retrieval and natural language processing that are tailored to the specific needs of this domain and can handle big data play a vital role in harnessing the power of these sources to advance precision medicine research and delivery. The session aims to provide a platform for researchers to share their latest investigations in text mining and visualization and advance the vision of precision medicine as a universal, accessible approach at the point of care.

References

1. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015;372(9):793-795. doi:10.1056/NEJMp1500523.
2. Understanding Data Sources | Agency for Healthcare Research & Quality. https://www.ahrq.gov/professionals/quality-patient-safety/talkingquality/create/understand.html. Accessed October 6, 2017.
3. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. 2013;41(Database issue):D36-42. doi:10.1093/nar/gks1195.
4. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207.
5. Wei C-H, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*. September 2017. doi:10.1093/bioinformatics/btx541.
6. Lin FP-Y, Pokorny A, Teng C, Epstein RJ. TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Sci Rep*. 2017;7(1):6918. doi:10.1038/s41598-017-07111-0.
7. Gao J, Lindsay J, Watt S, et al. Abstract 5277: The cBioPortal for cancer genomics and its application in precision oncology. *Cancer Res*. 2016;76(14 Supplement):5277-5277. doi:10.1158/1538-7445.AM2016-5277.
8. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signal*. 2013;6(269):pl1. doi:10.1126/SCISIGNAL.2004088.
9. Klonowska K, Czubak K, Wojciechowska M, et al. Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget*. 2016;7(1):176-192. doi:10.18632/oncotarget.6128.
10. Hu T, Chen Y, Kiralis JW, Moore JH. ViSEN: methodology and software for visualization of statistical epistasis networks. *Genet Epidemiol*. 2013;37(3):283-285. doi:10.1002/gepi.21718.
11. Moore JH, Lari RCS, Hill D, Hibberd PL, Madan JC. Human microbiome visualization using 3D technology. *Pac Symp Biocomput*. 2011:154-164. http://www.ncbi.nlm.nih.gov/pubmed/21121043. Accessed October 6, 2017.
12. Reif DM, Israel MA, Moore JH. Exploratory Visual Analysis of statistical results from microarray experiments comparing high and low grade glioma. *Cancer Inform*. 2007;5:19-24. http://www.ncbi.nlm.nih.gov/pubmed/19390666. Accessed October 6, 2017.
13. Zerva C, Batista-Navarro R, Day P, Ananiadou S. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*. July 2017. doi:10.1093/bioinformatics/btx466.
14. Batista-Navarro R, Carter J, Ananiadou S. Argo: enabling the development of bespoke workflows and

services for disease annotation. *Database (Oxford)*. 2016;2016. doi:10.1093/database/baw066.
15. Miwa M, Ohta T, Rak R, et al. A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. *Bioinformatics*. 2013;29(13):i44-52. doi:10.1093/bioinformatics/btt227.
16. Poon H, Quirk C, DeZiel C, Heckerman D. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*. 2014;30(19):2840-2842. doi:10.1093/bioinformatics/btu383.
17. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303.
18. Mallory EK, Zhang C, Ré C, Altman RB. Large-scale extraction of gene interactions from full-text literature using DeepDive. *Bioinformatics*. 2016;32(1):106-113. doi:10.1093/bioinformatics/btv476.
19. Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. *Bioinformatics*. 2014;30(12):i34-42. doi:10.1093/bioinformatics/btu282.

Improving precision in concept normalization

Mayla Boguslav[†], K. Bretonnel Cohen, William A. Baumgartner Jr., and Lawrence E. Hunter
Computational Bioscience Program, University of Colorado School of Medicine,
Aurora, CO 80045, USA
[†]*E-mail: Mayla.Boguslav@ucdenver.edu*
compbio.ucdenver.edu

Most natural language processing applications exhibit a trade-off between precision and recall. In some use cases for natural language processing, there are reasons to prefer to tilt that trade-off toward high precision. Relying on the Zipfian distribution of false positive results, we describe a strategy for increasing precision, using a variety of both pre-processing and post-processing methods. They draw on both knowledge-based and frequentist approaches to modeling language. Based on an existing high-performance biomedical concept recognition pipeline and a previously published manually annotated corpus, we apply this hybrid rationalist/empiricist strategy to concept normalization for eight different ontologies. Which approaches did and did not improve precision varied widely between the ontologies.

Keywords: ontologies; precision medicine; natural language processing; text mining; concept normalization

1. Introduction

Ambiguity is a fundamental feature of all known human languages.[1] That ambiguity is one of the fundamental challenges for natural language processing systems.[2] One form of ambiguity relevant to text mining for precision medicine is *polysemy*—the phenomenon of words having more than one meaning. For example, Johnson et al.[3] point out that the word *cone*, concept BTO:0000280 in the BRENDA Tissue Ontology, refers to an ovule- or pollen-bearing structure that participates in reproductive functions of pine trees. The word *cone* also appears in terms from the Gene Ontology, where it refers to a kind of cell found in the retina. Since it has more than one potential meaning, it is ambiguous—specifically, it is polysemous.

Concept normalization is the language processing task of mapping mentions of concepts in text to an independently defined terminology or ontology.[4,5] Polysemy-based ambiguity of terms in the terminology can cause errors in that mapping. These include both false negatives—a failure to recognize a mention of a concept—and false positives—incorrectly identifying a concept as being mentioned, when in fact it has not. In some applications in precision medicine, false positives are more harmful than false negatives.[6–11] For example, false positives in tumor profiling in cancer precision medicine can lead to somatic and germline mutation confusion.[12] Despite this general challenge, dictionary-based methods are effective for mapping text to ontologies. For example, in normalizing genes, chemicals, cell types, and tissue types.[13] This motivates the error-analysis-based strategy for reducing the false positive rate for concept normalization systems.

Most systems[14] (and most evaluations of them[4]) attempt to optimize a measure called F_1, which weights false positives (i.e. precision) and false negatives (i.e. recall) equally. However, there are applications of concept normalization in which equal weighting is not ideal. For example, emphasizing precision is useful in protein protein interaction tasks involving large corpora because false positives will degrade the accuracy and reliability of the knowledge inferred.[15] High precision is also important for user acceptance, as the presence of obvious errors in system output reduces user confidence, for example, as shown in a study of information retrieval from medical textbooks.[16] Further, large corpora are often redundant, the same information is present multiple times.[17] If the goal is to identify *types* of concepts in large corpora, then high precision, at the cost of some loss of recall, improves overall performance. There are situations in which recall can be more important.[18,19] For example, recall is important "for practical applications like semantic search, since such applications need to recognise as many events as possible [in biomedical event extraction]."[18] If the results of text mining will be further used in computation, it is advantageous to have high recall for further training steps,[18] and it may be possible to filter out false positives in later processing, making recall more important.[19]

Here though, we focus on precision because of the application: precision medicine. Concept recognition has found many applications in precision medicine, for example information extraction from cancer clinical trials[6,7] or cancer pathology reports,[10] prediction of cancer[8] or cancer stage[9] in electronic medical records (EMRs), and in patient-centered decision support.[11] In each of these applications, improved precision (even at the cost of some recall) has significant advantages. One concern with many false positives is user acceptance: a system with many errors is less likely to be implemented. Thus, reducing false positives increases the likelihood that these systems will be used for precision medicine.

As with many other phenomena in NLP, the distribution of false positives is often Zipfian: a relatively small number of errors occur frequently, while most occur rarely. Johnson, et al.[3] observed that precision can often be improved significantly by addressing the common errors: "The Zipf-like distribution of error counts across terms suggest that filtering a small number of terms would have a beneficial effect on the error rates due to... ambiguity-related errors."[3] Thus, fixing the top errors can have a beneficial effect on overall performance. This is the motivation here.

To assess the accuracy of a concept normalization system, a manually annotated gold standard corpus is generally required. Here, we use the Colorado Richly Annotated Full Text Corpus (CRAFT) of full text biomedical journal articles, annotated with concepts from eight different ontologies.[20] As a baseline concept normalization system, we used the best performing systems from Funk, et al.,[21] with the precision maximizing parameters for each ontology. For each ontology, we tested for a Zipfian distribution, identified the most common concept errors in PubMed Central Open Access, and tested a set of five different potential pre- and post-processing steps that could improve precision.

2. Materials and Methods

To assess whether biomedical concepts have a Zipf-like distribution, and to identify the most common false positives, we use the PubMed Central Open Access (PMCOA) corpus. PMCOA is under the Creative Commons or similar license that generally allows more liberal redistribution and reuse

than a traditional copyrighted work.[22] As of this writing there are 1,494,227 articles in this set.

To calculate precision and recall, we used The Colorado Richly Annotated Full Text Corpus (CRAFT): 67 public full-text biomedical articles from PMCOA that are manually annotated, with approximately 100,000 concept annotations to eight different biomedical ontologies, including Chemical Entities of Biological Interest (ChEBI), Cell Ontology (CL), Gene Ontology (GO) which includes biological processes (GO-BP), cellular components (GO-CC), and molecular function (GO-MF), NCBI Taxonomy (NCBI Taxon), Protein Ontology (PRO), and Sequence Ontology (SO).[20] Our concept normalization system is ConceptMapper, a high-performance customizable dictionary look-up tool implemented as a UIMA component.[23] Funk et al. determined that ConceptMapper is the best performing (highest F_1 measure) concept recognition software as compared to others.[21] They also determined the best parameter settings for it to obtain the highest F_1 measure, precision, or recall.[21] We use the high-precision parameter settings as baseline concept normalization. Also, the dictionaries for ConceptMapper are taken from ontology concept names and synonyms. See `https://github.com/UCDenver-ccp/ccp-nlp-pipelines` for the ConceptMapper pipeline.

To test the distributional characteristics of the PMCOA corpus, we ran the ConceptMapper system over it and calculated concept frequencies for each ontology. To identify the most frequent false positives, we manually reviewed the 20 most frequently occurring concepts for each of the eight ontologies. Based on this manual review, we applied five different pre- and post-processing strategies (and their combinations) in no particular order, in order to improve precision for each ontology. Each strategy's effectiveness was evaluated using CRAFT. The strategies included:

- Pre-processing ConceptMapper dictionaries to remove problematic concepts/synonyms

  Some false positives are due to errors in ConceptMapper dictionary entries, so delete concepts or synonyms if they grossly misidentify the definition of the concept. For example, CHEBI:90880, the tipiracil cation, had as a synonym "(1+)," which matched all instances of the numeral "1" in PMCOA. Although 91% of the papers in PMCOA had at least one match, none of those were references to the ChEBI term.
- Pre-processing to remove single word concepts that are in the general English dictionary

  One type of ambiguity stems from the fact that some concepts have acronyms that are general English words (single words), even though they are scientifically specialized words. For example, the PRO concept, PR:00003444, "carbonic anhydrase 2," with acronym "CAN," triggers annotation of PR:00003444 to all instances of the general English word "can". This repair filters out concepts and synonyms that are in a general English dictionary in hopes to only capture the scientifically specialized words in the ontologies.
- Post-processing to check for the presence of acronym apositives

  To ensure that acronyms are found correctly, keep concepts using capitalization: all upper case or first letter capitalized. For example, the NCBI Taxon concept NCBITaxon:3702, "Arabidopsis thaliana," with acronym "AT," not the general English word "at," is commonly recognized as a false positive. Further, some acronym apositives have the first letter uppercase such as the GO-MF concept GO:0043336, "site-specific telomere resolvase activity," with acronym "ResT." This filter keeps concepts with either full capitalization or the first letter capitalized.
- Post-processing to keep only the canonical form of the concept

Keep concepts that are in the canonical form in the ConceptMapper dictionaries. For example, the ChEBI concept CHEBI:27889, "lead(0)," in its canonical form, is kept, but its synonym "lead" is not.

- Post-processing based on the frequency of occurrence of the concept in the document

  Calculate how frequent a concept appears in each document and test whether keeping an annotation if the concept was more or less frequent than a threshold improved precision. For example, if found that deleting concept annotations that appeared under 40 times in a text improved precision, and the PRO concept PR:000009431, "kinetochore-associated protein 1," with acronym "ROD," appeared 28 times in a document, then the PRO concept annotation would be removed. To determine the threshold for each ontology, optimize precision over a range of thresholds from 0 to 50 on CRAFT. 50 was the upper threshold because manually, the max threshold found was 40. Note that each ontology may have its own threshold.

3. Results

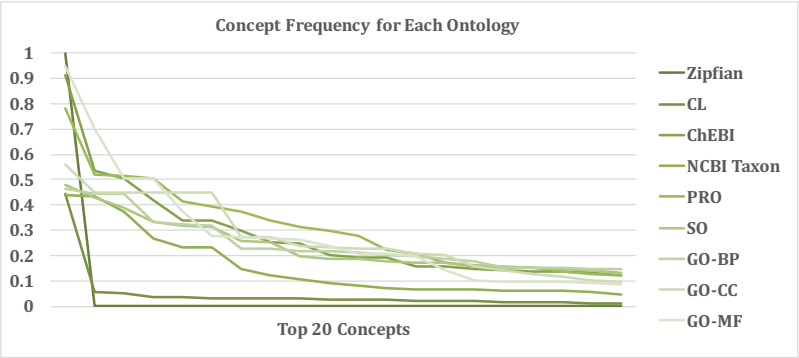


Fig. 1: The concept frequencies for all 8 ontologies for the top 20 concepts. The rates decline rapidly in a Zipfian distribution using the Kolmogorov-Smirnov test[24] based on no significant difference between each ontology distribution and the Zipf distribution (p-value > 0.2).

The distribution of the 20 most frequent concepts in the PMCOA literature obeys a Zipfian distribution (see figure 1) using the Kolmogorov-Smirnov test[24] (igraph package in R[25]) based on seeing no significant difference between each ontology distribution and a Zipfian distribution (see `github.com/mboguslav/Concept-Normalization`). Thus, fixing the most-frequent errors could be expected to increase precision. The 20 most frequently recognized concepts from each ontology in the PMCOA corpus were evaluated manually. Many of the most frequently occurring concepts were errors. The top two terms for each ontology are listed below with false positives in red and the ontology concept ID and frequency in parentheses. The top 20 are available at `https://github.com/mboguslav/Concept-Normalization`. This manual evaluation was used to determine the pre- and post-processing strategies that aim to get rid of false positives and improve precision.

- CL: cell (CL_0000000, 44%); T cell (CL_0000084, 5.6%)
- ChEBI: (1+) (CHEBI_90880, 91%); group (CHEBI_24433, 54%)

- NCBI Taxon: Homo Sapiens (NCBITaxon_9606, 44%); order (NCBITaxon_order, 43%)
- PRO: carbonic anhydrase 2 (PR_000034449, 78%); Golgi-associated PDZ and coiled-coil motif-containing protein (PR_000008147, 52%)
- SO: single (SO_0000984, 48%); region (SO_0000001, 43%)
- GO-BP: developmental process (GO_0032502, 47%); cell aging (GO_0007569, 44%)
- GO-CC: cell and encapsulating structures (GO_0005623, 56%); cell part (GO_0044464, 45%)
- GO-MF: 3-methyl-2-oxobutanoate dehydrogenase (ferredoxin) activity (GO_0043807, 95%); enoyl-CoA hydratase activity (GO_0004300, 70%)

The manually determined problematic concepts in the top two are in red above, such as the "(1+)" issue already mentioned. With this list of problematic terms, an error analysis (looking to the original text) was performed to determine the cause of the false positives. This demonstrated that many of the false positives were also terms in general English (e.g. can, lead, fig). We then assessed all the terms in each ontology that were in a general English dictionary using the Enchant spell-checker library[26] (see figure 3). With this information, we tried to post-process the general English words (as mentioned in methods), but the precision did not increase and this was better captured by the other post-processing methods.

The highest precision criteria were chosen, and statistical significance of the difference was calculated according to Yeh.[27] This criteria was implemented on PMCOA to get the new top 20 most frequent concepts, in hope that the new concept frequencies make more sense then the previous (see `https://github.com/mboguslav/Concept-Normalization` for the new top 20 concepts).

In general, only two ontologies needed pre-processing due to problematic concepts or synonyms: ChEBI and PRO. This increased precision significantly for ChEBI and had little impact on PRO (see section 3.1 with figure 2).

Within post-processing, focusing on precision only, the best criteria to improve precision was checking for the presence of appositives: keeping concept annotations if the annotation was all upper case (an acronym) or the first letter of the annotation was uppercase (acronym or proper noun), and discarding them otherwise (see section 3.3 with figure 4a). With these criteria, precision improved significantly for 6 out of the 8 ontologies. However, recall plummets at the same time, except for PRO: the change in recall is minimal. Thus different criteria is needed that improves precision and maintains recall (remains the same or slightly decreases).

Combining the post-processing that keeps only the canonical forms and the post-processing based on the frequency of occurrence, improves precision and only slightly lowers recall. Thus keep concept annotations if they are in the canonical form of the ontology concept or the number of times the annotation concept appears per document is over a specific threshold as explained above. With the optimal thresholds for the highest precision for each ontology, ConceptMapper ran over CRAFT (see thresholds in table 1). This significantly improved precision for 5 of the 8 ontologies and did nothing for the others (see section 3.4). Overall precision improves, while recall decreases slightly leading to a decrease in F_1 measure.

3.1. *Results from deleting concepts and synonyms*

Since deleting concepts or synonyms is likely to cause an increase in false negatives, this approach is likely to lead to performance improvements only in the case of obvious ontological errors, such as the case of "(1+)" as a synonym for the tipiracil cation. These errors were found in ChEBI and PRO. Changes led to a modest improvement for ChEBI, but not for PRO. This is likely because the erroneous terms do not appear in CRAFT (although they did in PMCOA). For example, the PRO concept PR:000015574, "small proline-rich protein 2A," has the synonym "2-1," which was deleted, but the synonym is not in CRAFT. In no case did this reduce performance (see figure 2).

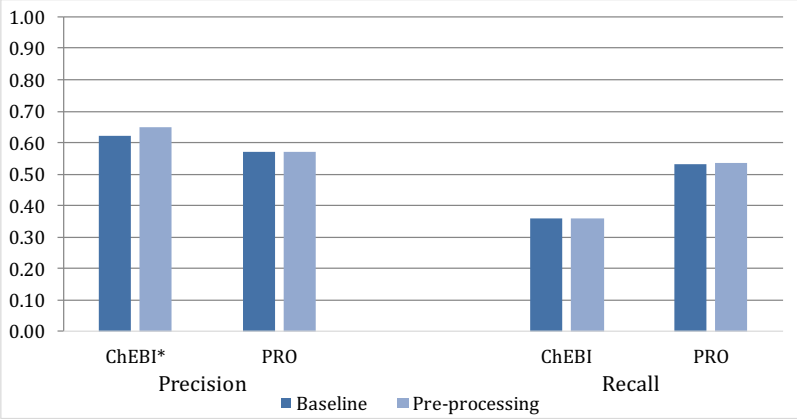


Fig. 2: Precision and recall after the pre-processing step as compared to the baseline. *Statistically significant difference in precision between pre-processed and baseline using χ^2 test.[27]

3.2. *Results from general English dictionary*

Some of the ambiguity issues involve concept acronyms that are words in the general English dictionary. Thus, we quantified the number of single word concepts that are in the general English dictionary for each ontology. As figure 3 shows, GO-BP has the most single word concepts that are general English words, followed by GO-CC. The caveat to this is that some concepts are general English words and correctly identify the ontology concept, such as sulfation in GO-BP. Due to this caveat, more concept annotations were deleted that were identifying the correct concept, rather than identifying the wrong concept, leading to a decrease in precision. Further, the issue with single word concepts as general English words is somewhat solved in the other post-processing criteria.

3.3. *Results from Case post-processing*

Requiring annotations to be acronyms based on case, improved precision significantly for six ontologies, but hurt two (NCBI Taxon and GO-CC). The first letter capitalized also finds proper nouns, which are also more likely to be concepts of interest. For all but NCBI Taxon, precision was above 0.74 with most between 0.8 and 0.98 (see figure 4a). Looking at F_1 measure, overall it drastically declined, except for PRO which increased from 0.55 to 0.58, suggesting that this may be a good fix for PRO at least (see figure 5).

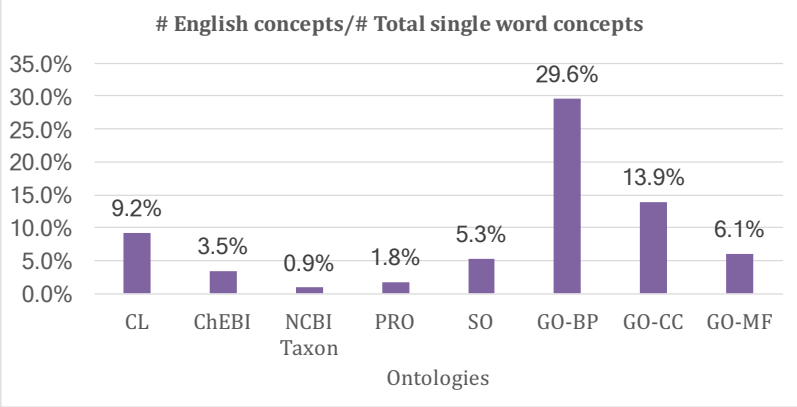
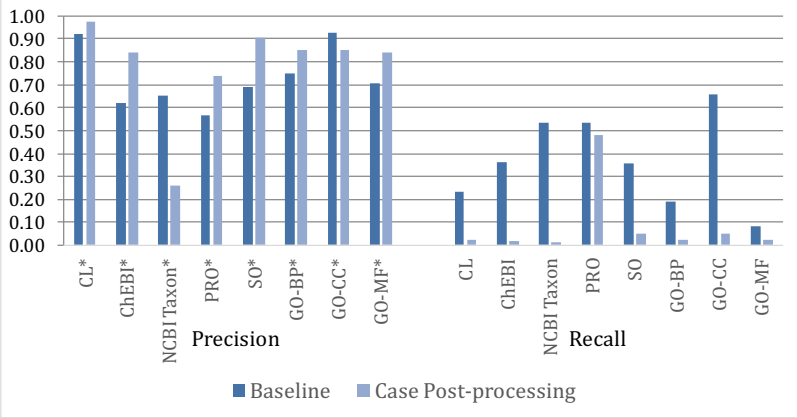
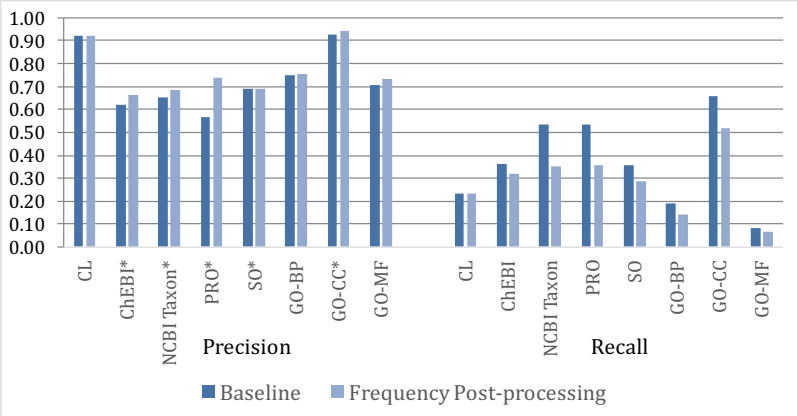


Fig. 3: The percentage of single word concepts for each ontology in the general English dictionary.



(a) Case post-processing



(b) Frequency post-processing

Fig. 4: Precision and recall for both case post-processing and frequency post-processing as compared to the baseline. *Statistically significant difference in precision between post-processing and baseline using χ^2 test.[27]

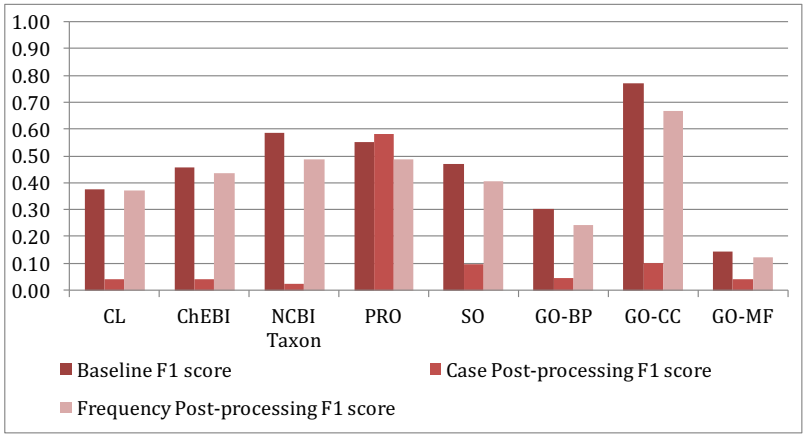


Fig. 5: F_1 scores for both post-processing steps as compared to baseline.

3.4. *Results from canonical form and frequency-based post-processing*

Requiring the canonical form of the concept is helpful: using the canonical forms improved precision because of its specificity to the concept. For example, in the "lead(0)" example above, the synonym "lead" causes false positive issues, finding annotations of the verb ("lead to") instead of the noun. Thus, if we only keep the annotations that have the "lead(0)" form, then we are more likely to annotate the concept correctly. At the same time, keeping only the canonical form drastically decreased recall because concept synonyms are widely used, including "lead". Therefore, the canonical form post-processing must be used in combination with another post-processing technique: frequency-based post-processing. Using the concept annotation frequency per document (see table 1 for thresholds for each ontology) and the specificity of the canonical forms, enabled us to improve precision significantly for five ontologies (ChEBI, NCBI Taxon, PRO, SO, and GO-CC), not hurt the other three, and while only slightly lowering recall (see figure 4b). Precision also improved for the two ontologies, NCBI Taxon and GO-CC, where precision decreased using case post-processing (see figure 4a). Overall, F_1 measure remained the same or decreased slightly for each ontology (see figure 5), an improvement on the case post-processing.

Table 1: The concept frequency threshold for each ontology based on maximizing precision with thresholds between 0 and 50.

Ontology	Frequency threshold	Ontology	Frequency threshold
CL	2	SO	42
ChEBI	22	GO-BP	43
NCBI Taxon	40	GO-CC	22
PRO	43	GO-MF	3

4. Repeatability and reproducibility

The ConceptMapper pipeline is available at:
`github.com/UCDenver-ccp/ccp-nlp-pipelines`. The intermediate data files for the analysis are available at `github.com/mboguslav/Concept-Normalization`.

Since ontologies are often updated, the results here are unlikely to be reproduced exactly the same in the future. In a more general sense, however, the effects of applying the pre- and post-processing rules to the concept annotations can change any time the contents of the ontologies change. For example, we reported the tipiracil cation error to the ChEBI maintainers. They have since replaced *(1+)* with *tipiracil(1+),* obviating the need for our handling of the false positives that resulted from the *(1+)* synonym. Although this could potentially change the measured performance of the current implementation, on balance it is difficult to see this state of affairs as anything other than good.

These are weaknesses of the current version of the implementation (and probably any future ones). Nonetheless, the elaboration of the hybrid rationalist/empiricist methodology of using distributional aspects of the errors to prioritize which ones to address, with knowledge-based approaches being applied to recover from them, remains a contribution. There is an oft-heard trope that purely statistical approaches to language processing "work" better than knowledge-driven ones. That claim is rarely, if ever, substantiated. The work presented here is consistent with what one actually observes in practice: hybrid systems are a reasonable approach to the challenges of the ambiguity of natural language.

5. Discussion and conclusions

Table 2: Summary of findings for each pre- or post-processing step. Y means the method increased precision, No means the method did not change precision, and Decrease means that the method decreased precision.

Ontology	Pre-processing	Case post-processing	Frequency post-processing
CL		**Y**	No
ChEBI	**Y**	**Y**	**Y**
NCBI Taxon		**Decrease**	**Y**
PRO	No	**Y**	**Y**
SO		**Y**	No
GO-BP		**Y**	No
GO-CC		**Decrease**	**Y**
GO-MF		**Y**	No

At least one method improved precision for each ontology tested (see table 2), since there is a "Y" for at least one method for each ontology. Further, the ontology with the worst baseline precision was PRO and precision significantly improved using both post-processing methods, and F_1 measure even improved using the case method (see figure 4). The ontology with the second lowest precision was ChEBI, and precision improved significantly for all methods (see "Y" for all methods for ChEBI

in table 2). Note that there is no single method that only increases or only decreases precision for all ontologies. Each method presented here though, helps at least one ontology, specifically pre-processing, the case post-processing, and frequency post-processing methods. So, we suggest trying these methods to improve precision either on a manual annotated corpus like CRAFT, or an unknown set of documents that one can review manually after, before singling out each ontology and finding its specific errors and fixes. Another option is to analyze the most frequent concepts as we did here to see if they are false positives and follow a Zipfian distribution.

Although these results are from the biomedical literature, there are reasons to believe this will apply to concept normalization in electronic patient records as well. For example, a general text mining NLP tool and ontologies have been used to extract information from electronic medical records.[28]

5.1. *Limitations*

Here we only evaluated our metrics (precision, recall, and F_1 measure) using CRAFT and do not know how a different corpus may affect these metrics.[29] Further, the limited size (and domain) of the CRAFT corpus means that the estimates of precision and recall will not perfectly reflect the changes in performance over the entire PMCOA corpus.[30] However, we believe that some of the changes that showed little effect in CRAFT are actually useful over the entire PMCOA corpus, suggested by the fact that the new top 20 concepts for each ontology have more realistic concept frequencies. Further, there are other possible post-processing rules to evaluate that were not included in this study. For example, for the PRO concept PR:000008147, "Golgi-associated PDZ and coiled-coil motif-containing protein," with acronym "FIG," which currently recognizes the same shorthand for figures in text, context could inform post-processing. For figures in the text, a number usually appears after it, whereas it would probably not for the protein. Future directions include trying other pre- and post-processing techniques that have the potential to improve precision.

5.2. *Conclusion*

The work here shows that it is possible to improve precision without losing much recall for existing systems, ConceptMapper, by exploring both pre- and post-processing methods of concepts and concept annotations, respectively. Further, this work provides some evidence that there is no single fix that will improve precision for concept recognition for all ontologies, based on exploring five different methods to do so. At the same time, these methods improve precision for at least one ontology, suggesting that trying these methods before examining each ontology closely could be worthwhile. For example, the case post-processing method improves both precision and F_1 measure for the protein ontology suggesting that this is a necessary step in concept normalization for it. Thus trying these different combinations of pre- and post-processing steps on other ontologies may prove fruitful for concept recognition as a whole.

6. Acknowledgements

Boguslav was supported by the Deans Fund at University of Colorado Anschutz Medical and now is supported by the training grant T15 LM009451. Cohen is supported by NIH grants LM008111,

LM009254, and NSF IIS-1207592 to Lawrence E. Hunter, and by generous funding from Labex DigiCosme (project ANR11LABEX0045 DIGICOSME) operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11 IDEX000302), as well as by a Jean dAlembert fellowship. Hunter is supported by the grants R01 LM008111 and R01 LM009254, which also supports Baumgartner. The work was aided by discussions with Michael Bada, Negacy Hailu, and Tiffany Callahan; all remaining faults are the authors.

7. Author contributions

MB: ran all experiments, analyzed the data, and wrote the final version of the paper. KBC: analyzed the data and wrote the first draft of the paper. WAB: performed precursor work and participated in running experiments. LEH: conceived of and directed the project. All authors participated in analyzing the data and approved the final version of the paper.

References

1. D. of Linguistics, *Language Files: Materials for an Introduction to Language and Linguistics.*, 12 edn. (The Ohio State University Press, 2016).
2. D. Jurafsky and J. H. Martin, *Speech and language processing* (Pearson London, 2014).
3. H. L. Johnson, K. B. Cohen and L. Hunter, A fault model for ontology mapping, alignment, and linking systems, in *Pacific symposium on biocomputing.*, 2007.
4. F. Rinaldi, T. R. Ellendorff, S. Madan, S. Clematide, A. Van der Lek, T. Mevissen and J. Fluck, *Database* **2016** (2016).
5. K. B. Cohen, G. K. Acquaah-Mensah, A. E. Dolbey and L. Hunter, Contrast and variability in gene names, in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, 2002.
6. J. Zeng, Y. Wu, A. Bailey, A. Johnson, V. Holla, E. V. Bernstam, H. Xu and F. Meric-Bernstam, *AMIA Summits on Translational Science Proceedings* **2014**, p. 126 (2014).
7. J. Xu, H.-J. Lee, J. Zeng, Y. Wu, Y. Zhang, L.-C. Huang, A. Johnson, V. Holla, A. M. Bailey, T. Cohen *et al.*, *Journal of the American Medical Informatics Association* **23**, 750 (2016).
8. M. Hoogendoorn, P. Szolovits, L. M. Moons and M. E. Numans, *Artificial intelligence in medicine* **69**, 53 (2016).
9. J. L. Warner, M. A. Levy, M. N. Neuss, J. L. Warner, M. A. Levy and M. N. Neuss, *Journal of oncology practice* **12**, 157 (2015).
10. A. E. Wieneke, E. J. Bowles, D. Cronkite, K. J. Wernli, H. Gao, D. Carrell and D. S. Buist, *Journal of pathology informatics* **6** (2015).
11. M. Becker and B. Böckmann, *Studies in health technology and informatics* **235**, p. 271 (2017).
12. A. Garofalo, L. Sholl, B. Reardon, A. Taylor-Weiner, A. Amin-Mansour, D. Miao, D. Liu, N. Oliver, L. MacConaill, M. Ducar *et al.*, *Genome medicine* **8**, p. 79 (2016).
13. H. L. Johnson, K. B. Cohen, W. A. Baumgartner Jr, Z. Lu, M. Bada, T. Kester, H. Kim and L. Hunter, Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies, in *Pacific Symposium on Biocomputing*, 2006.
14. K. B. Cohen and D. Demner-Fushman, *Biomedical natural language processing* (John Benjamins Publishing Company, 2014).
15. J. Lee, S. Kim, S. Lee, K. Lee and J. Kang, *BMC medical informatics and decision making* **13**, p. S7 (2013).
16. D. C. Berrios, R. J. Cucina and L. M. Fagan, *Journal of the American Medical Informatics Association* **9**, 637 (2002).

17. K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner Jr, E. White, H. Tipney and L. Hunter, *Computational intelligence* **27**, 681 (2011).
18. M. Miwa and S. Ananiadou, *BMC bioinformatics* **16**, p. S7 (2015).
19. W. A. Baumgartner, K. B. Cohen and L. Hunter, *Journal of biomedical discovery and collaboration* **3**, p. 1 (2008).
20. K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer and L. E. Hunter, *The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain*, in *Handbook of Linguistic Annotation*, (Springer, 2017), pp. 1379–1394.
21. C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter and K. Verspoor, *BMC bioinformatics* **15**, p. 59 (2014).
22. J. A. Fain, Nih public access policy how does information get uploaded to pubmed central and by whom? (2015).
23. M. A. Tanenblatt, A. Coden and I. L. Sominsky, The conceptmapper approach to named entity recognition, in *LREC*, 2010.
24. R. Wilcox, *Encyclopedia of biostatistics* (2005).
25. T. Nepusz and G. Csardi, R igraph manual pages (2003), `http://igraph.org/r/doc/fit_power_law.html`.
26. D. Lachowicz, Enchant (2008), `https://abiword.github.io/enchant/`.
27. A. Yeh, More accurate tests for the statistical significance of result differences, in *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 2000.
28. R. Batool, A. M. Khattak, T. S. Kim and S. Lee, Automatic extraction and mapping of discharge summarys concepts into SNOMED CT, in *Conf Proc IEEE Eng Med Biol Soc*, 2013.
29. S. Mehrabi, A. Krishnan, A. M. Roch, H. Schmidt, D. Li, J. Kesterson, C. Beesley, P. Dexter, M. Schmidt, M. Palakal *et al.*, *Studies in health technology and informatics* **216**, 604 (2015).
30. J. G. Caporaso, N. Deshpande, J. L. Fink, P. E. Bourne, K. B. Cohen and L. Hunter, Intrinsic evaluation of text mining tools may not predict performance on realistic tasks, in *Pacific symposium on biocomputing.*, 2008.

VisAGE: Integrating external knowledge into electronic medical record visualization

Edward W Huang, Sheng Wang, and ChengXiang Zhai†

Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, IL, USA
†*Email: czhai@illinois.edu*

In this paper, we present VisAGE, a method that visualizes electronic medical records (EMRs) in a low-dimensional space. Effective visualization of new patients allows doctors to view similar, previously treated patients and to identify the new patients' disease subtypes, reducing the chance of misdiagnosis. However, EMRs are typically incomplete or fragmented, resulting in patients who are missing many available features being placed near unrelated patients in the visualized space. VisAGE integrates several external data sources to enrich EMR databases to solve this issue. We evaluated VisAGE on a dataset of Parkinson's disease patients. We qualitatively and quantitatively show that VisAGE can more effectively cluster patients, which allows doctors to better discover patient subtypes and thus improve patient care.

Keywords: Electronic medical records; Data integration; Knowledge graphs; Visualization.

1. Introduction

In modern healthcare settings, doctors record the details of patient visits in electronic medical records (EMRs), which are then collected in databases. At first, EMR systems were poorly implemented; initial studies reported that they reduced physician productivity and lacked data sharing capabilities.[1] However, recent advances have improved recordkeeping and decision support. For example, EMR systems have enhanced productivity in physician workloads[2] and have increased the delivery of health behavior counseling.[3]

Despite these advances, EMR systems can still benefit from human interpretation, which allows for exploratory analysis and more control over decision-making.[4] However, human interaction with EMR systems has been hindered. 37% of participants in a previous study reported that interacting with their EMR databases was too time consuming.[5] Another study showed that when using EMRs, nurses face challenges that can threaten quality and safety of care.[6] Both shortcomings can be addressed with information visualization, which can aid doctors in processing and understanding complex, high-dimensional EMR data.

In particular, EMR visualization in a two-dimensional space is useful for observing disease subtypes in patient clusters. Coherent clusters may elucidate a patient's most significant characteristics by visualizing his or her proximity to successfully diagnosed patients.[7] For example, thoracic aortic dissections are commonly misdiagnosed as acute myocardial infarctions (MIs).[8] Misdiagnosis in these cases is extremely harmful, as patients with aortic dissections treated for MIs have mortality rates similar to that of untreated patients. Despite this risk, patients with aortic dissections have a misdiagnosis rate of 39%.[9] Fortunately, the most telltale signs

of aortic dissection (age, onset of pain, and syncope) are readily available in EMRs.[10,11] An effective visualization would utilize these features to place an undiagnosed aortic dissection patient near similar patients, reducing the chance of misdiagnosis.

Unfortunately, designing an effective visualization system is a complicated task, as EMRs are high-dimensional sources of data that consist of thousands of features. Because there are many potential medical tests that a patient can take, but only a handful are relevant to each patient's condition, EMRs also tend to be sparse. Additionally, EMRs are typically fragmented or incomplete due to human error. These reasons make it challenging to correctly group together similar patients, leading to poor or even misleading visualizations.

To address these challenges, we present **Vis**ualization **A**ssisted by Knowledge **G**raph **E**nrichment (VisAGE), a method that enriches patient records with a knowledge graph built from external databases. These databases include protein-protein interactions, genomic data, and drug-chemical associations. Performing network embedding on the knowledge graph allows us to infer associations among different types of data to accommodate inexact matchings of related features, which can alleviate data sparsity in EMRs. A major novelty of VisAGE is that it is the first to use all of these data sources in EMR visualization. In the rest of the paper, we describe our dataset, the details of VisAGE, and our evaluation process.

2. Data Description

While VisAGE is a general method that can be applied to any set of EMRs, we chose the Parkinson's Progression Markers Initiative (PPMI) dataset[12] for the evaluation process. The PPMI dataset contains a mix of Parkinson's disease (PD) patients and control patients suffering from other diseases. We chose this dataset for two reasons: (1) it contains many feature types, and (2) Parkinson's disease is a complicated disorder the causes of which have been attributed to complex combinations of genetic and environmental factors. We only considered the 1,579 patients with baseline visits. This dataset includes 6,013 biospecimen, genetic, drug, symptom, diagnosis, medical test, and demographic features. Feature types include binary, numerical, and categorical features. We binarized the categorical features. On average, each patient only has 261 of the 6,013 available features, which supports our previous assertion that EMRs are typically sparse.

3. Patient Profile Matrix

In general, each EMR can be formally represented as a high-dimensional feature vector. Features can be any attributes of interest in an EMR. We first generated a $p \times n$ patient profile matrix, M, from the data, where p is the number of patients and n is the number of features. Thus, each row corresponds to a patient record and each column corresponds to a feature. In our dataset, $p = 1,579$ and $n = 6,013$. Existing visualization methods use M directly as input (see **Related Work**). However, as previously stated, M is typically sparse and thus suboptimal for visualization. The main idea of VisAGE is to enrich M before visualization by leveraging associations inferred from a knowledge graph. The enriched profile matrix, M', then replaces M as the input to any visualization method. We later show that M' gives better visualizations in several applications on our dataset.

4. Patient Profile Matrix Enrichment

Our proposed method for enriching a profile matrix consists of three steps (Figure 1). The first constructs a knowledge graph with external data sources and EMRs. The second performs embedding on the constructed graph to learn a similarity matrix. Lastly, the third step multiplies M by the similarity matrix to obtain M'. We now describe each step in more detail.

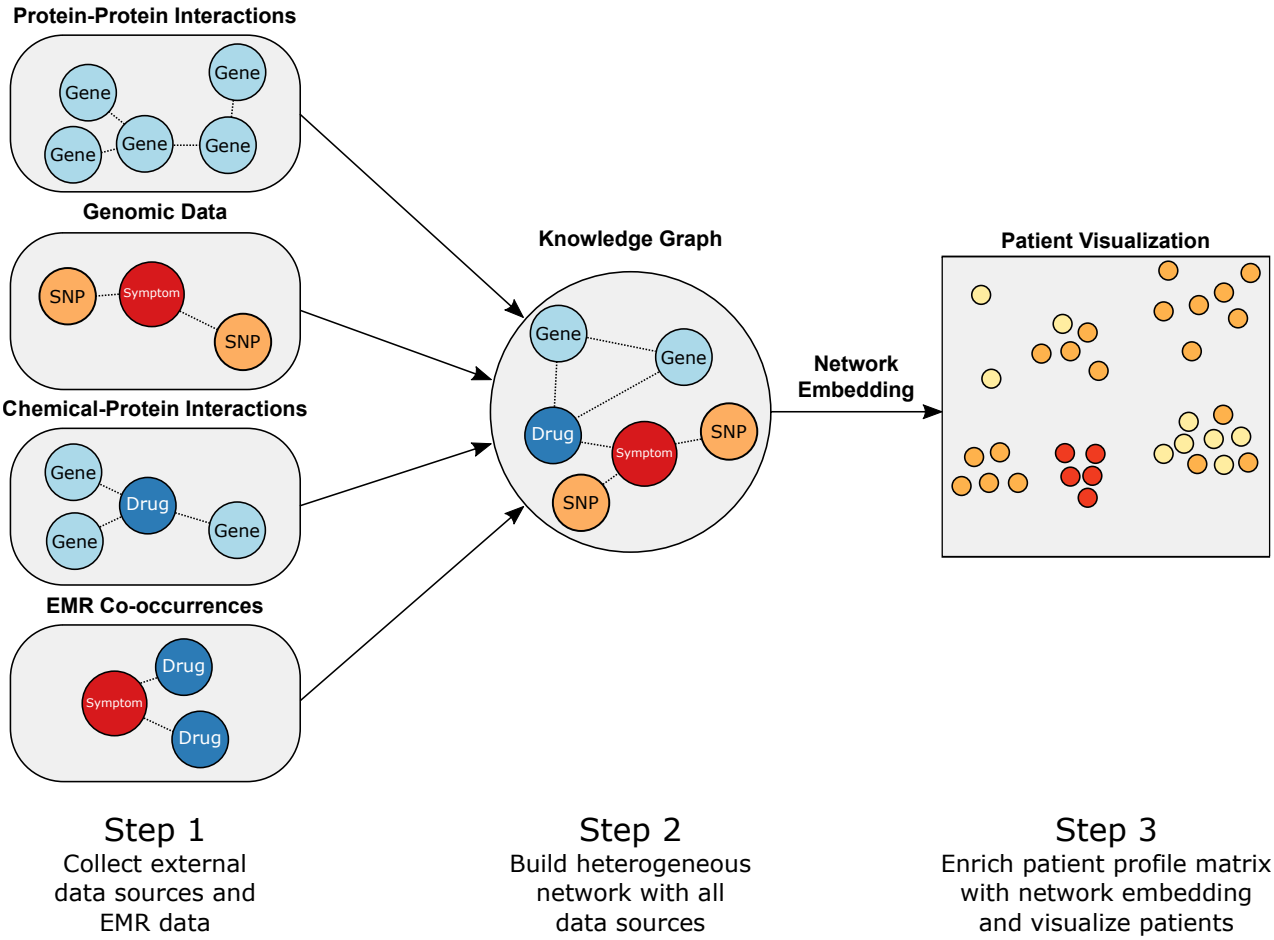


Fig. 1: The VisAGE pipeline. We first create a knowledge graph from multiple data sources. We then perform network embedding on this knowledge graph, enrich the patient profile matrix, and then visualize each patient in a two-dimensional space.

4.0.1. *Knowledge Graph Construction*

The knowledge graph is a heterogeneous network containing edges from four data sources.

(1) **Protein-protein interaction network.** We used the inBioMap database[13] of protein-protein interaction (PPI) edges. For a functional linkage between two proteins p_1 and p_2, we created a node for p_1, a node for p_2, and an undirected edge $\{p_1, p_2\}$ in the network. There were 17,327 proteins and 606,194 edges in this network.

(2) **Single-nucleotide polymorphism enrichment.** We integrated genomic data in the form of single-nucleotide polymorphisms (SNPs), which are single variations in the human genome. We identified SNPs that are highly enriched in PD patients in the dataset by using a one-sided Fisher's exact test.[14] Overall, we found 3,900 SNPs with p-values < 0.05. We then selected the nonsynonymous SNPs and determined if specific symptoms were enriched in SNPs with another one-sided Fisher's exact test. For each PD-enriched SNP g, we created a node for g and an edge $\{g, s\}$ if s was significantly enriched in g with a p-value < 0.01. There were 34,324 SNP-symptom edges.

(3) **Chemical-protein interaction network.** We used STITCH, a database of known and predicted interactions between chemicals and proteins.[15] STITCH includes computationally predicted associations in addition to those aggregated from other databases. For each drug d in the EMR data, if d's active ingredient interacts with a protein p in the STITCH database, then we created a node for d, a node for p, and an undirected edge $\{d, p\}$. There were 7,218 drug-protein edges in this network.

(4) **Electronic medical records.** We directly added co-occurrence edges from each medical record. For example, if a patient was diagnosed with symptom s and prescribed a drug d, then we created a node s, a node d, and an undirected edge $\{s, d\}$. We repeated this for all elements in each patient's medical record.

The resulting network contained 23,886 nodes and 17,108,116 edges. The knowledge graph's purpose is to utilize the "guilt by association" rule: although many related medical concepts may not directly co-occur in any medical records, they may indirectly share neighbors in the knowledge graph through the protein-protein, SNP-symptom, and drug-protein edges.

4.1. *Similarity Matrix Learning*

We used the recently proposed method ProSNet to infer the relationships among the entities in the knowledge graph.[16] ProSNet performs a dimensionality reduction algorithm on heterogeneous networks to optimize a low-dimensional vector representation for each node. The vectors of two nodes will be co-localized in the low-dimensional space if the nodes are near each other in the heterogeneous network. After generating these vectors, we enriched the patient profile matrix. The similarities between the node vectors capture the latent relationships among the external data sources and the information in the EMRs. VisAGE constructs an $n \times n$ similarity matrix, S, where n is the number of features and S_{ij} is the cosine similarity between feature i's low-dimensional vector and feature j's low-dimensional vector.

4.1.1. *Enriching the Profile Matrix*

After obtaining S, we generated the enriched patient profile, M', with the following operation:

$$M' = M \times S \tag{1}$$

This multiplication allows features that are related but not semantically identical to partially match through similarity scores. As stated before, each patient in the original patient profile matrix, M, only had an average of 261 nonzero features. On the other hand, each patient in M' had 1,732 nonzero features.

5. Evaluation

We wished to determine whether M' would lead to better visualization results than the original patient profile matrix, M. Thus, we compared the results between using M' versus using M in meaningful downstream visualization applications. Specifically, following previous studies,[17,18] we used t-distributed stochastic neighbor embedding (t-SNE),[19] an algorithm that can efficiently model high-dimensional objects as two-dimensional points, which makes it especially well-suited for visualizing our dataset. We generated our visualization by running t-SNE with default settings on the patient profile matrix M for the baseline and M' for VisAGE. This created a new $p \times 2$ matrix, so that each patient was finally reduced to two dimensions. We plotted this matrix as a set of points. We now discuss our results when visualizing M' and M in various applications.

5.1. *Two-Dimensional Visualization with UPDRS*

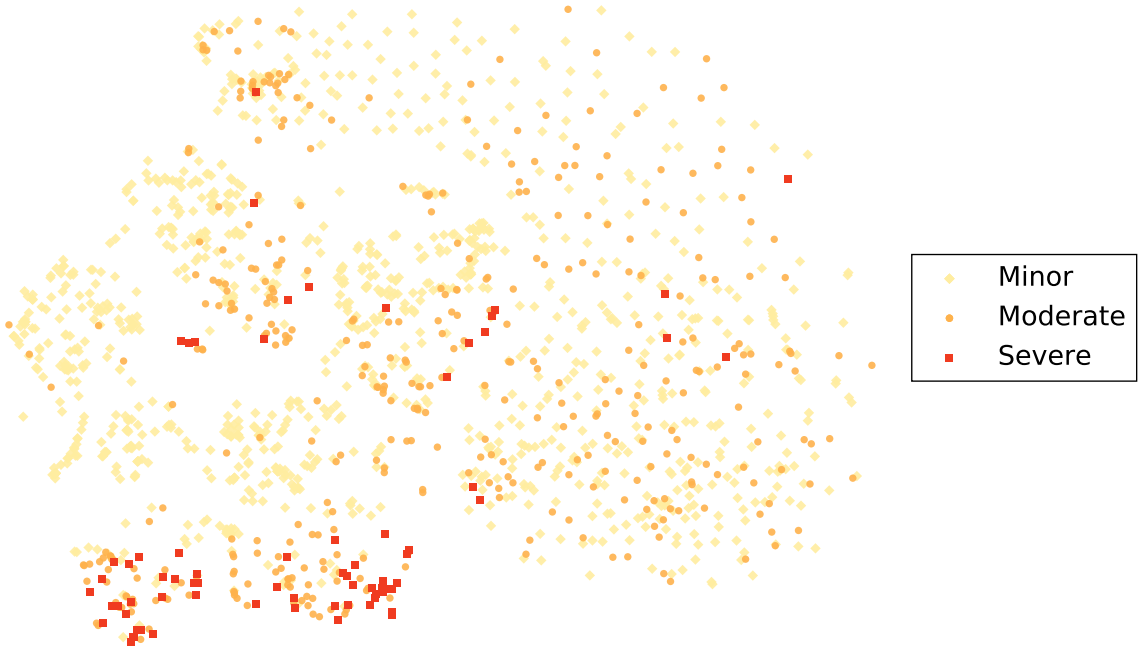
Using the two-dimensional representations of patient records, we labeled each record according to its unified Parkinson's disease rating scale (UPDRS) scores. The UPDRS consists of six sections, each containing survey questions that evaluate a patient's physical and mental condition.[20] The questions deal with topics ranging from anxiety to sleeping problems, with scores scaled from 0 to 4. A higher score indicates more severe impairment or disability. As in previous work, we labeled each patient with the sum of his or her UPDRS scores.[21]

The main difference between VisAGE and the baseline is that in Figure 2, moderately impaired patients (orange circles) on the right side of the plots were clustered more distinctly in the VisAGE visualization, while the same patients were less structured in the baseline visualization. The most severe Parkinson's disease patients are marked by red squares, and were clustered more tightly together in the VisAGE visualization than in the baseline. The baseline's worse performance can be attributed to the data sparsity of the EMRs.

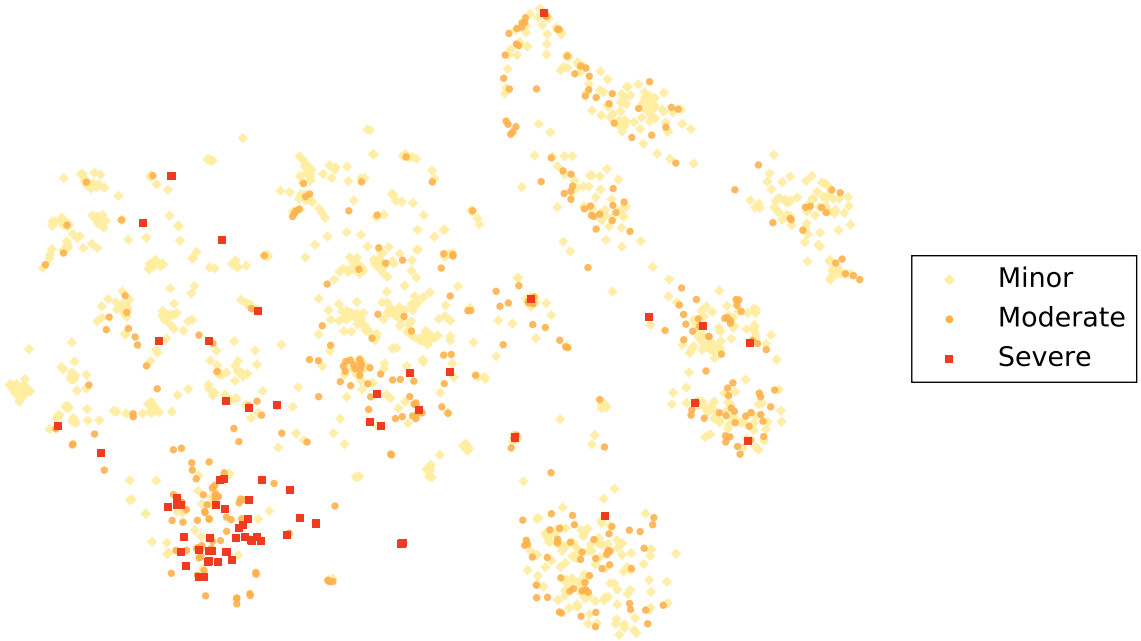
5.2. *Qualitative Evaluation: Drug and Symptom Enrichment*

We qualitatively evaluated the visualization results by computing drug and symptom enrichments for each cluster. We used symptoms and drugs because they are strongly connected to patient statuses and diagnoses. Thus, if a cluster is highly enriched in a symptom or drug, then doctors will have a general idea of the cluster's disease subtype. We first clustered the two-dimensional patient representations with DBSCAN,[22] which is robust to outliers and does not need to specify the number of clusters. Because the PPMI dataset contains control patients to simulate noise, DBSCAN's robustness to outliers is especially desirable. Additionally, not having to specify the number of clusters *a priori* is useful for our application, as we do not know the exact number of patient subtypes beforehand.

For the DBSCAN parameters, we set $\epsilon = 1$ and $minPts = 10$. In the baseline method, patients were placed into 10 clusters. With VisAGE, patients were placed into 18 clusters. For each cluster c and each symptom or drug b, we computed Fisher's exact test to determine if c was significantly enriched in b. We only used symptoms and drugs with binary values to avoid medical tests for which all patients had non-zero values (e.g., the Epworth Sleepiness Scale[23]).



(a) Baseline visualization



(b) VisAGE visualization

Fig. 2: The two-dimensional representations of patient records, plotted with color labels determined by each record's UPDRS scores. VisAGE's visualization identifies more clusters for moderately impaired patients, and more tightly groups severely impaired patients.

5.3. *VisAGE Discovers More Patient Subtypes*

We show the two-dimensional plot of patients in Figures 3 and 4, with each color-shape combination corresponding to a unique cluster generated by DBSCAN. We also show the two most enriched symptoms for each cluster in the legends. With the baseline, many of the points in the upper right quadrant of the plot were determined to be noise (black circles). As a result, no patients can be deemed to be similar to these noise points. On the other hand, VisAGE was able to properly classify many of these patients into distinct clusters.

We saw that both methods mostly grouped together the patients with the highest UPDRS scores (Figure 2). The corresponding DBSCAN clusters that overlapped with these high-UPDRS patients were most enriched in parkinsonism and Parkinson's disease, as expected.

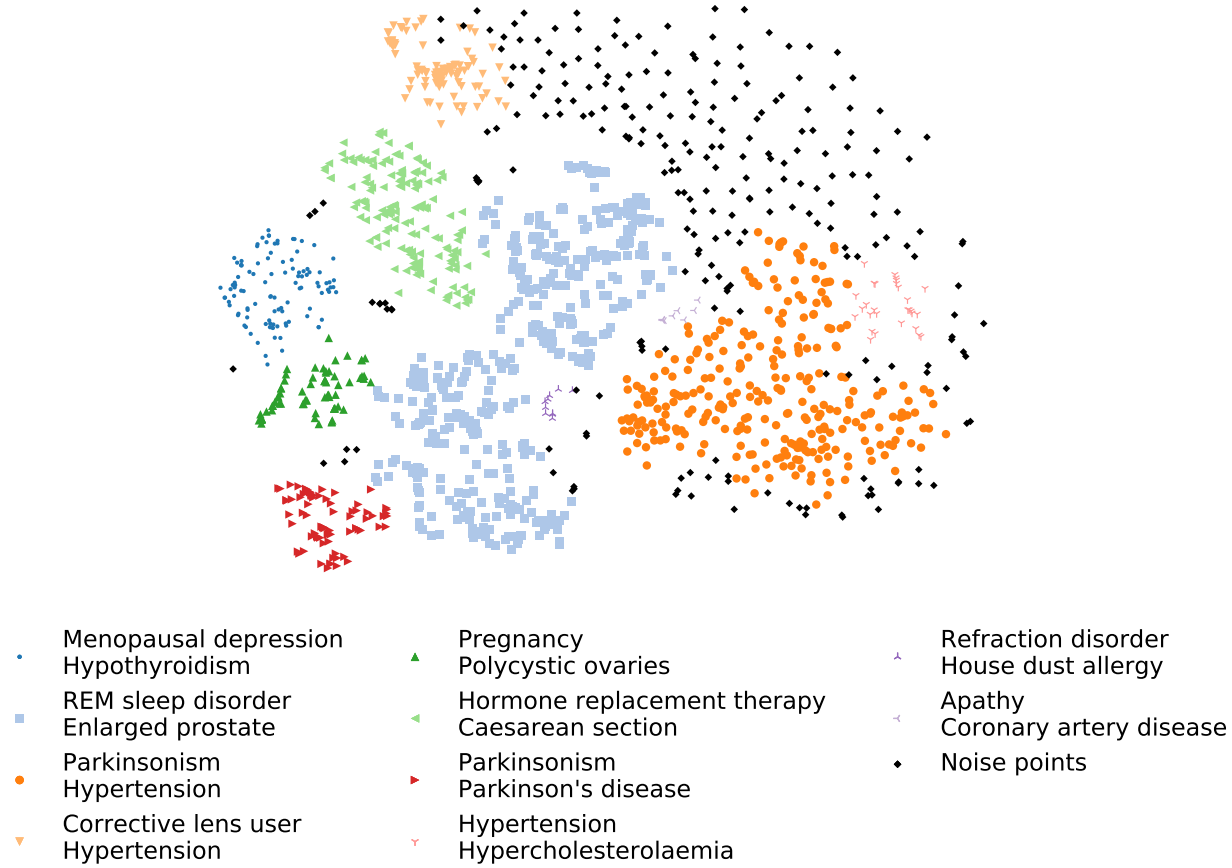


Fig. 3: The baseline's two-dimensional representation of patient records, with colors determined by the DBSCAN clustering.

Both methods identified a cluster of patients enriched in parkinsonism and hypertension (orange circles in the baseline and dark green triangles pointing up in VisAGE). Indeed, hypertension is commonly known to be prevalent in PD patients.[24] However, VisAGE identified four additional clusters that were significantly enriched in PD/parkinsonism and another informative symptom. On the other hand, the baseline method combined these clusters into

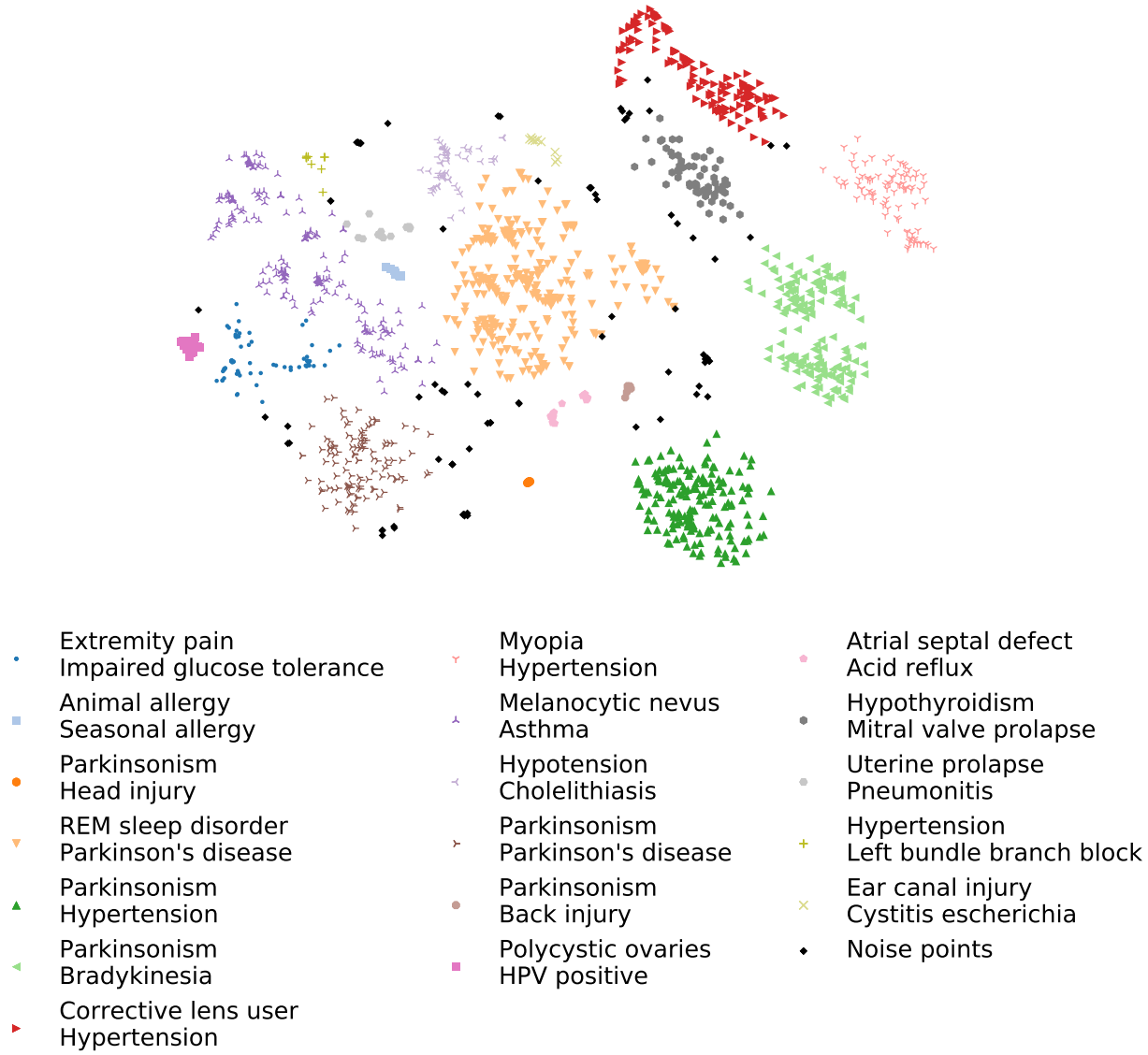


Fig. 4: VisAGE's two-dimensional representations of patient records, with colors determined by the DBSCAN clustering.

larger ones, losing information in the process. We interpreted these additional clusters as PD patient subtypes that required special treatment. We now discuss these four clusters.

(1) **Parkinsonism and head injury.** The cluster of dark orange circles contained 16 patients, and was enriched in parkinsonism and head injury with p-values of 3.110×10^{-4} and 0.01013, respectively. This is consistent with previous work, as head trauma is one of the most common candidates for PD causes.[25] This cluster was highly enriched in entacapone, levodopa, and carbidopa with p-values of 1.085×10^{-25}, 9.030×10^{-10}, and 7.099×10^{-5}, respectively. While levodopa/carbidopa (LC) is the most common drug prescribed to PD patients, entacapone is often prescribed as a supplementary drug to improve the efficacy

of LC.[26] As expected, these patients are also labeled as "Severe" in Figure 2, which would explain the need for this supplement. Furthermore, entacapone has been proposed as a possible treatment for traumatic brain injury.[27] In the baseline, this group of patients was incorrectly combined with the cluster most enriched in parkinsonism and hypertension.

(2) **REM sleep disorders and Parkinson's disease**. The cluster of light orange, down-pointing triangles contained 292 patients, and was enriched in rapid eye movement (REM) sleep behavior disorder, which is most often associated with PD (p-values of 1.477×10^{-5} and 7.246×10^{-3}, respectively).[28] In addition to the standard levodopa prescription (p-value $= 9.787 \times 10^{-23}$), the cluster was also highly enriched in clonazepam (p-value $= 0.004377$). Clonazepam administered with levodopa at bedtime has been shown to reduce REM sleep disorder symptoms.[29] In the baseline, the corresponding cluster contained nearly twice as many patients (458), and was not highly enriched in Parkinson's disease.

(3) **Parkinsonism and bradykinesia.** In VisAGE's visualization, the cluster of light green, left-pointing triangles contained 159 patients, and was enriched in parkinsonism and bradykinesia with p-values of 1.526×10^{-8} and 3.974×10^{-8}, respectively. As expected, bradykinesia is a key symptom of parkinsonism.[30] Additionally, this cluster was highly enriched in ropinirole with a p-value of 1.246×10^{-7}. Ropinirole stimulates mesolimbic D_3 receptors, which alleviates bradykinesia.[31] In the baseline, this group of patients was mixed with patients exhibiting parkinsonism and hypertension.

(4) **Parkinsonism and back injury**. The cluster of light brown circles contained 17 patients, and was enriched in parkinsonism and back injury with p-values of 1.320×10^{-5} and 1.092×10^{-4}, respectively. A previous study showed that spinal cord injuries are associated with increased risk of PD.[32] In addition to the standard levodopa/carbidopa prescription, this cluster was significantly enriched in amantadine (p-value $= 6.09 \times 10^{-5}$). Amantadine is not only an antiparkinsonian agent, but has also been shown to act as a non-competitive N-Methyl-D-aspartate (NMDA) receptor antagonist.[33] NMDA receptor antagonists have been shown to treat acute spinal cord injuries.[34] Like in the cluster that was enriched in parkinsonism and head injury, this cluster also contained many patients with severe UPDRS scores. In the baseline, these patients were again mixed with the cluster enriched in parkinsonism and hypertension.

5.4. *Quantitative Evaluation: False Discovery Rate*

For each method, we compared the number of clusters highly enriched in drugs and symptoms. To this end, we excluded drugs and symptoms from both M and M'. Additionally, we excluded these features from VisAGE's knowledge graph in order to limit data leakage. We then re-computed the enrichments for drugs and symptoms, taking the drug or symptom with the lowest p-value to represent each cluster. With these p-values, we counted the number of clusters that were significantly enriched in at least one drug or symptom.

To create a fair comparison, we used the Benjamini-Hochberg procedure[35] to control the false discovery rate at different levels of α (Figure 5). We see that VisAGE identified more enriched clusters than the baseline at every level of α, which is consistent with our earlier observation that the baseline method is incapable of distinguishing among patients with less

severe symptoms. Thus, we conclude that VisAGE also performs better quantitatively.

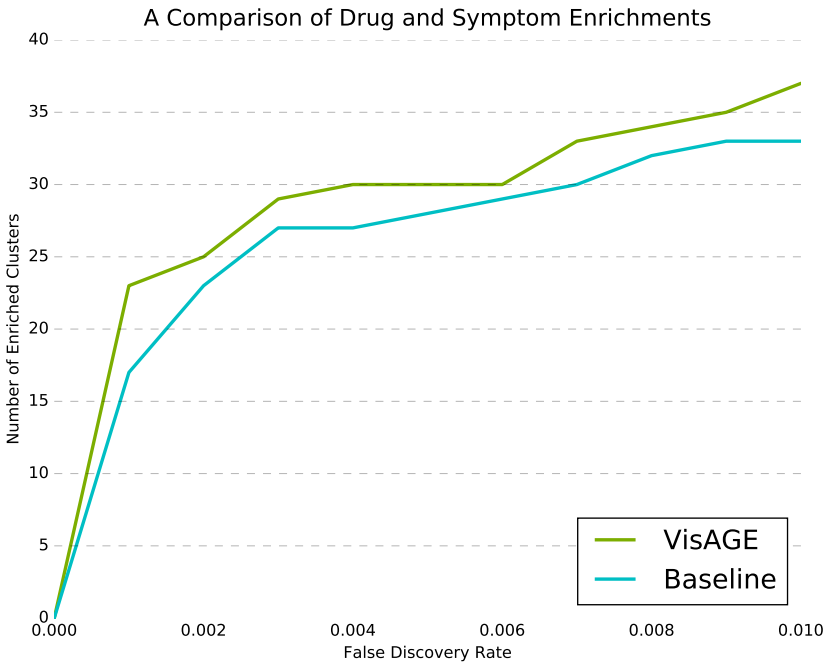


Fig. 5: A plot comparing the baseline and VisAGE. VisAGE dominates the baseline in the number of clusters enriched with at least one drug or symptom at each level of α.

6. Related Work

A previous study visualized high-dimensional data with a technique called LargeVis. However, it builds a k-NN network directly from the data, and then reduces the network to two dimensions without using external information.[18] Another study built upon LargeVis to visualize single cells, but still also directly computed embeddings from a k-NN network without utilizing external data.[17] Marlin *et al.* visualized a pattern discovery model's clustering parameters in the context of EMR analysis.[36] However, they focused on longitudinal data and predicting mortality outcomes rather than patient subtyping. Gotz *et al.* performed interactive visualization of EMR data, but worked with time series data to analyze patterns over time.[37] The Dynamic Icons (DICON) system clusters EMRs that are similar to a given patient, visualizing the clusters. However, it does not utilize molecular interaction networks or genomic data to compute similarities between EMRs.[38] Lastly, Perer *et al.* developed Care Pathway Explorer to visualize EMR data to investigate correlations with patient outcomes.[39] However, they use sequential pattern mining, which relies on historical EMR data to extract patterns.

7. Conclusions and Future Work

In this paper, we presented VisAGE, a method of improving EMR visualization by enriching EMRs with external knowledge sources. Evaluations on a PD patient dataset showed that VisAGE can generate visualizations such that similar patients are clustered together more tightly than in a baseline that does not alter the original database. We also evaluated our visualization with enrichments of drugs and symptoms, and showed that VisAGE can produce a higher quantity of fine-grained partitions of PD patients.

One limitation of our work is that the evaluation is done on only one dataset, which is mainly due to the necessity of expensive patient annotations. In the future, it is important to further evaluate the proposed enrichment method on more datasets as they become available. We also plan to build software that can implement our visualization in real application environments. Since VisAGE is a general method, the software would serve as a framework for an interactive component that can enrich any EMR database. For example, in a clinical setting, previously treated patients can serve as guidelines for doctors treating new patients. Doctors can identify these similar, previously treated patients in the two-dimensional space using the visualization tool and optimize treatment for the current patient.

8. Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144245. This research was also supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

References

1. S. T. Mennemeyer, N. Menachemi, S. Rahurkar and E. W. Ford, *Journal of the American Medical Informatics Association* **23**, 375 (2016).
2. J. Bae and W. E. Encinosa, *BMC health services research* **16**, p. 172 (2016).
3. J. Bae, J. M. Hockenberry, K. J. Rask and E. R. Becker, *Health care management review* **42**, 258 (2017).
4. A. Rind, P. Federico, T. Gschwandtner, W. Aigner, J. Doppler and M. Wagner, Visual analytics of electronic health records with a focus on time, in *New Perspectives in Medical Records*, (Springer, 2017) pp. 65–77.
5. A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, B. Shneiderman *et al.*, *Foundations and Trends® in Human–Computer Interaction* **5**, 207 (2013).
6. M. Ozkaynak, B. Reeder, L. Hoffecker, M. B. Makic and K. Sousa, *CIN: Computers, Informatics, Nursing* **35**, 465 (2017).
7. S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow and C. Neti, Predicting patients trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics, in *AMIA annual symposium proceedings*, 2010.
8. G. Hals and F. Lovecchio, *Emergency Medicine Reports* **30**, 145 (2009).
9. M. S. Hansen, G. J. Nogareda and S. J. Hutchison, *The American journal of cardiology* **99**, 852 (2007).
10. P. G. Hagan, C. A. Nienaber, E. M. Isselbacher, D. Bruckman, D. J. Karavite, P. L. Russman, A. Evangelista, R. Fattori, T. Suzuki, J. K. Oh *et al.*, *Jama* **283**, 897 (2000).

11. B. K. Nallamothu, R. H. Mehta, S. Saint, A. Llovet, E. Bossone, J. V. Cooper, U. Sechtem, E. M. Isselbacher, C. A. Nienaber, K. A. Eagle *et al.*, *The American journal of medicine* **113**, 468 (2002).
12. K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury *et al.*, *Progress in neurobiology* **95**, 629 (2011).
13. T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. M. Mercer, G. Slodkowicz, C. Workman, O. Regina, K. Rapacki, H.-H. Staerfeldt *et al.*, *bioRxiv* , p. 064535 (2016).
14. R. A. Fisher, *Journal of the Royal Statistical Society* **85**, 87 (1922).
15. D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork and M. Kuhn, *Nucleic acids research* **44**, D380 (2015).
16. S. Wang, M. Qu and J. Peng, Prosnet: Integrating homology with molecular networks for protein function prediction, in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2016.
17. J. Kim, N. Russell and J. Peng, Scalable visualization for high-dimensional single-cell data, in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, 2017.
18. J. Tang, J. Liu, M. Zhang and Q. Mei, Visualizing large-scale and high-dimensional data, in *Proceedings of the 25th International Conference on World Wide Web*, 2016.
19. L. v. d. Maaten and G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).
20. C. Ramaker, J. Marinus, A. M. Stiggelbout and B. J. Van Hilten, *Movement Disorders* **17**, 867 (2002).
21. C. Shi, Z. Zheng, Q. Wang, C. Wang, D. Zhang, M. Zhang, P. Chan and X. Wang, *PloS one* **11**, p. e0155758 (2016).
22. M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, A density-based algorithm for discovering clusters in large spatial databases with noise., in *Kdd*, (34)1996.
23. M. W. Johns, *sleep* **14**, 540 (1991).
24. A. A. Ejaz, I. S. Sekhon and S. Munjal, *European journal of internal medicine* **17**, 417 (2006).
25. S. M. Goldman, C. M. Tanner, D. Oakes, G. S. Bhudhikanok, A. Gupta and J. W. Langston, *Annals of neurology* **60**, 65 (2006).
26. R. A. Hauser, M. Panisset, G. Abbruzzese, L. Mancione, N. Dronamraju and A. Kakarieka, *Movement Disorders* **24**, 541 (2009).
27. R. D. Zafonte, J. Lexell and N. Cullen, *The Journal of head trauma rehabilitation* **16**, 112 (2001).
28. J. J. Gugger and M. L. Wagner, *Annals of Pharmacotherapy* **41**, 1833 (2007).
29. M. Stacy, *Drugs & aging* **19**, 733 (2002).
30. M. Hallett and S. Khoshbin, *Brain* **103**, 301 (1980).
31. W. H. Jost and D. Angersbach, *CNS drug reviews* **11**, 253 (2005).
32. T. Yeh, Y. Huang, H. Wang and S. Pan, *Spinal cord* **54**, 1215 (2016).
33. J. Kornhuber, G. Quack, W. Danysz, K. Jellinger, W. Danielczyk, W. Gsell and P. Riederer, *Neuropharmacology* **34**, 713 (1995).
34. A. I. Faden, J. Ellison and L. Noble, *European journal of pharmacology* **175**, 165 (1990).
35. Y. Benjamini and Y. Hochberg, *Journal of the royal statistical society. Series B (Methodological)* , 289 (1995).
36. B. M. Marlin, D. C. Kale, R. G. Khemani and R. C. Wetzel, Unsupervised pattern discovery in electronic health care data using probabilistic clustering models, in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012.
37. D. Gotz, J. Sun, N. Cao and S. Ebadollahi, Visual cluster analysis in support of clinical decision intelligence, in *AMIA Annual Symposium Proceedings*, 2011.
38. N. Cao, D. Gotz, J. Sun and H. Qu, *IEEE transactions on visualization and computer graphics* **17**, 2581 (2011).
39. A. Perer, F. Wang and J. Hu, *Journal of biomedical informatics* **56**, 369 (2015).

GeneDive: A gene interaction search and visualization tool to facilitate precision medicine

Paul Previde[1], Brook Thomas[1], Mike Wong[2], Emily K Mallory[3], Dragutin Petkovic[12], Russ B Altman[456], Anagha Kulkarni[1*]

[1]*Department of Computer Science*
[2]*Center for Computing for Life Sciences*
San Francisco State University
San Francisco, California 94132, U.S.A.
E-mail: {pprevide,bthomas}@mail.sfsu.edu, {mikewong, petkovic, ak}@sfsu.edu

[3]*Biomedical Informatics Training Program*
[4]*Department of Bioengineering,*
[5]*Department of Genetics,*
[6]*School of Medicine,*
Stanford University,
Stanford, California 94305, U.S.A.
E-mail: {emily.mallory,russ.altman}@stanford.edu

Obtaining relevant information about gene interactions is critical for understanding disease processes and treatment. With the rise in text mining approaches, the volume of such biomedical data is rapidly increasing, thereby creating a new problem for the users of this data: information overload. A tool for efficient querying and visualization of biomedical data that helps researchers understand the underlying biological mechanisms for diseases and drug responses, and ultimately helps patients, is sorely needed. To this end we have developed *GeneDive*, a web-based information retrieval, filtering, and visualization tool for large volumes of gene interaction data. GeneDive offers various features and modalities that guide the user through the search process to efficiently reach the information of their interest. GeneDive currently processes over three million gene-gene interactions with response times within a few seconds. For over half of the curated gene sets sourced from four prominent databases, more than 80% of the gene set members are recovered by GeneDive. In the near future, GeneDive will seamlessly accommodate other interaction types, such as gene-drug and gene-disease interactions, thus enabling full exploration of topics such as precision medicine. The GeneDive application and information about its underlying system architecture are available at http://www.genedive.net.

Keywords: Gene Interactions; Retrieval and Visualization; Gene Sets; Gene Networks.

*To whom correspondence should be addressed.

1. Introduction

A complete and accurate database of gene interactions (GI) is needed to understand cellular processes, disease states, and drug responses. Compiling such a database manually is a labor-intensive task, requiring highly-trained curators to identify each gene interaction and assess the quality of the supporting evidence. This approach is expensive both in time and in effort. Consequently, database curators have to limit scope by starting with key genes, drugs, or literature central to a given pathology or species of interest. This focus leads to curation bias towards well-studied or high-impact work but may miss important relationships that are not well-studied. Modern data mining techniques can scale to meet the challenge, and are automating the process of extracting interactions from biomedical literature.[1,9,13,15,19,20] However, the volume of extracted data is large, and the quality is mixed. These extracted interactions must thus be made accessible to curators and researchers in such a way as to facilitate analysis, curation, exploration, and discovery of well-understood and poorly-understood disease mechanisms and drug responses and, ultimately, to inform and improve patient care.

Current GI curation tools such as Cytoscape[17] and VisANT[4] have multi-species support for integrating and visualizing curated datasets (*e.g.*, GO, KEGG pathways). Literome[14] uses natural language processing techniques to identify and display GIs extracted from literature. Literome provides excerpts from supporting literature, so curators can assess and vote on the interaction accuracy. However, these tools lack other features that curators find useful. To mitigate curation bias, curators need to be able to to identify GIs that are not well-studied. To our knowledge, no current tool offers an automated solution to meet this need. Another feature helpful to curators is the ability to investigate local topologies of GI networks around a gene of interest. Cytoscape and VisAnt offer comprehensive visualization, interaction selection and network analysis, while Literome has more rudimentary features.

To address the challenges of curation bias and provide needed network analysis features, we introduce GeneDive, a web-based information retrieval, filtering, topology discovery, and visualization tool that is designed to process millions of interactions efficiently. GeneDive users can sift through large volumes of data and zoom in onto a small subset of interactions of their interest. GeneDive's graphical rendering leverages Cytoscape's comprehensive visualization library, which is already familiar to many users.

GeneDive is currently powered by a repository of over three million gene-gene interactions (GGIs) extracted using DeepDive, a scalable text-mining system.[11] The interactions were extracted from a corpus consisting of 100,000 PLOS articles and 340,000 PMC articles, and training data was extracted from BioGRID, ChEA, and Negatome.[9] Each extracted GGI includes an article excerpt that mentions a gene pair, and a computed probability that the relation is a true interaction. GGIs with probability of 0.9 or higher are typically true positive interactions. GGIs with slightly lower probabilities are likely to be true interactions that lack strong evidence in literature. In general, these marginal probabilities provide a unique mechanism to quickly identify poorly-studied but important GIs. GeneDive also provides other supporting evidence, such as the excerpt text, the number of instances extracted for a given GGI, and number of articles that mention the GGI.

In the near future, GeneDive's capabilities will be expanded beyond GGIs to include gene-

drug and gene-disease interactions. Such a comprehensive and large repository of gene interactions can help unlock previously poorly-understood disease mechanisms and drug responses. Overall, GeneDive is designed to facilitate data-driven discovery and hypothesis testing for precision medicine researchers.

2. GeneDive

GeneDive was developed using the agile software development approach,[3] which entails iterative development by a cross-functional team consisting of computer scientists, developers, biomedical researchers and curators. The high-level goal was to design and develop an easy-to-use and scalable web application with powerful search functionality and multi-modal result presentation capabilities. The resulting web application and its features are described next.

(a) Search input with autocomplete (b) Ambiguity resolution interface (c) Available search modalities

Fig. 1. GeneDive search entry screenshots.

Ambiguity Resolution: GeneDive users can search the interactions database by entering one or more gene symbols (from Entrez/NCBI) or gene set names (from Reactome, KEGG and many others). Figure 1(a) provides an illustration of search input entry. Gene symbols are sometimes ambiguous: for example, PSP-A may refer to either of two genes, SFTPA-1 and SFTPA-2, that encode distinct pulmonary surfactant proteins. In this case, information about the genes is provided, and the user is prompted to choose the intended gene (Figure 1(b)). All gene symbol ambiguities, if any, have to be resolved before proceeding.

Topology Search: GeneDive provides four search modes to help users investigate the GGI non-directional network topology: 1-hop, 2-hop, 3-hop, and Clique, as shown in Figure 1(c). The default mode is 1-hop, which searches for immediate neighbors, that is, genes that interact directly with the gene of interest. 1-hop retrieves all neighbors for a single-gene search, and pairwise GGIs for searches of two or more genes. For example, if the user has entered genes A, B, and C, then 1-hop search will return GGIs for the following gene pairs: AB, AC, and BC. 2-hop and 3-hop search modes require at least two gene symbols as input and retrieve

intermediary genes along a pathway. 2-hop shows paths with one or no intermediaries, and 3-hop shows paths with two or fewer genes between the genes of interest. The intermediate genes revealed in 2-hop, 3-hop, and Clique search modes may refine our understanding of a pathway or discover new pathways. Clique search mode accepts only one gene symbol, and retrieves all the *complete networks* that contain the input gene. A complete network is formed when every member gene directly interacts with every other gene in the network.

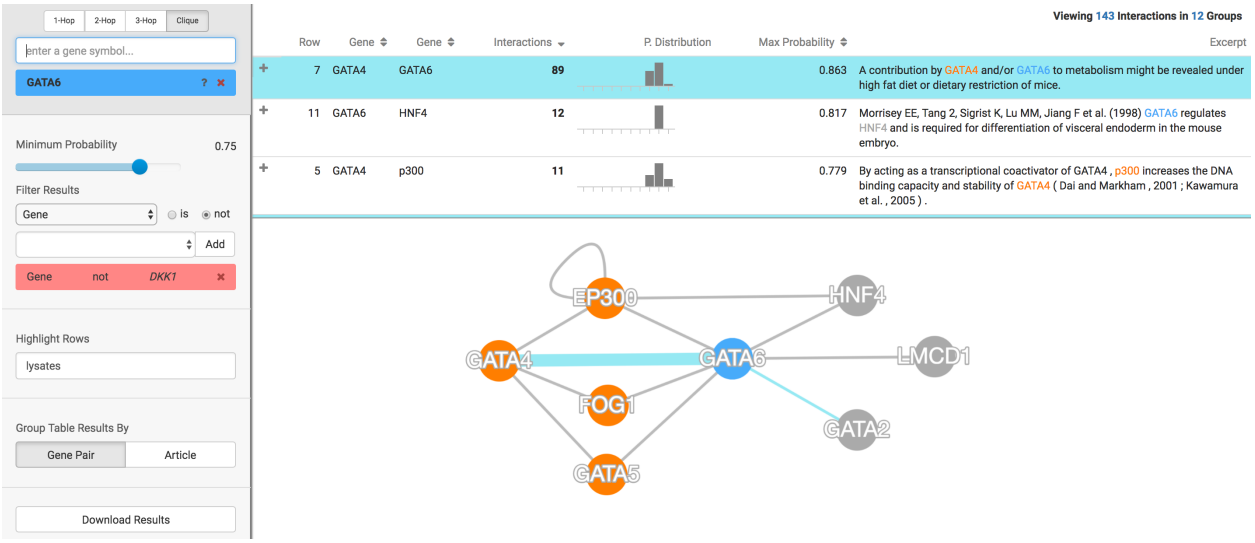


Fig. 2. Tabular (top) and graphical (bottom) views of GGIs. (Search & controls panel on left.)

Graphical and Tabular Views: Search results are presented in both tabular and graphical views, shown simultaneously and in parallel. Figure 2 shows a representative view of the results of a search for gene GATA6 in the Clique mode. Users may choose to view the resulting interactions grouped by gene pairs (default) or by articles. Grouping by gene pairs collates all GGIs for every unique gene-gene pair in the retrieved results into a single row in the Table view. The *Interactions* column specifies the number of GGI mentions over all articles. The number of unique articles from which these mentions are extracted is specified under *Articles*. A histogram of the probabilities for the interaction group is shown in the *P. Distribution* column, with the highest value shown in *Max Probability*. The source sentence from which the highest-probability GGI mention was extracted is shown in the *Sample Excerpt* column.

Grouped rows may be expanded to show individual interactions, each with a probability, extracted excerpt, name and hyperlink to the source article, and relevant section from the article. Nearly all of the columns are sortable in both the grouped and expanded views of the table, with descending probability as the default sort order. Gene mentions are color-coded by set membership; genes that belong to multiple sets are displayed in multiple colors. Color-coding is consistent between the graph and table views. The lower panel presents the search results in a graphical format using Cytoscape's force-directed layout engine, with nodes representing genes and edges representing interactions; edge thickness is proportional to the number of mentions for retrieved GGIs. Users can modify or change their search inputs by

clicking on a node to search for that gene, or shift-clicking on one or more nodes to search for multiple genes. This allows for easy navigation of the gene interaction network.

Results Filtering and Highlighting: The user may filter the search results using filters for one or more of the following: gene name, journal, specific article, document section (*e.g.*, abstract, methods, results), excerpt partial text matching, and minimum probability cutoff (the default cutoff is 0.7). Figure 2 shows the filter controls in the left panel and illustrates the negation filter: interactions containing *DKK1* as one of the genes will not be retrieved. The filters are additive; users can add or remove filters as desired. Users can highlight interactions in both views, calling attention to interactions that match the user-provided criteria (*e.g.*, excerpts containing specific terms). In the graphical view, edges that match the highlighting criteria are rendered with a highlight color; in the tabular view, matching rows are also highlighted in the same color. Figure 2 demonstrates highlighting for the term *lysates*. For example, the first row, *GATA4* and *GATA6*, is highlighted because at least one of the excerpts in that group contains the term *lysates*. The edge between the nodes for *GATA4* and *GATA6* is also highlighted in the graphical view.

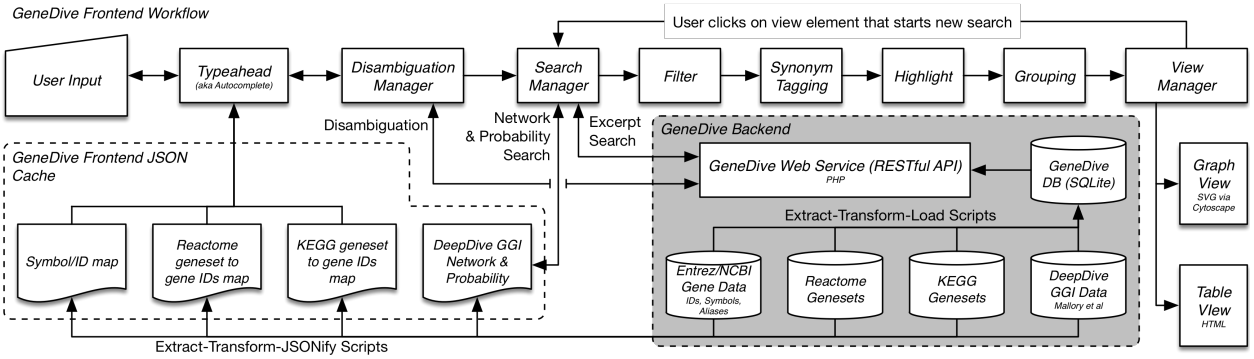


Fig. 3. GeneDive architecture diagram showing data sources and workflow.

Saving Results: GeneDive saves the system state; search, filter, and highlight parameters and results; graphical views; and table contents as a state file, which can be re-uploaded to resume where the user stopped during a prior session. Tabular views can be exported to Excel. Graphical views can be saved in scalable vector graphics (SVG) format, providing users with publication-ready images of interaction networks.

2.1 System Architecture

GeneDive is a single-page web application following the model-view-controller (MVC) architecture. User account management and web services are implemented in PHP. As shown in Figure 3, the web service comprises a RESTful API layer to a SQLite relational database containing complete interaction and gene data, extracted from Entrez/NCBI, Reactome, KEGG, and DeepDive-identified GGI data. All gene references are mapped to Entrez gene IDs, and all gene sets are filtered using the GGI data; if a gene in a set does not have a DeepDive-identified

interaction with any other gene in the same set, that gene is removed from the set. To improve responsiveness, most of the processing is done client-side through modules connected in serial, assisted by cached lookup tables. The modules manage filtering, coloring, synonym tagging, highlighting, grouping, and view rendering. Compressed caches include a GGI adjacency matrix with probabilities, and gene and gene set lookup tables for auto-complete.

2.2 Feature Comparison with Other Tools

Table 1 provides a side-by-side comparison of GeneDive features with those of three other widely-used search and/or visualization tools: Literome, Cytoscape, and VisANT, all of which were discussed in Section 1. This comparison considers 12 different features, such as whether the tool provides supporting evidence in the form of the excerpt from which the GGI was extracted, or whether the tool can search for GGIs associated with a known pathway, or whether it can support various levels of gene network analysis. As shown in Table 1, GeneDive supports more features than any of the other tools. GeneDive provides features that are traditionally supported by manually-curated databases as well as features typical of an automated database created using data mining.

Table 1. Feature matrix of gene interaction visualization tools.

Feature	GeneDive	Literome	Cytoscape	VisANT
GGI Probability	✓			
GGI Description		✓		
Supporting Literature Excerpt	✓	✓		
KEGG Pathways	✓		✓	✓
Reactome Pathways	✓		✓	✓
GO Term Enrichment			✓	✓
Tabular View	✓	✓	✓	
Graphical View	✓		✓	✓
Direct GGI (1-hop)	✓	✓	✓	✓
Intermediate GGI (2-hop)	✓	✓	✓	✓
2 Intermediates (3-hop)	✓		✓	✓
Clique	✓		✓	✓

3. Prominent Use Cases of GeneDive

The following use cases demonstrate the features and benefits of GeneDive for research and biomedical literature curation: single-gene searches, multi-gene searches, and searches involving related genes (gene sets or pathways).

Use Case 1: Single-Gene Query. To demonstrate GeneDive's features and capabilities for a single-gene search, we queried the system for interactions with the gene NOD2. NOD2 is involved in immune system pathways and is associated with Crohn's disease.[21] We input NOD2 in the search bar and limited the results to probability greater than 0.80. In the gene

pair group view, the result table contained interactions between NOD2 and 17 other genes. Each individual gene pair result row displayed the highest-scoring sentence for the given pair. By expanding an interaction row with a single gene, a user can display and browse supporting evidence for the interaction. For example, a user can select the row with RIPK2 and read eleven sentences that support the interaction from two different articles. In addition, the user can filter the results by specific article sections if they are more interested in sentence evidence from abstracts or results and discussion. The user also has the option to switch to the article group view in order to review all sentences with interactions from individual articles at once. This feature can hasten full-article curation. Using the gene pair group view, we discovered genes related to immune response (RIPK2,[16] TRAF6,[10] DUOX2[5]). After uncovering genes potentially related to the underlying mechanism for Crohn's disease, a user can query the resulting genes and construct networks with their associations. This example highlights the utility of single-gene searches for understanding the underlying mechanisms of a disease.

Use Case 2: Multi-Gene Query. Given a group of disease-associated genes, we would like to discover potential connections between the genes in the query set or between their interacting genes. For this example, we focused on genes related to follicular lymphoma. Translocation and subsequent overexpression of BCL2 is found in follicular lymphoma,[8] but an additional seven genes (EZH2, ARID1A, MEF2B, EP300, FOXO1, CREBBP, and CARD11) are used for predicting disease risk.[12] We queried the system using their official symbols for the seven risk genes. When entering MEF2B in the search bar, a response box is triggered, requesting the user to resolve an ambiguous symbol. In this case, MEF2B is a symbol for MEF2B and BORC28-MEF2B. For the seven genes, GeneDive returned one sentence for CREBBP and EP300 as evidence for an interaction. This sentence described their individual interactions with a third gene, c-Myb, but not between themselves. Next we investigated whether any of the risk genes had interactions three or less hops (*i.e.*, 3-hop) from the known causative gene BCL2. Using a probability cutoff of 0.80, we discovered four risk genes connected to BCL2 via a 3-hop search (EZH2, EP300, FOXO1, CARD11). From here, users can explore the intermediate genes and any interactions with pathway gene sets included in GeneDive.

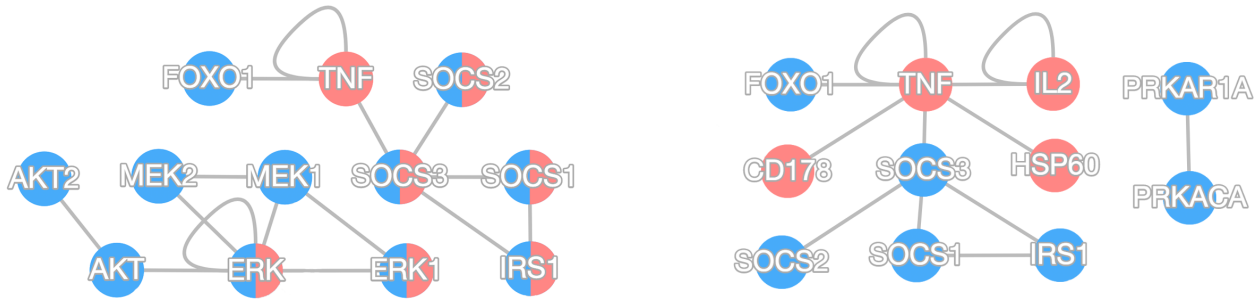


Fig. 4. Comparison of partial network overlap of two gene sets. Genes in the insulin signaling pathway set are colored blue. Genes with both colors belong to both sets. **Left:** Genes in the Type 2 diabetes mellitus pathway are pink. **Right:** Genes in the Type 1 diabetes mellitus pathway are pink.

Use Case 3: Pathway Gene Set Query. Type 2 diabetes mellitus (T2DM) is a complex dis-

ease characterized by insulin resistance, and over the last decade, a number of genetic variants have been shown to impact disease risk.[2] To demonstrate GeneDive's ability to detect potential pathway overlaps with a set of disease-related genes, we queried the system for two KEGG pathways included in the MSigDB gene sets: KEGG_TYPE_II_DIABETES_MELLITUS and KEGG_INSULIN_SIGNALING_PATHWAY. While a user can input a gene set of interest, we demonstrate this feature using a pre-loaded gene set for the T2DM pathway in KEGG.[b] From the network view (Figure 4, left), numerous genes and their interactions from the T2DM pathway overlap directly with the insulin signaling pathway. This contrasts with the KEGG pathway for Type 1 diabetes mellitus (T1DM) (KEGG_TYPE_1_DIABETES_MELLITUS). Here, T1DM genes interact with those from the insulin signaling pathway, but the T1DM genes are not found in the pathway itself (Figure 4, right). The lack of overlap between T1DM genes and the insulin signaling pathway is due to T1DM resulting from error in insulin production, not resistance.[6] By including pathway gene sets in GeneDive, users have the ability to not only query their own genes or gene sets of interest, but also to investigate possible disease and biological pathway associations.

4. Recall of Curated Gene Sets

While a good user interface is helpful to speed curator efforts, the quality of the underlying data is critical. The GeneDive data currently consists of 1,312,685 GGIs extracted from 340,000 PubMed Central ("PMC") articles, and 1,322,459 GGIs extracted from 100,000 Public Library of Science ("PLOS") articles. To assess the quality of this data, we attempt to *retrieve* curated gene sets using the GGI data. For this evaluation we use the C2 sub-collection of the GSEA MSigDB dataset, which consists of 4,731 curated gene sets[c]. The retrieval efficacy of C2 gene sets from GeneDive data is evaluated using precision and recall measures, with results shown in Table 2. We count false positives (FP) as follows: for each GGI in GeneDive data, if one gene is in any C2 gene set, and the other gene is not in any C2 gene set, then the other gene is labeled as FP. If both genes in the GGI are in the same C2 gene set, then they are identified as true positives (TP); each unique gene is counted as TP only once, even if it appears in multiple sets. If a gene in any gene set is not labeled as TP, it is labeled as a false negative (FN). Precision is calculated as $\frac{TP}{TP+FP}$, recall as $\frac{TP}{TP+FN}$, and F1 score as $\frac{2 \cdot precision \cdot recall}{precision+recall}$.

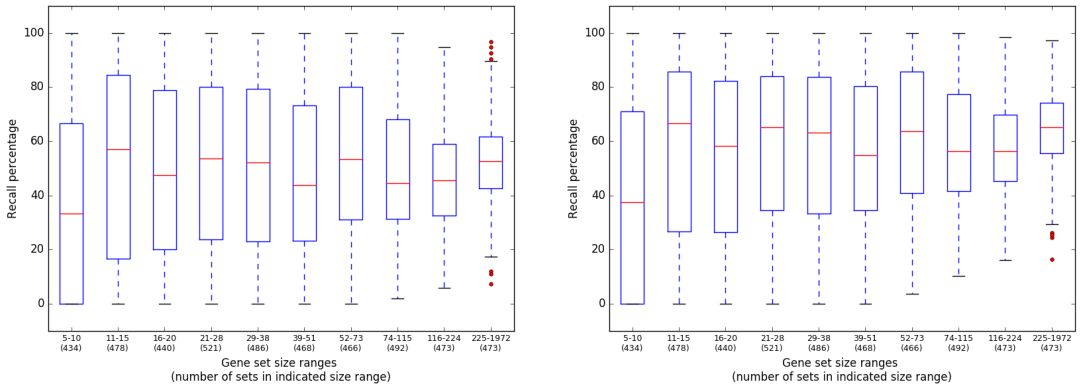
Table 2. Precision, Recall, and F1 Scores for PLOS and PMC corpora.

Cutoff	PLOS:	Precision	Recall	F1	PMC:	Precision	Recall	F1
0.0		0.941	0.651	0.770		0.941	0.561	0.703
0.5		0.953	0.521	0.674		0.953	0.460	0.620
0.7		0.977	0.246	0.393		0.971	0.259	0.409
0.9		0.985	0.096	0.175		0.983	0.124	0.220

These numbers suggest that the GeneDive data is fairly clean but not complete in the

[b] `http://www.kegg.jp`.
[c] `http://software.broadinstitute.org/gsea/msigdb/collections.jsp`.

context of C2 gene sets, so we investigate further into the sets themselves to better understand the completeness gap. 27% of these gene sets are contributed by either the Reactome pathway database (674 gene sets contributed), the Biocarta pathway collection (217 sets), the Pathway Interaction Database ("PID", 196 sets), or the Kyoto Encyclopedia of Genes and Genomes ("KEGG", 186 sets). The remaining gene sets are from over nine hundred other institutions and individual contributors. C2 gene sets are of highly variable size: the smallest contain 5 genes, and the largest 1,972. The median and the mean gene set size are 39.0 and 93.2, respectively. As a result, we bin the gene sets based on their sizes in the following analysis. For each gene set X in C2, its member genes x_1 and x_2 are labeled true positives if GeneDive data contains one or more interactions between x_1 and x_2.

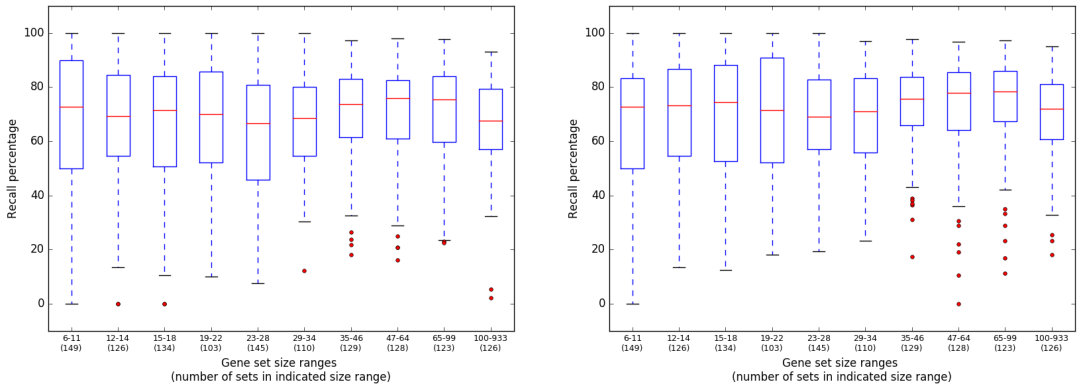


(a) PMC (b) PLOS

Fig. 5. Gene set recall plots for 4,731 gene sets in the MSigDB C2 collection, using GGIs extracted from (a) PMC, and (b) PLOS articles. The gene sets (binned by size) are along the X-axis, and recall percentage along the Y-axis. The median recall percentage for each bin is denoted by the horizontal line in the corresponding box.

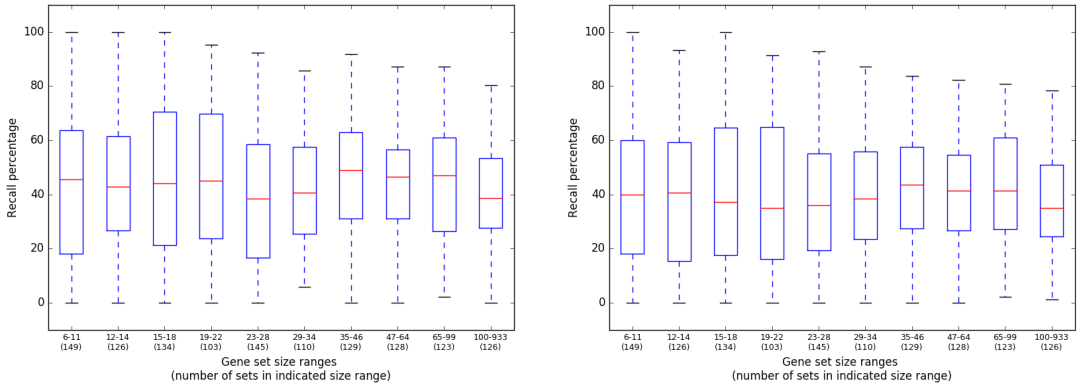
The results for the gene set recall analysis are given in Figure 5. The left plot shows that for half of the gene sets in a bin, about 50% or more of its members are recovered by PMC interactions. The recall percentages with PLOS interactions are consistently higher than those with PMC. The median recall in the first bin (size: 5 to 10 members) is the lowest for both PMC and PLOS. The reason for this observation becomes clearer when the sources of the gene sets are factored into this analysis. The probabilities assigned to the interactions also significantly influence gene set recall percentages. As such, the remainder of this section investigates recall percentages in the context of specific gene set sources and probability cutoffs.

The distribution of GeneDive probability scores resembles a normal distribution. If we filter out interactions with low probability of accuracy, then at a probability cutoff of 0.5, 563,383 interactions from PMC qualify, and 524,140 from PLOS. At 0.7 cutoff, only about 10% of the original collection remains: 133,567 interactions from PMC, and 112,282 from PLOS. Predictably, in regard to matching C2 gene sets, recall percentages decrease as more data is filtered out: at 0.5 cutoff, about 30% to 40% to the gene set members are recovered for half of the gene sets in the bin. At 0.7 this drops to 10% to 15%.



(a) PMC (b) PLOS

Fig. 6. Gene set recall plots for MSigDB gene sets from Reactome, Biocarta, PID, and KEGG, with probability cutoff of 0.5, extracted from (a) PMC, and (b) PLOS collections.



(a) PMC (b) PLOS

Fig. 7. Gene set recall plots for MSigDB gene sets from Reactome, Biocarta, PID, and KEGG, with probability cutoff 0.7, extracted from (a) PMC, and (b) PLOS collections.

If we restrict the analysis to the top four gene set contributors (*i.e.*, Reactome, Biocarta, PID, and KEGG), then the recall percentages, even when using only the high-quality interactions, are substantially higher than when using all the gene sets. At 0.5 cutoff, the recall percentage for all the gene sets was 30% to 40%, while for the gene sets from the top four contributors it is 70% to 75% (Figure 6). At 0.7 cutoff, the recall percentage for the gene sets contributed by the top four sources is about 40% (Figure 7), while for all gene sets it is 10% to 15%. This trend also continues in a more fine-grained analysis: at probability cutoff of 0.7, across all gene set sizes, 237 gene sets have 70% or greater recall; of those, 207 (87.3%) are contributed by Reactome, Biocarta, PID, or KEGG. Conversely, out of the 132 gene sets that had a recall percentage of less than 10%, only 2 (1.5%) are contributed by the four sources.

The cause for this trend is illustrated next using two gene sets, one with low recall and the other with high. An exemplary six-member gene set[d] contributed by Kasler *et al.* relates to

[d]`http://software.broadinstitute.org/gsea/msigdb/geneset\_page.jsp?geneSetName=`

the set of genes up-regulated in DO11.10 hybridoma cells by expression of HDAC7, a signal-dependent repressor of gene transcription during T-cell development. Using either the PLOS or PMC interaction data, the recall percentage of this set is 0% at a 0.7 probability cutoff. Even the primary publication[7] for this gene set does not explicitly mention, by symbol or by numerical gene identifier, any of the six genes in that set. Many other articles mention the gene set name, but the member genes are rarely mentioned. As such, the text mining approach used to extract GIs is unable to discover any of the member genes with high confidence.

A high-recall gene set is the Reactome ethanol oxidation set. This ten-member gene set had 100% recall at 0.7 probability cutoff with PMC interaction data. The Reactome documentation[e] of this gene set includes publications as early as 1949 and as recent as 2007, nearly 60 years of research. Thus, the interactions of this gene set's members are well-documented, leading to the high recall. The significantly lower recall percentages associated with small gene sets (gene sets with 5-10 members) are also explained by consideration of gene set sources. Figures 6 and 7 show that recall percentage for small gene sets contributed by the top four sources is on par with recall of larger gene sets. The smaller gene sets from other sources, however, experience close to 0% recall. Collectively, Reactome, Biocarta, PID, and KEGG contribute only 55 small gene sets, while all other sources contribute 379 small gene sets.

5. Conclusions

We have developed GeneDive, a web application facilitating efficient retrieval and exploration of a large number of gene interactions. GeneDive can reveal direct and indirect relationships between two or more genes in both tabular and graphical views. Since every interaction has a confidence score (probability) and an excerpt from a cited article, the user can easily collate evidence to support or refute the hypothesis being tested. GeneDive has been designed and developed in a data-agnostic fashion so as to seamlessly process different types of interactions. Gene-drug and gene-disease interactions are slated to be integrated in GeneDive soon, making GeneDive a powerful tool for precision medicine researchers that facilitates better understanding of genetic contributions to disease and treatment outcomes.

Funding

This work was partially supported by NIH grant LM005652, and by Center for Computing for Life Sciences at San Francisco State University. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

1. Davis AP, Wiegers TC, Roberts PM, King BL, Lay JM, Lennon-Hopkins K, Sciaky D, Johnson R, Keating H, Greene N et al. A CTD-Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug-disease and drug-phenotype interactions. Database: The Journal of Biological Databases and Curation (2013), bat080. PMC. Web. 25 May 2017.

KASLER_HDAC7_TARGETS_2_UP.
[e]http://www.reactome.org/content/detail/71384.

2. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI et al. Type 2 diabetes mellitus. Nature reviews Disease primers (2015), 1:15019.
3. Highsmith J, Cockburn A. Agile software development: The business of innovation. Computer (2001), 34(9):120-127.
4. Hu Z, Mellor J, Wi J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics (2004), 5(1):17.
5. Joo JH, Ryu JH, Kim CH, Kim HJ, Suh MS, Kim JO, Chung SY, Lee SN, Kim HM, Bae YS et al. Dual oxidase 2 is essential for the toll-like receptor 5-mediated inflammatory response in airway mucosa. Antioxidants & redox signaling (2012), 16(1):57-70.
6. Katsarou A, Gudbjornsdottir S, Rawshani A, Dabelea D, Bonifacio E, Anderson BJ, Jacobsen LM, Schatz DA, Lernmark A. Type 1 diabetes mellitus. Nature reviews Disease primers (2017).
7. Kasler HG, Verdin E. Histone deacetylase 7 functions as a key regulator of genes involved in both positive and negative selection of thymocytes. Molecular and Cellular Biology (2007), 27(14).
8. Kridel R, Sehn LH, Gascoyne RD. Pathogenesis of follicular lymphoma. The Journal of clinical investigation (2012), 122(10):3424-3431.
9. Mallory EK, Zhang C, Re C, Altman RB. Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics (2016), 32(1):106-113.
10. Manna SK, Ramesh GT. Interleukin-8 induces nuclear transcription factor-kappaB through a TRAF6-dependent pathway. The Journal of biological chemistry (2005), 280(8):7010-7021.
11. Niu F, Zhang C, Re C, Shavlik J. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. VLDS, 884, 2528.
12. Pastore A, Jurinovic V, Kridel R, Hoster E, Staiger AM, Szczepanowski M, Pott C, Kopp N, Murakami M, Horn H et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. The Lancet Oncology (2015), 16(9):1111-1122.
13. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2012).
14. Poon H, Quirk C, DeZiel C, Heckerman D; Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics (2014), 30(19):2840-2842.
15. Poon H, Toutanova K, Quirk C. Distant supervision for cancer pathway extraction from text. Pacific Symposium on Biocomputing (January 2015), 20:12031.
16. Ruefli-Brasse AA, Lee WP, Hurst S, Dixit VM. Rip2 participates in Bcl10 signaling and T-cell receptor-mediated NF-kappaB activation. The Journal of biological chemistry (2004), 279(2).
17. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, (November 2003), 13(11):2498-504
18. Subramanian A, Tamayo A, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102.43 (2005), 15545-15550.
19. Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering drugdrug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics (2010), 26(18).
20. Vazquez M, Krallinger M, Leitner F, Valencia A. Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Molecular Informatics (June 2011), 30(6-7):506-519,
21. Yamamoto S, Ma X. Role of Nod2 in the development of Crohn's disease. Microbes and infection (2009), 11(12):912-918.

Annotating gene sets by mining large literature collections with protein networks

Sheng Wang[1,*], Jianzhu Ma[2,*], Michael Ku Yu[2], Fan Zheng[2], Edward W Huang[1], Jiawei Han[1], Jian Peng[1,#], Trey Ideker[2,#]

[1]*Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA*
[2]*School of Medicine, University of California San Diego, San Diego, CA, USA*
[*]*These authors contributed equally to this work*
[#]*Email: jianpeng@illinois.edu, trey@bioeng.ucsd.edu*

Analysis of patient genomes and transcriptomes routinely recognizes new gene sets associated with human disease. Here we present an integrative natural language processing system which infers common functions for a gene set through automatic mining of the scientific literature with biological networks. This system links genes with associated literature phrases and combines these links with protein interactions in a single heterogeneous network. Multiscale functional annotations are inferred based on network distances between phrases and genes and then visualized as an ontology of biological concepts. To evaluate this system, we predict functions for gene sets representing known pathways and find that our approach achieves substantial improvement over the conventional text-mining baseline method. Moreover, our system discovers novel annotations for gene sets or pathways without previously known functions. Two case studies demonstrate how the system is used in discovery of new cancer-related pathways with ontological annotations.

Keywords: text mining, functional annotations, knowledge network, gene interactions

1. Introduction

With significant advances in ‘omics technologies, it has become increasingly routine to identify functionally related sets of genes based on different biological patterns. For example, a gene set may be computationally derived based on differential expression[1,2], based on associations to the same phenotypes[3,4], or based on a high density of molecular interactions among the genes[5–8]. Because of their functional relationships, these gene sets can often be interpreted as cellular pathways or protein complexes, enabling a systems approach to studying human diseases beyond individual genes[2-5].

Given a gene set of interest, a critical task is to learn what is its overall function as a pathway or complex in the cell. There are two major approaches to address this task. The first approach is to search for significant overlap with known pathways in manually curated databases such as the Gene

Ontology (GO)[9] and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[10]. However, it is very likely that little or no overlap can be found due to the limited coverage of these databases, especially when querying with gene sets related to a rare disease.

The second approach is to search for scientific articles that describe each gene in the set, and then summarize these articles to describe the aggregate function of the gene set. Manually performing this process requires substantial domain knowledge and does not scale to large pathways. While automatic summarization of free text has been proposed by many text-mining methods[11-12], these methods can describe only one gene rather than a gene set. In particular, automatic summarization for a gene set requires addressing several new challenges. First, the increased number of free text articles introduces diverse and noisy annotations compared to individual genes. Second, the relationship between pathway functions and gene interactions should be considered, since genes can perform very different functions when participating in different biological processes. Third, literature contains many potential and diverse function annotations, only some of which are relevant. Thus researchers need systematic approaches to filter, organize and display the most useful information in literature to better understand the biological pathways represented by a gene set. Many related approaches mine literature data to study the functions of a group of genes together. CoCiter tests the significance of co-citation of a gene set either from a user-defined queried gene sets or a known pathway[13]. Since the functions of this gene set are provided by user, CoCiter is not able to automatically mine new functional annotations to describe the gene set. Martini is a gene set comparison tool which assesses the similarity of two gene sets by using keywords extracted from Medline abstracts[14]. Although gene sets are compared using keywords, the functional description for each gene set is not explicitly generated.

Here we develop a novel approach to automatically mine functional annotations of pathways from a large corpus of literature supported by biological networks. Our approach has two major advantages over previous text mining methods. First, it integrates semantic information derived from literature with biological information derived from experimental and interactome data. In this framework, annotations and genes are linked through a comprehensive similarity network. By propagating information in the network, an annotation can be assigned to a gene even when the two were never mentioned together in the same literature. Second, we adopt a new way to organize and visualize functional annotations using a data structure called a “Hierarchical Concept Ontology”. This ontology reduces redundant information and visual complexity to display the complex structure embedded in the network. We evaluate our method on both manually-curated pathway annotations and gene sets derived from computational tools. We observe substantial improvement in predicting the manually curated annotations in comparison to a text-mining baseline (non-network) approach. We further explore two case studies to demonstrate how our method can combine text mining, molecular networks and advanced visualization to discover new pathways related to cancer.

2. Methods

Our method consists of four major steps (**Fig. 1**). First, it constructs a vocabulary of high quality phrases (a sequence of one or more words) by processing a large corpus of PubMed journal articles[15] using a software AutoPhrase[16]. Second, phrases are connected within a weighted network based on their probability of co-occurrence within the same articles. Third, our method builds a phrase-gene similarity network by joining the phrase-phrase network with an existing gene-gene network derived from experimental data. Fourth, phrases are ranked by how well they describe the function of a gene set. Finally, top-ranked phrases are projected into a low-dimensional space and hierarchically clustered to create a Concept Ontology.

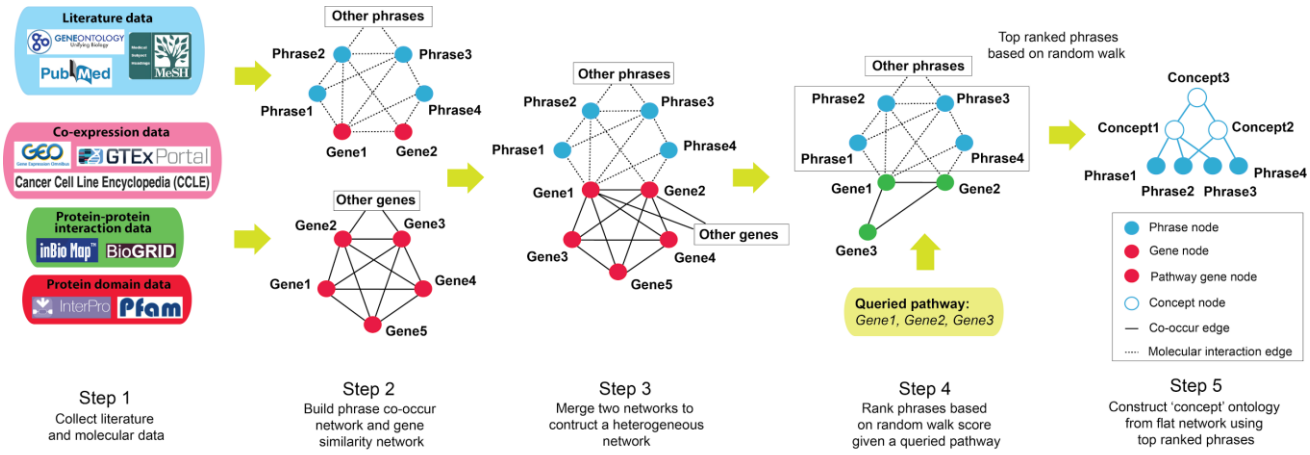


Figure 1. Diagram of our method

2.1 Constructing a phrase-gene network

We construct a weighted network to quantify the functional similarities between both phrases and genes. The edge weight w_{AB} between phrase *A* and phrase *B* is defined as:

$$w_{AB} = \frac{Pr(A,B)}{Pr(A)Pr(B)} \quad (1)$$

where *Pr(A)* is the marginal probability that phrase *A* appears in any article and *Pr(A, B)* is the probability that phrase *A* and phrase *B* co-occur in the same article. Intuitively, two phrases receive a large edge weight if they co-occur together more often than expected given their individual probabilities. In practice, non-informative phrases such as 'cell lines' and 'system biology' have many network neighbors with low edge weights; thus we retain only the top 50 edges for each phrase. To calculate the edge weight between two genes, we integrate multiple heterogeneous data sources, including gene co-expression, protein-protein interaction, protein-domain co-occurrence and genetic interaction (see **section 3.1**). We perform this integration in an unsupervised fashion using a network-

fusion-based algorithmic framework[17]. To calculate the edge weight between a phrase and a gene, the name of the gene is considered as a phrase and the weight is then calculated by **Eqn. 1**. In this way, the phrase-phrase and gene-gene networks are joined into a single network consisting of both phrases and genes as nodes.

2.2 Ranking candidate annotations of a pathway

Based on connections in this initial phrase-gene network, we further identify non-obvious links between phrases and genes through a random walk transformation of the network. An association score between gene *A* and phrase *B* is defined as the probability of randomly walking from *A* to *B* in the network, with restart probability = 0.5. Similarly, the association score between a queried gene set (pathway) and a phrase is defined as the average association score between the phrase and all genes in the set. We then rank pathways based on these scores. To efficiently rank a large number of phrases in a reasonable time, we only consider phrases that are within a distance of <3 to any of the genes in a queried pathway. Use of this filter in practice did not result in any significant decrease in performance (as evaluated below). Finally, we select all phrases with scores above a threshold as the candidate annotations of the queried pathway. We will discuss how to empirically pick this threshold in the below 'Experimental results' section.

2.3 Visualizing results as a Concept Ontology

The number of candidate annotations returned by the previous step can be very large, especially for large pathways. In general, synonyms are connected by the strongest weights because they are exchangeable in the literature. Phrases related to the same topic such as 'tumor suppressor' and 'driver mutations' will also be assigned strong weights but weaker than synonyms. Such intuition encouraged us to organize the flat phrase networks into a data-driven hierarchical 'concept' ontology[18-19]. For this purpose we adopt a network embedding approach[17] in which phrases are projected into a low-dimensional space and the cosine of two phrase embedding vectors is used as their pairwise distance. Given this new distance matrix, we then apply a network clustering approach, CLiXO[18], to transform the flat phrase network into a data-driven 'concept' ontology, where leaf nodes are phrases and internal nodes are clusters of similar phrases suggestive of higher order 'concepts'. Low-level concepts tend to be relatively concrete, because all phrases are strongly connected with each other, while high-level concepts tend to be more abstract, because phrases are more loosely connected with each other. Similar to a manually curated ontology, we assign each concept a name using a representative phrase having minimum distance with all the other phrases in the same concept cluster. Cytoscape[20] is then applied to visualize the data-driven Concept Ontology.

3. Experimental results

3.1 Dataset and experimental settings

We obtained 33,462,308 journal articles from PubMed published between 1994 to 2017. For each article, we only used the abstract and title rather than the whole article. We obtained 41,367 gene descriptions and gene name synonyms from NCBI[15]. The lengths of descriptions ranged from 100 to 300 words. AutoPhrase[16] then identified 727,289 phrases from the text corpus combining both gene descriptions and journal articles.

To calculate gene similarities, we aggregated various types of molecular networks using a Random Forest (RF) model trained to best recover the GO semantic distance between gene pairs. The trained model can be viewed as a nonlinear weighting of different kinds of features to reflect statistical pairwise correlations between two genes. The integrated data sources include coexpression networks, protein-protein interaction networks and protein-domain co-occurrences and genetics interactions, as follows. For co-expression networks, we used 980 genome-wide datasets extracted from the Gene Expression Omnibus (GEO) database[21]. We also used co-expression networks from the Genotype-Tissue Expression (GTEx)[22] project in which both global and tissue-specific co-expression are considered. In addition, we calculated a co-expression matrix on both the Human Protein Atlas and the Cancer Cell Line Encyclopedia[23,24]. For protein-protein interaction networks, we included all interactions in InBioMap[25] and only physical interactions in BioGRID[26]. In addition, we included genetic interaction data inferred from radiation hybrid genotypes[27] and domain co-occurrence data from InterPro[28] and PFAM[29].

3.2 Performance

3.2.1 Recovering curated names in GO

We examined the ability of our method to recover the names of known biological processes and cellular components in GO, given only information about their sets of annotated genes. For each GO term, we looked for its curated name among all candidate phrases ranked according to their association scores to the genes in the term (Section 2.2). Gene-term annotations were taken using experimental evidence codes (EXP, IDA, IPI, IMP, IGI, and IEP) but not *in silico* codes (e.g. IEA) to avoid potential leakage of labels.

We found that for 40% of terms in the biological process branch of GO, the curated name was among the top 50 candidates (**Fig. 2a**). Similarly, for 50% of terms in the cellular component branch, the curated name was among the top 50 candidates (**Fig. 2b**). More generally, we calculated the proportion of GO terms for which the curated name was among the top *K* candidate names of the term. For comparison, we set up a baseline approach in which a phrase is scored and ranked simply by the number of articles that mention this phrase together with any of the genes in the gene set. This simple but intuitive baseline mimics a search engine that ranks documents based on word frequency[30].

Our method substantially outperformed this baseline approach in naming terms across all three branches of GO (**Fig. 2a-c**). Here, for each term we only considered the curated name itself and did not reward returning the names of ancestors or descendants. In practice, however, we also observed the names of ancestors and descendants among the top ranked phrases (**Fig. S1-3**).

Further examining these results, we observed that the rank of the curated name identified by our method was positively correlated with the size of the gene set (**Fig. 2d**). That is, our method predicted more accurately when the gene set was small. For sets with fewer than 250 genes, our method found the correct curated term among the top 10 phrases the majority of the time. When the gene set was larger than 750 genes, our method could only detect the curated name among the top ~75 phrases. An explanation for this result is that large gene sets tend to cover broad or diverse functions and thus are more difficult to summarize by a short phrase.

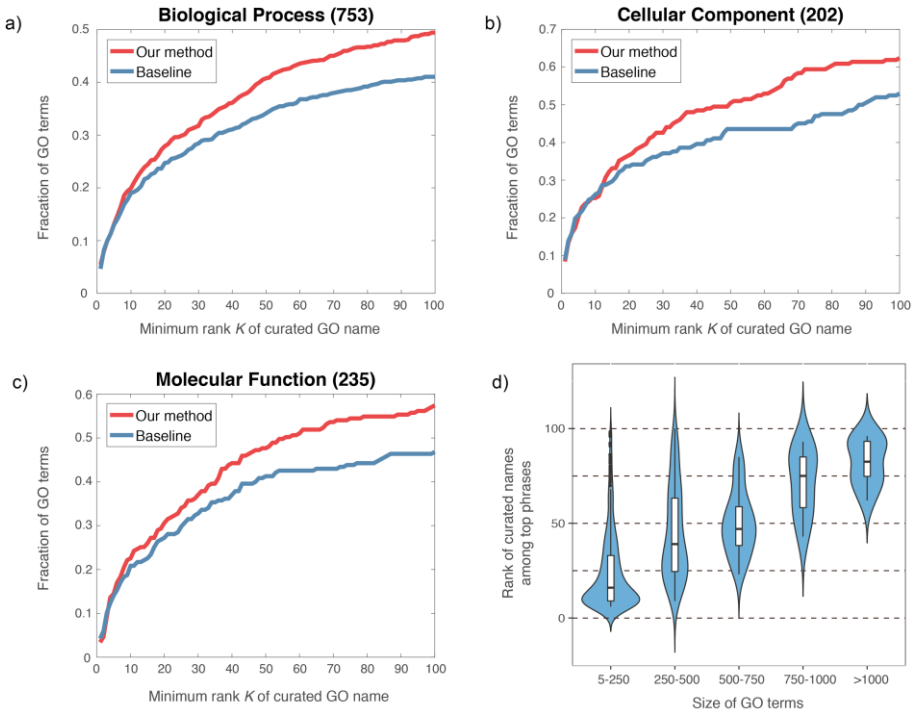


Figure 2. Comparison of our method and baseline on recovering term names of three Gene Ontology categories: Biological Process (a), Cellular Component (b) and Molecular Function (c). The fraction of terms for which the curated name was among the top *K* candidate phrases. (d) The correspondence between the rank of term names and the sizes of terms. The Y-axis shows the distribution of ranks of curated names with varying sparsity levels shown in the X-axis.

Next, we studied another critical problem: Given a ranking of phrases, how do we determine the threshold to select the most relevant phrases? To address this problem, for each GO term, we compared the ranking of its curated name with its association score. As shown in **Figs. 3a-b**, better rankings of curated names were generally tied to stronger association scores. This implies that the association scores across different GO terms are comparable. Therefore, we applied a universal threshold on the association score to determine final annotations for every GO term. We found that when the score is larger than -6 (log domain), we could always find the curated name among top 40

ranked phrases, regardless of term size (**Fig. 3b**). Therefore, we used -6 as our universal threshold to determine annotations.

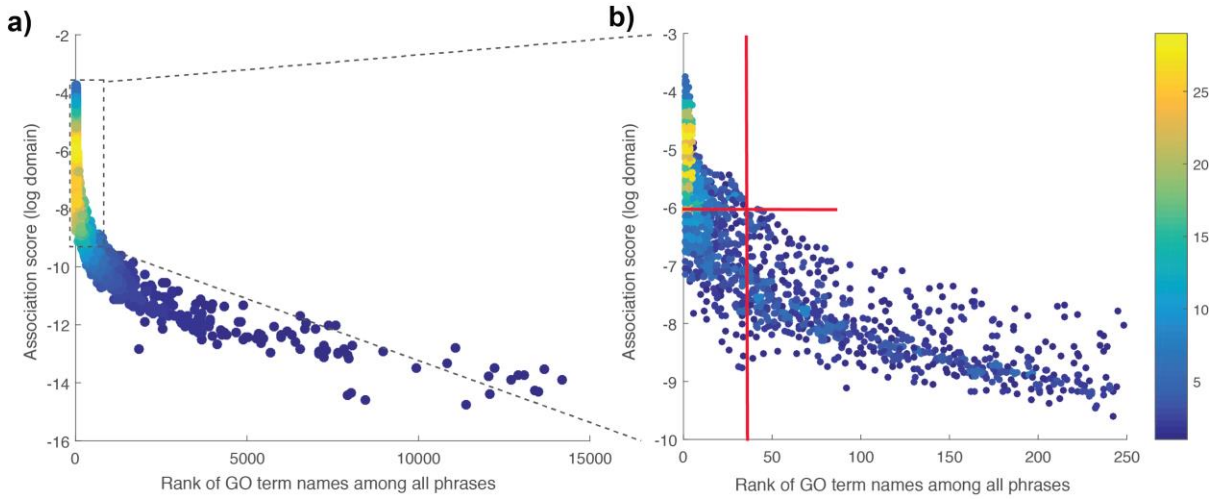
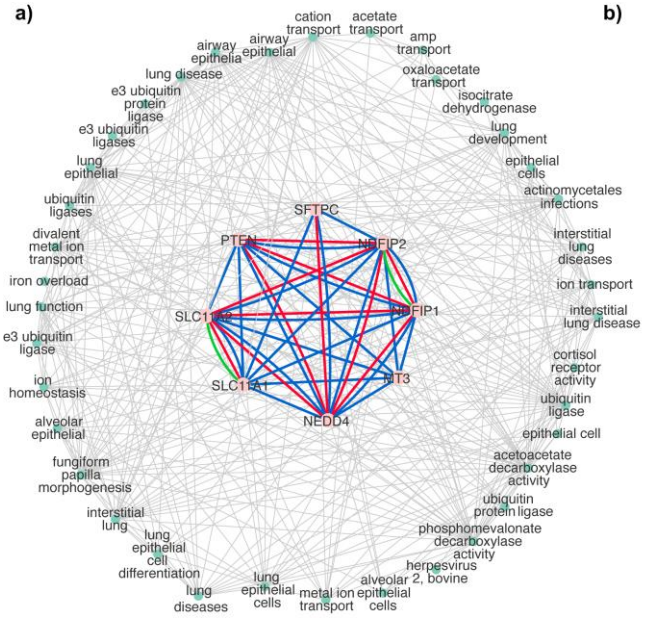


Figure 3. Selecting relevant phrases based on the association score. (a) For each GO term, the association score of its curated name is plotted with the rank of this score among all candidate phrases. (b) Zoom-in of panel (a) reveals that applying a threshold of $\geq$ -6 on the association score guarantees that the curated name of a term is ranked among the top 40 candidate phrases.



b)

Cancer type	MutsigCV Enrichment Pvalue
BLCA	0.00122
CESC	0.00030
COAD	0.00030
KICH	0.00010
LGG	0.00969
LUSC	0.00345
PRAD	0.00447
SKCM	0.00071
UCEC	0.00535

Figure 4. Discovery and characterization of a new pathway by our method. (a) The pathway is defined by eight genes related by protein interactions, co-expression and protein shared domains. These functions of these genes are collectively described by 38 phrases. (b) Cancer types in which these genes are significant mutated in The Cancer Genome Atlas (TCGA).

3.2 Functional annotations for unknown cancer pathways

Encouraged by the ability of our method to recover the curated names of known pathways, we set out to assign names to new gene sets inferred from molecular data. We analyzed a total of 2,132 gene sets detected by the hierarchical clustering algorithm CLiXO[18] based on a human gene similarity network with 19,035 genes and 181,156,095 edges (Section 3.1). Of these, we only considered those that did not significantly overlap with known pathways. In this section, we chose two example gene sets which were suggested to be highly related to cancer by our approach to demonstrate how our method can help to discover new biological knowledge.

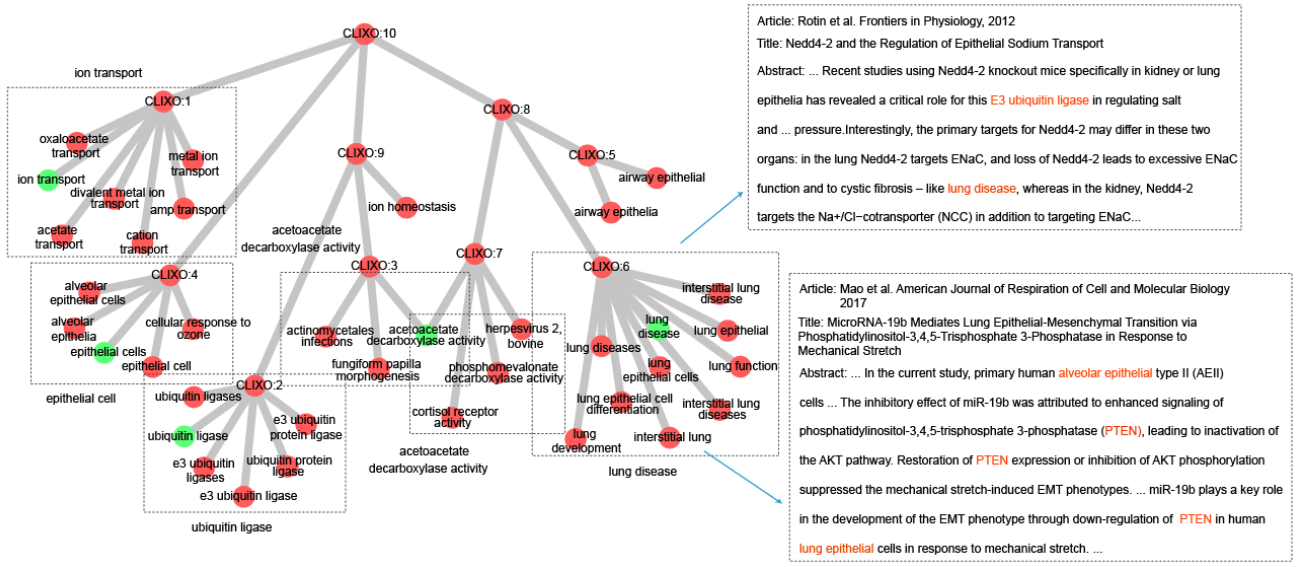


Figure 5. Summarization of biological function by a Concept Ontology. The 38 phrases describing the pathway in **Fig. 4** were hierarchically clustered based on their semantic relations using the CLiXO algorithm. These phrases were organized into six major concepts. We list two of the journal articles, Mao et al.[32] and Rotin et al.[29,31], contributing to the concepts 'lung disease' and 'epithelial cell'.

As a first case study, we examined a pathway consisting of eight strongly interacting genes: *NEDD4*, *PTEN*, *SLC11A2*, *SLC11A1*, *SFTPC*, *MT3*, *NDFIP1* and *NDFIP2* (**Fig. 4a**). To our knowledge, this pathway was previously unknown, as it has poor overlap with all catalogued pathways in GO and KEGG (Jaccard Index ≤ 0.25). Our method identified 38 literature phrases associated with this set of genes (**Fig. 4a**). Although each of these phrases might represent a distinct biological function, we found that some were highly related to one another, forming a hairball-like subnetwork of gene-phrase linkages (**Fig. 4a**). Thus, it would be very challenging for a human to summarize the overall functions of this pathway. To address this challenge, we applied CLiXO to hierarchically organize these phrases into a Concept Ontology. Visualization of this ontology revealed six major functions at multiple

scales (**Fig. 5**). On a molecular level, this pathway has functions related to 'ion transport', 'acetoacetate decarboxylase activity' and 'ubiquitin ligase'. On a cellular and organismal level, it is involved in 'epithelial cells' and 'lung disease'. These descriptions were supported by direct associations between phrases and genes in multiple articles, such as 'lung disease' and *NEDD4* in Rotin et al.[29,31] and 'lung epithelial cell' and *PTEN* in Mao et al.[32], and by indirect associations learned through the random walk transformation. As validation of these descriptions, we found that genes in this pathway were recurrently mutated in lung squamous cell carcinoma (LUSC), a disease in epithelial cells[33], based on MutSigCV scores[34] in The Cancer Genome Atlas (TCGA)[35]. All evidence suggested that this is a novel functional pathway related to lung cancer.

As a second case study, we examined another pathway, consisting of eight genes *FBXW7*, *ARL2*, *FBXW11*, *FBXW2*, *BTRC*, *PWP2*, *COPA* and *FBXW10*. These genes strongly interacted with each other primarily through domain co-occurrence, suggesting their proteins share similar 3D structures (**Fig. 6a**). This pathway was also previously unknown (Jaccard Index ≤ 0.1 in GO and KEGG). Our method described its functions with 37 phrases, which could be hierarchically organized into six major concepts (**Fig. 7**). An interesting concept was 'acute monoblastic leukemia', suggesting this pathway was cancer-associated. As shown in **Fig. 7**, validation for this pathway was achieved by tracing back the actual literature referencing these genes and diseases simultaneously. One of the articles, Gelbard et al.[36], related *FBXW7* to sinonasal carcinoma, a kind of head and neck cancer. This is consistent with our finding that these genes were recurrently mutated in the HNSC and UCEC patient cohorts in TCGA (**Fig. 6b**). These two examples demonstrate how pathways can be automatically discovered and annotated by integrating years of biomedical knowledge with 'omics datasets.

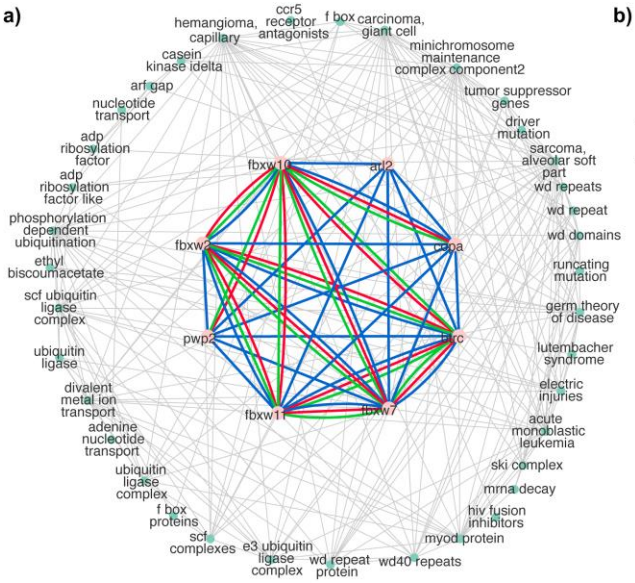


Figure 6. Discovery and characterization of another new pathway by our method. (a) The pathway is defined by eight genes related by protein interactions, co-expression, and protein shared domains. The functions of these genes are collectively described by 37 phrases set out around the periphery. (b) Cancer types in which these genes are recurrently mutated in TCGA.

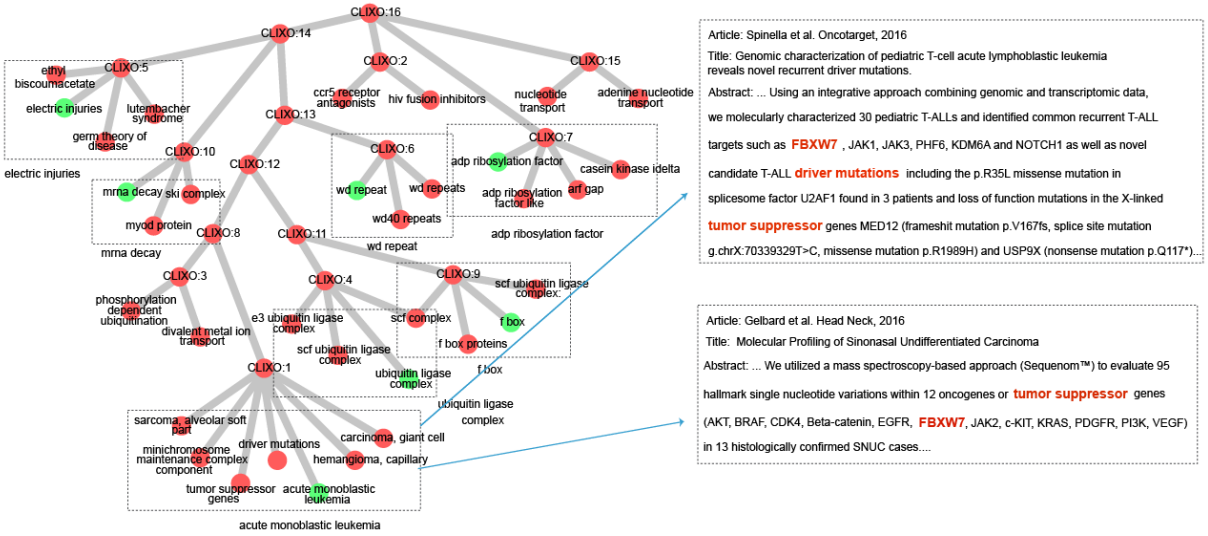


Figure 7. Summarization of biological function by a Concept Ontology. The 37 phrases describing the pathway in **Fig. 6** were hierarchically clustered based on their semantic relations, using the CLiXO algorithm. These phrases were organized into six major concepts. We list two of the journal articles, Spinella et al.[37] and Gelbard et al.[36], from which the concept 'acute monoblastic leukemia' was inferred.

4. Conclusion

In this work, we have developed a novel text mining and visualization tool for automated pathway functional annotation. Our main idea is to integrate literature and molecular interaction information into a large heterogenous network and then use a random walk-based approach to rank candidate pathway descriptions. In the final step, we use a Concept Ontology to visualize annotations as a more informative alternative to a flat network of biomedical phrases. In this work our primary focus is to annotate gene set, however, our framework can be well generalized to other applications. For instance, if the user provides a set of drugs, targets and their corresponding interaction networks, our method should be able to return the potential downstream and upstream pathways where these drugs might influence. Another application is that we can replace gene set with a group of disease symptoms and replace molecular network with symptom similarity network. Then our method might help to define the potential pathways and genes that lead to such symptoms.

One of the major limitations of our work is currently we can not accept users' input to specify a particular context. For example, the user might want to know the roles of these genes in brain or the user only want to know the location information. Theoretically speaking, these information is all included in our result, however, they might not rank high enough to pass our filter. There are many interesting directions to explore in the future. To name a few, we plan to automatically generate sentences instead of phrases for new pathways. Sentences are more widely accepted and carry more

information than phrases. Another direction is to improve our algorithm to move beyond the abstract and title to scanning complete articles and even figures. A more challenging direction is to link functional descriptions more deeply with molecular data. In our current method, the types of interactions among genes do not influence the final functional annotations. However, in practice, a rich protein-protein interactions and genetic interactions usually suggest a protein complex.

Supplementary Data: http://swang141.web.engr.illinois.edu/PSB/NetAnt_PSB2018_suppl.pdf

Reference

1. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. *Bioinformatics* **18 Suppl 1,** S233–40 (2002).
2. Ideker, T. & Krogan, N. J. *Mol. Syst. Biol.* **8,** 565 (2012).
3. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. *Nat. Genet.* **44,** 841–847 (2012).
4. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
5. Nepusz, T., Yu, H. & Paccanaro, A. *Nat. Methods* **9,** 471–472 (2012).
6. Leiserson, M. D. M. *et al. Nat. Genet.* **47,** 106–114 (2015).
7. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. *Nat. Methods* **10,** 1108–1115 (2013).
8. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. *Genome Biol.* **9 Suppl 1,** S4 (2008).
9. Ashburner, M. *et al. Nat. Genet.* **25,** 25–29 (2000).
10. Kanehisa, M. *Nucleic Acids Res.* **28,** 27–30 (2000).
11. Jin, F., Huang, M., Lu, Z. & Zhu, X. in *BioNLP '09* (2009). doi:10.3115/1572364.1572377
12. Ling, X. *et al. Inf. Process. Manag.* **43,** 1777–1791 (2007).
13. Qiao, N., Huang, Y., Naveed, H., Green, C. D. & Han, J.-D. J. *PLoS One* **8,** e74074 (2013).
14. Soldatos, T. G. *et al. Nucleic Acids Res.* **38,** 26–38 (2010).
15. NCBI Resource Coordinators. *Nucleic Acids Res.* **45,** D12–D17 (2017).
16. Liu, J., Shang, J. & Han, J. (Morgan & Claypool Publishers, 2017).
17. Cho, H., Berger, B. & Peng, J. *Cell Syst* **3,** 540–548.e5 (2016).
18. Kramer, M., Dutkowski, J., Yu, M., Bafna, V. & Ideker, T. *Bioinformatics* **30,** i34–42 (2014).
19. Dutkowski, J. *et al. Nat. Biotechnol.* **31,** 38–45 (2013).
20. Shannon, P. *et al. Genome Res.* **13,** 2498–2504 (2003).

21. Edgar, R. *Nucleic Acids Res.* **30,** 207–210 (2002).
22. GTEx Consortium. *Science* **348,** 648–660 (2015).
23. Uhlen, M. *et al. Nat. Biotechnol.* **28,** 1248–1250 (2010).
24. Barretina, J. *et al. Nature* **483,** 603–607 (2012).
25. Li, T. *et al. Nat. Methods* **14,** 61–64 (2017).
26. Stark, C. *Nucleic Acids Res.* **34,** D535–D539 (2006).
27. Lin, A., Wang, R. T., Ahn, S., Park, C. C. & Smith, D. J. *Genome Res.* **20,** 1122–1132 (2010).
28. Finn, R. D. *et al. Nucleic Acids Res.* **45,** D190–D199 (2017).
29. Finn, R. D. n *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* (2005).
30. Zhai, C. & Lafferty, J. in *SIGIR '01* (2001). doi:10.1145/383952.384019
31. Rotin, D. & Staub, O. *Front. Physiol.* **3,** 212 (2012).
32. Mao, P. *et al. Am. J. Respir. Cell Mol. Biol.* **56,** 11–19 (2017).
33. Sutherland, K. D. & Berns, A. *Mol. Oncol.* **4,** 397–403 (2010).
34. Lawrence, M. S. *et al. Nature* **499,** 214–218 (2013).
35. Cancer Genome Atlas Research Network *et al. Nat. Genet.* **45,** 1113–1120 (2013).
36. Gelbard, A. *et al. Head Neck* **36,** 15–21 (2014).
37. Spinella, J.-F. *et al. Oncotarget* **7,** 65485–65503 (2016).

The diversity and disparity in biomedical informatics (DDBI) workshop

William M. Southerland[1] and S. Joshua Swamidass[2]
[1]*Department of Biochemistry & Molecular Biology*
Howard University College of Medicine
Washington, DC
Email: wsoutherland@howard.edu

[2]*Institute for Informatics*
Washington University in St. Louis
St. Louis, Missouri
Email: swamidass@gmail.com

Philip R. O. Payne
Institute for Informatics
Washington University in St. Louis
St. Louis, Missouri
Email: prpayne@wustl.edu

Laura Wiley
Division of Biomedical Informatics and Personalized Medicine
University of Colorado
Denver, ,Colorado
Email: laura.wiley@ucdenver.edu

ClarLynda Williams-DeVane
Biomedical Biotechnology Research Institute
North Carolina Central University
Durham, North Carolina
Email: clarlynda.williams@nccu.edu

The Diversity and Disparity in Biomedical Informatics (DDBI) workshop will be focused on complementary and critical issues concerned with enhancing diversity in the informatics workforce as well as diversity in patient cohorts. According to the National Institute of Minority Health and Health Disparities (NIMHD) at the NIH, diversity refers to the inclusion of the following traditionally underrepresented groups: African Americans/Blacks, Asians (>30 countries), American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Latino or Hispanic (20 countries). Gender, culture, and socioeconomic status are also important dimensions of diversity, which may define some underrepresented groups. The under-representation of specific groups in both the biomedical informatics workforce as well as in the patient-derived data that is being used for research purposes has contributed

to an ongoing disparity; these groups have not experienced equity in contributing to or benefiting from advancements in informatics research. This workshop will highlight innovative efforts to increase the pool of minority informaticians and discuss examples of informatics research that addresses the health concerns that impact minority populations. This workshop topics will provide insight into overcoming pipeline issues in the development of minority informaticians while emphasizing the importance of minority participation in health related research. The DDBI workshop will occur in two parts. Part I will discuss specific minority health & health disparities research topics and Part II will cover discussions related to overcoming pipeline issues in the training of minority informaticians.

Keywords: Diversity; Disparity; Biomedical Informatics.

1. Rationale for Diversity in Biomedical Informatics

The Diversity and Disparity in Biomedical Informatics (DDBI) workshop will be focused on complementary and critical issues concerned with enhancing diversity in the informatics workforce as well as diversity in patient cohorts. According to the National Institute of Minority Health and Health Disparities (NIMHD) at the NIH, diversity refers to the inclusion of the following traditionally underrepresented groups: African Americans/Blacks, Asians (>30 countries), American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Latino or Hispanic (20 countries). Gender, culture, and socioeconomic status are also important dimensions of diversity, which may define some underrepresented groups. The under-representation of specific groups in both the biomedical informatics workforce as well as in the patient-derived data that is being used for research purposes has contributed to an ongoing disparity; these groups have not experienced equity in contributing to or benefiting from advancements in informatics research. With the recent and growing importance of Precision Medicine initiatives in healthcare strategies and with the increasing contribution of biomedical informatics to Precision Medicine activities, it has become imperative that all segments of society become fully engaged and represented in biomedical informatics research efforts. Further, a diverse informatics workforce is an essential component to achieving diversity in patient cohorts.

2. The Need for Diversity in the Biomedical Informatics Workforce

We face a two-fold challenge: the lack of diversity in the biomedical and health informatics workforce and lack of diversity in study cohorts which leads to disparities in contributions from diverse groups which results in experimental results that are not fully reflective of demographics. These healthcare challenges present wide-ranging biomedical informatics opportunities that will be addressed in this workshop. These challenges are quite evident in the Precision Medicine (All of Us) initiative which emphasizes that the participation of individuals from diverse social, racial/ethnic, ancestral, geographic, and economic backgrounds is critical. It has become imperative that all segments of society become fully engaged and represented in informatics efforts since a diverse informatics workforce is essential to achieving diversity in patient cohorts and in scientific contributions. For example, this is especially true in the area of genomics (more specifically genome informatics) which is emerging as a guide in the treatment of disease. This workshop will highlight innovative efforts to increase the pool

of minority informaticians and discuss examples of informatics research that addresses the health concerns that impact minority populations. Collectively, these workshop topics will provide insight into overcoming pipeline issues in the development of minority informaticians while emphasizing the importance of minority participation in health related research.

3. Strategies for Addressing Diversity and Disparity in Informatics

The DDBI workshop will occur in two parts. Part I will discuss specific minority health & health disparities research topics and Part II will cover discussions related to overcoming pipeline issues in the training of minority informaticians.

3.1. *Minority Health and Health Disparities Research*

This phase of the workshop will focus on two scientific presentations, one presentation will focus on biomedical informatics and the other on clinical informatics. The biomedical presentation: "On Using Local Ancestry to inform eQTL mapping in African Americans" will present data on the impact of local ancestry on eQTL mapping of admixed populations such as African Americans. This presentation recognizes the importance of eQTL data in understanding the function of non-coding variants in European and Asian populations and explores the significance of this data in admixed populations such as African Americans that are disproportionately impacted by a variety of diseases. The clinical informatics presentation will focus on methodological approaches to Howard University Hospital (HUH) EHR data analysis. The HUH patient population is approximately 90% minority. Detailed analysis and understanding of this data promises to shed highly granular insights into unique parameters of minority health and health disparity in the Washington, DC metropolitan area. This presentation will focus on the development of an innovative interactive tool designed to probe HUH EHR data. HUeMR (Hoard University electronic Medical Records) is a secure web-based i2b2 plugin that supports complex Boolean search operations. It has a highly interactive user interface that allows rapid data analysis for cohort discovery. Data is displayed using editable interactive charts. Users can create multiple rows of charts that contain different types of data. Users can refine queries by clicking on the charts and then select one or more additional query parameters. Additionally, users can search for diagnosis by keyword or International Classification of Disease (ICD) codes. This presentation will feature use-case examples from the Howard University Hospital (HUH) de-identified EHR data.

3.2. *Overcoming Pipeline Issues in Developing Minority Informaticians*

The lack of diversity in the field of bioinformatics is a reflection of a larger problem of exposure of underrepresented students to opportunities and role models in fields that require computational ability. In order to increase the diversity of the informatics workforce there needs to be increased exposure of more diverse students to computational thinking and practices, role models, and opportunities. This exposure needs to be deliberate and targeted to ensure that students understand the career outlook and opportunities available in the field of bioinformatics. In fact, jobs for Computer and Information Research Scientist are projected to grow 11 percent faster than the average for all

occupations over the next seven years[1]. In order to achieve this awareness researchers in the field of computer science, the natural sciences and industry partners are encouraged to form alliances that ensure that interventions are identified and implemented at critical junctures to increase the diversity of the bioinformatics workforce. Minority Serving Institutions (MSIs) should play a significant role in the advancement of this cause and need to take steps to develop their capacity to offer programs that prepare their students for this field[2].

This session will feature a panel presentation that will discuss challenges, opportunities, and success stories involving industry-academia partnerships focused on overcoming pipeline barriers in the training of minority informaticians. The Howard-Google collaboration, referred to as 'Howard West' will be a prominent discussion topic of the panel. The 'Howard West' initiative provides Howard University undergraduate students hands-on training and exposure in informatics at the Google campus in Mountain View, California with the aim of stimulating interest in computational methods as career options. The participating students are also accompanied by Howard faculty during the Google experience. The 'Howard West' initiative will be discussed from the perspectives of participating faculty, students, and Google partners. Positive outcomes, challenges and lessons learned will be discussed during this panel presentation. The panel will explore how partner stakeholders can work together to develop interdisciplinary programs to increase exposure of students at critical junctures. Additionally, the social, psychological and environmental issues related to 'belonging' that minorities face in the computing related workforce will be a topic of discussion. This panel will focus on collaborative and creative interdisciplinary pathways that serve to increase the diversity of the bioinformatics workforce. It is essential that promising interventions be shared, scaled and transferred across disciplines and institutions in educating and preparing diverse students. This panel comes at an opportune time to share challenges and solutions in the creation of approaches that address the lack of diversity in the bioinformatics workforce.

Acknowledgments

This project was supported (in part) by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number G12MD007597.

References

1. Computer and Information Research Scientists : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics: *https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm*. Accessed: 2017-08-01.
2. Mendez, R.G., Torres, J., Ishwad, P., Nicholas, H.B., Jr. and Ropelewski, A. 2016. Assisting Bioinformatics Programs at Minority Institutions: Needs Assessment, and Lessons Learned – A Look at an Internship Program. *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* (New York, NY, USA, 2016), 52:1–52:8.

INTEGRATING COMMUNITY-LEVEL DATA RESOURCES FOR PRECISION MEDICINE RESEARCH

William S. Bush, PhD, Dana C. Crawford, PhD, Farren Briggs, PhD, Darcy Freedman, PhD
Department of Population and Quantitative Health Sciences, Case Western Reserve University
Cleveland, OH, 44106, USA
Email: wsb36@case.edu; dana.crawford@case.edu; farren.briggs@case.edu; daf96@case.edu

Chantel Sloan, PhD
Department of Health Science, Brigham Young University
Provo, UT, 84602, USA
Email: chantel.sloan@byu.edu

Precision Medicine focuses on collecting and using individual-level data to improve healthcare outcomes. To date, research efforts have been motivated by molecular-scale measurements, such as incorporating genomic data into clinical use. In many cases however, environmental, social, and economic factors are much more predictive of health outcomes, yet are not systematically used in clinical practice due to the difficulties in measurement and quantification. Advances in both the availability of electronic health information, environmental exposure data, and the more systematic use of geo-coding now provide ways to systematically assess community-level indicators of health, and link these factors to electronic health records for evaluating their influence on disease outcomes. In this workshop, we discuss new electronic sources of community-level data, and provide insight into their utility and validity when compared with gold-standard data collection approaches.

1. Introduction

From the earliest efforts to identify genetic polymorphisms influencing drug response (Meyer 2004), a fundamental goal of medical practice is to tailor clinical care to the precise, individual attributes of a patient. Recent years has seen the expansion of precision medicine, with the collection of full genomic sequence data for pharmacogenomics studies (Bush et al. 2016; Rasmussen-Torvik et al. 2014). Given the dramatic and rapid expansion of knowledge in this area, it is now widely accepted that physicians cannot digest the literature fast enough to implement research findings in clinical practice (Johansen Taber and Dickinson 2014). Resources such as the Clinical Pharmacogenomics Implementation Consortium (CPIC) assist by creating clinical practice guidelines for drug prescriptions (Caudle et al. 2014). Clinical decision support systems also provide ways to implement pharmacogenomics in the clinic at point of care (Pulley et al. 2012).

While precision medicine research has focused largely on making detailed molecular measurements for each patient, efforts in the Precision Medicine Initiative will capture additional data elements including behavioral, psychosocial, and environmental factors (Collins and Varmus 2015). Part of the motivation for the expansion of scope is the increasing recognition that social, environmental, and behavioral factors are likely highly influential in disease etiology, progression, and treatment (Woolf et al. 2007). In fact, a growing body of work illustrates that considering biological and social risk factors together as a *system* may lead to better intervention strategies

(Vaughn, DeLisi, and Matto 2013). This idea of a "cells to society" model provides a framework to understand and prioritize the multitude of influential factors that comprise patient health trajectories. To date, precision medicine research has focused much more heavily on the "cells" side of this model, despite the tremendous potential of understanding and addressing the "society" side. In this workshop, we consider multiple scales on which we can derive social, environmental, and behavioral risk factors; information on the individual level, information on the community level, and information on the geographic level.

2. Information on the Individual Level

As electronic health record systems (EHRs) have increased in their adoption, information gathered from a patient encounter is now recorded digitally. Much like their paper predecessors, EHRs serve as a clearinghouse of patient-centric data gathered from clinical resources – laboratory values, vital signs, diagnoses, and procedures. Other select factors may be patient-reported, such as race/ethnicity, income, and employment, but these are rarely used to support clinical care or decision making (Community health centers leveraging the social determinants of health 2012). Despite their established importance in health, these, and other nonclinical factors, are often recorded manually and are not standardized (DeVoe et al. 2016). The National Academy of Medicine has noted the limited collection of social and behavioral factors, and has suggested new standardized data fields (based on validated instruments) for systematic adoption (Adler and Stead 2015), however these are part of a third tier of EHR meaningful use criteria, which is likely many years away.

Some information on social determinants of health is reported in ancillary EHR data through routine clinical communications (i.e. intake questions). Indicators of socio-economic status, such as measures of occupational prestige, unemployment, education, and homelessness (Hollister et al. 2016), along with country-of-origin (Farber-Eger et al. 2017) have been extracted from clinical free-text. Smoking status and alcohol use have also been extracted using natural language processing (Chen and Garcia-Webb 2014; Savova et al.). While text extraction has limitations, the use of structured elements (billing codes, etc.) may also be subject to reporting bias as well (Men 2015). Furthermore, data extracted from clinical text ultimately relies on patient self-report, which is also subject to reporting bias and differences in health perception (Campos-Castillo and Anthony 2015; Sen 2002). For many of measures like alcohol use (Bradley et al. 2011), data suggest that patients may be reluctant to divulge sensitive information associated with social stigma, also known as social desirability bias (Althubaiti 2016).

3. Information on the Community Level

Much like pharmacogenomics, the concept of using community data to inform clinical care in the US dates back to the 1960s (Adashi, Geiger, and Fine 2010), however early basic concepts, such as organizing patient charts by family and neighborhood (Froom 1977) were never widely adopted and have not transitioned to modern EHR systems. Now, due to the digitization and public availability of data, there are unprecedented opportunities to gather community-level data for precision medicine studies.

The Accelerating Data Value Across a National Community Health Center Network (ADVANCE) pilot study has begun conducting electronic assessments of the built environment,

environmental exposures, and neighborhood economic conditions to synthesize a "community vital sign" (Bazemore et al. 2016). This work is similar to other assessments, like the SocioEconomic Status Index (Roblin 2013) and the Neighborhood Deprivation Index (Messer et al. 2006). Another critical component of the built environment is access to healthy food; multiple studies have examined the impact of perceived availability of fast food and healthy food options (Barnes et al. 2016, 2017). While some information on community status can also be obtained through self-reporting, perceptual biases may be especially problematic when defining community-level characteristics – how do you self-report the status of your community relative to others? External, objective sources of community-level data may provide better estimates of their health impact.

4. Information on the Geographic Level

Geographic Information Systems (GIS) have become an extremely useful tool for analyzing publicly available geospatial information (Steiniger and Bocher 2009), such as census and environmental exposure data. These data resources have been used to derive community-level metrics integrated with EHR data, including characteristics of the walkable built environment in Pennsylvania (Nau et al. 2015), Massachusetts (D. T. Duncan et al. 2014), and Ohio (Roth et al. 2014). Similar approaches have been applied to characterize the food environment (Fiechtner et al. 2015, 2016). There are also multiple sources of atmospheric pollution in the US (B. Duncan 2014), which can provide insights into pulmonary conditions (C. D. Sloan and Johnston 2016), along with other transient, spatio-temporal factors like weather (C. Sloan et al. 2017).

Geographic data has also been combined with electronic health record information to assess the distribution of preventable emergency department visits (Fishman 2015), to identify geographic risk factors for sexually transmitted infections (Comer et al. 2011), and to assess asthma risk (Xie et al. 2017). Importantly, studies that have performed geospatial mapping of EHR data find high concordance to traditionally collected studies like the Centers for Disease Control 500 Cities Project (Birkhead 2017; CDC 2017).

5. Closing

It is clear that effective precision medicine will require a complete understanding of the patient's current health status, including risk factors from cells to society, to forecast disease development and implement treatment response. Standards for data collection are common for clinical data derived from physical examinations, and even the collection of genetic data must now adhere to Clinical Laboratory Improvement Amendments (CLIA)-certified processes, but many EHRs are missing complete and uniform documentation of environmental, social, and behavioral contexts despite their strong influence in disease processes and treatment outcomes. Despite this, the realization of a cells to society vision of precision medicine is within reach. EHRs continue to evolve, along with the infrastructure to collect, store, and integrate these community-level data into EHRs and research databases to enable the *precision* of precision medicine on multiple scales.

6. References

Adashi, Eli Y., H. Jack Geiger, and Michael D. Fine. 2010. "Health Care Reform and Primary Care — The Growing Importance of the Community Health Center." *New England Journal of Medicine* 362(22): 2047–50.

Adler, Nancy E., and William W. Stead. 2015. "Patients in Context — EHR Capture of Social and Behavioral

Determinants of Health." *New England Journal of Medicine* 372(8): 698–701.
Althubaiti, Alaa. 2016. "Information Bias in Health Research: Definition, Pitfalls, and Adjustment Methods." *Journal of multidisciplinary healthcare* 9: 211–17. https://www.dovepress.com/information-bias-in-health-research-definition-pitfalls-and-adjustment-peer-reviewed-article-JMDH (October 2, 2017).
Barnes, Timothy L et al. 2016. "Geographic Measures of Retail Food Outlets and Perceived Availability of Healthy Foods in Neighbourhoods." *Public Health Nutrition* 19(8): 1368–74.
Barnes, Timothy L. et al. 2017. "Do GIS-Derived Measures of Fast Food Retailers Convey Perceived Fast Food Opportunities? Implications for Food Environment Assessment." *Annals of Epidemiology* 27(1): 27–34.
Bazemore, Andrew W et al. 2016. "*'Community Vital Signs'* : Incorporating Geocoded Social Determinants into Electronic Records to Promote Patient and Population Health." *Journal of the American Medical Informatics Association* 23(2): 407–12.
Birkhead, Guthrie S. 2017. "Successes and Continued Challenges of Electronic Health Records for Chronic Disease Surveillance." *American journal of public health* 107(9): 1365–67.
Bradley, Katharine A et al. 2011. "Quality Concerns with Routine Alcohol Screening in VA Clinical Settings." *Journal of general internal medicine* 26(3): 299–306.
Bush, William S et al. 2016. "Genetic Variation among 82 Pharmacogenes: The PGRN-Seq Data from the eMERGE Network." *Clinical pharmacology and therapeutics*.
Campos-Castillo, Celeste, and Denise L Anthony. 2015. "The Double-Edged Sword of Electronic Health Records: Implications for Patient Disclosure." *Journal of the American Medical Informatics Association* 22(e1): e130–40.
Caudle, Kelly E et al. 2014. "Incorporation of Pharmacogenomics into Routine Clinical Practice: The Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline Development Process." *Current drug metabolism* 15(2): 209–17.
CDC. 2017. "500 Cities | About the Project." *8/28/2017*. https://www.cdc.gov/500cities/about.htm (October 2, 2017).
Chen, Es, and M Garcia-Webb. 2014. "An Analysis of Free-Text Alcohol Use Documentation in the Electronic Health Record: Early Findings and Implications." *Applied clinical informatics* 5(2): 402–15.
Collins, Francis S., and Harold Varmus. 2015. "A New Initiative on Precision Medicine." *New England Journal of Medicine* 372(9): 793–95.
Comer, Karen Frederickson et al. 2011. "Incorporating Geospatial Capacity within Clinical Data Systems to Address Social Determinants of Health." *Public health reports (Washington, D.C. : 1974)* 126 Suppl 3(Suppl 3): 54–61.
Community Health Centers Leveraging the Social Determinants of Health. 2012. http://www.altfutures.org/pubs/leveragingSDH/IAF-CHCsLeveragingSDH.pdf.
DeVoe, Jennifer E et al. 2016. "Perspectives in Primary Care: A Conceptual Framework and Path for Integrating Social Determinants of Health Into Primary Care Practice." *Annals of family medicine* 14(2): 104–8.
Duncan, Brian. 2014. "Satellite Data of Atmospheric Pollution for U.S. Air Quality Applications: Examples of Applications, Summary of Data End-User Resources, Answers to FAQs, and Common Mistakes to Avoid." *Atmospheric Environment* 94: 647–62.
Duncan, Dustin T et al. 2014. "Characteristics of Walkable Built Environments and BMI Z-Scores in Children: Evidence from a Large Electronic Health Record Database." *Environmental health perspectives* 122(12): 1359–65.
Farber-Eger, Eric et al. 2017. "Extracting Country-of-Origin from Electronic Health Records for Gene- Environment Studies as Part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study." *AMIA Joint Summits on Translational Science* 2017: 50–57.
Fiechtner, Lauren et al. 2015. "Food Environments and Childhood Weight Status: Effects of Neighborhood Median Income." *Childhood Obesity* 11(3): 260–68.
———. 2016. "Effects of Proximity to Supermarkets on a Randomized Trial Studying Interventions for Obesity." *American journal of public health* 106(3): 557–62.
Fishman, Jamie Andrew. 2015. "Geospatial Analysis of Preventable Emergency Department Visits in Chicago, IL." https://www.ideals.illinois.edu/handle/2142/78400 (October 2, 2017).
Froom, J. 1977. "An Integrated Medical Record and Data System for Primary Care. Part 1: The Age-Sex Register: Definition of the Patient Population." *The Journal of family practice* 4(5): 951–53.
Hollister, Brittany M et al. 2016. "DEVELOPMENT AND PERFORMANCE OF TEXT-MINING ALGORITHMS TO EXTRACT SOCIOECONOMIC STATUS FROM DE-IDENTIFIED ELECTRONIC HEALTH

RECORDS." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 22: 230–41.
Johansen Taber, Katherine, and Barry Dickinson. 2014. "Pharmacogenomic Knowledge Gaps and Educational Resource Needs among Physicians in Selected Specialties." *Pharmacogenomics and Personalized Medicine* 7: 145.
Men, Jessica. 2015. "Obesity and Smoking, Alcohol Use Underreported in National Database." *AJMC*. http://www.ajmc.com/newsroom/obesity-and-smoking-alcohol-use-underreported-in-national-database-
Messer, Lynne C. et al. 2006. "The Development of a Standardized Neighborhood Deprivation Index." *Journal of Urban Health* 83(6): 1041–62.
Meyer, Urs A. 2004. "Pharmacogenetics - Five Decades of Therapeutic Lessons from Genetic Diversity." *Nature reviews. Genetics* 5(9): 669–76.
Nau, Claudia et al. 2015. "Community Socioeconomic Deprivation and Obesity Trajectories in Children Using Electronic Health Records." *Obesity* 23(1): 207–12.
Pulley, J M et al. 2012. "Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project." *Clinical pharmacology and therapeutics* 92(1): 87–95.
Rasmussen-Torvik, L J et al. 2014. "Design and Anticipated Outcomes of the eMERGE-PGx Project: A Multicenter Pilot for Preemptive Pharmacogenomics in Electronic Health Record Systems." *Clinical pharmacology and therapeutics* 96(4): 482–89.
Roblin, Douglas W. 2013. "Validation of a Neighborhood SES Index in a Managed Care Organization." *Medical Care* 51(1): e1–8.
Roth, Caryn, Randi E Foraker, Philip R O Payne, and Peter J Embi. 2014. "Community-Level Determinants of Obesity: Harnessing the Power of Electronic Health Records for Retrospective Data Analysis." *BMC medical informatics and decision making* 14(1): 36.
Savova, Guergana K et al. "Mayo Clinic NLP System for Patient Smoking Status Identification." *Journal of the American Medical Informatics Association : JAMIA* 15(1): 25–28.
Sen, Amartya. 2002. "Health: Perception versus Observation." *BMJ (Clinical research ed.)* 324(7342): 860–61.
Sloan, Chantel et al. 2017. "The Impact of Temperature and Relative Humidity on Spatiotemporal Patterns of Infant Bronchiolitis Epidemics in the Contiguous United States." *Health & Place* 45: 46–54.
Sloan, Chantel D., and James D. Johnston. 2016. "Can Extreme Air Pollution Events Provide a Window into Incident Asthma?" *American Journal of Respiratory and Critical Care Medicine* 194(12): 1440–41.
Steiniger, Stefan, and Erwan Bocher. 2009. "An Overview on Current Free and Open Source Desktop GIS Developments." *International Journal of Geographical Information Science* 23(10): 1345–70.
Vaughn, Michael G., Matt. DeLisi, and Holly C. Matto. 2013. *Human Behavior : A Cell to Society Approach*.
Woolf, Steven H, Robert E Johnson, Robert L Phillips, and Maike Philipsen. 2007. "Giving Everyone the Health of the Educated: An Examination of Whether Social Change Would Save More Lives than Medical Advances." *American journal of public health* 97(4): 679–83..
Xie, Sherrie, Rebecca Greenblatt, Michael Z Levy, and Blanca E Himes. 2017. "Enhancing Electronic Health Record Data with Geospatial Information." *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2017: 123–32.

Machine learning and deep analytics for biocomputing: call for better explainability

Dragutin Petkovic[*] [†]
Computer Science Department, San Francisco State University (SFSU)
1600 Holloway Ave, San Francisco, CA 94132
Petkovic@sfsu.edu

Lester Kobzik
Department of Environmental Health, Harvard University
665 Huntington Avenue, Boston, Massachusetts 02115
lkobzik@hsph.harvard.edu

Christopher Re
Department of Computer Science, Stanford University
353 Serra Mall, Stanford, CA 94305
chrismre@cs.stanford.edu

The goals of this workshop are to discuss challenges in *explainability* of current Machine Leaning and Deep Analytics (MLDA) used in biocomputing and to start the discussion on ways to improve it. We define explainability in MLDA as *easy to use* information explaining *why and how the MLDA approach made its decisions*. We believe that much greater effort is needed to address the issue of MLDA explainability because of: 1) the ever increasing use and dependence on MLDA in biocomputing including the need for increased adoption by non-MLD experts; 2) the diversity, complexity and scale of biocomputing data and MLDA algorithms; 3) the emerging importance of MLDA-based decisions in patient care, in daily research, as well as in the development of new costly medical procedures and drugs. This workshop aims to: a) analyze and challenge the current level of explainability of MLDA methods and practices in biocomputing; b) explore benefits of improvements in this area; and c) provide useful and practical guidance to the biocomputing community on how to address these challenges and how to develop improvements. The workshop format is designed to encourage a lively discussion with panelists to first motivate and understand the problem and then to define next steps and solutions needed to improve MLDA explainability.

Keywords: Machine Learning, explainability, interpretability, workshop

1. Introduction, Background and Motivation

The goals of this workshop are to discuss challenges in *explainability* of current Machine Leaning and Deep Analytics (MLDA) used in biocomputing and to explore ways to improve it.

*Travel partially supported by Mobilize Center, Stanford University

We define explainability in MLDA as *easy to use* information explaining *why and how the MLDA approach made its decision.* Successful explainability will offer much deeper insights into MLDA operation compared to what is available today. Algorithms and software implementing MLDA decision models are inherently complex and notoriously difficult to understand and communicate. This creates barriers to their adoption by non-experts and challenges in their validation, reproducibility and benchmarking in the research community. The input data ("training databases") have a critical influence on MLDA results but are complex, and they change with time as more and more or better measurements and samples are added. Frequently, a "gold standard" or ground truth is not easily available. All this makes it very challenging to understand, evaluate, verify and even reproduce results of published MLDA work. At the same time, a review of the literature shows that very few research efforts and methods focus specifically on MLDA explainability.

Interest in explaining how ML systems work is growing within not only funding agencies and potential adopters but also in general public reflecting the penetration of ML in all aspects of our lives. Specifically, we believe that improved explainability of MLDA in biocomputing will result in the following benefits: increased credibility and confidence in its application; improved ability to objectively evaluate, audit and verify MLDA solutions; and possible discovery of new knowledge and ideas enabled by better understanding of how MLDA works on specific problems.

2. Workshop Format and Organization

Six workshop panelists will represent all four constituencies in the biocomputing ecosystem: 1) *computational researchers who are experts in MLDA* and who develop and use the technology; 2) *biocomputing practitioners* who are using MLDA but are not experts; 3) *editors/ evaluators* who need to decide what to publish; 4) and members of *funding agencies* who evaluate research results and use the funding to influence the direction of research. The 3-hour workshop is organized in the form of two main panels, followed by a discussion. The panelists are:

- A. Esteva *(Ph. D. Candidate, Stanford University)*
- Dr. R. Ghanadan (*Google since September 2017; formerly Program Manager, Defense Sciences Office, DARPA*)
- Dr. W. Kibbe (*Chief for Translational Biomedical Informatics in the Department of Biostatistics and Bioinformatics and chief data officer for the Duke Cancer Institute, Professor, Duke University since August 2017; formerly Dir. of NCI Center for Biomedical Informatics and Inf. Technology since*)
- Dr. B. Percha *(Assistant Professor, Icahn School of Medicine at Mount Sinai; Head of R&D, Health Data and Design Innovation Center (HD2i) Institute for Next-Generation Healthcare)*
- Dr. R. Roettger *(Assistant Professor, University of Southern Denmark, Odense)*
- Dr. R. Scheuermann (*Dir. Of Bioinformatics, J. Craig Venter Institute*),

Panel 1: Needs for Explainability in ML and Deep Analytics - View of "Users "
Goal of this section is that panelists who are users but not necessarily experts or developers of MLDA: a) outline their experience, needs and motivation for better explainability in MLDA; and b) encourage and *challenge* developers of MLDA technology to provide better explainability.
Panel moderator: Dr. Lester Kobzik
Panelists: Dr. R. Ganadhan , Dr. B. Percha, Dr. W. Kibbe

Panel 2: Toward Better Explainability in ML and Deep Analytics – View of "Developers"
The goal of this section is for panelists on the development and research side of MLDA techniques to: a) present examples of the state-of-the-art in MLDA explainability; and b) discuss ways to address the challenges outlined by the previous panelists.
Panel moderator: Dr. Christopher Re
Panelists: A. Esteva, Dr. R. Roetger, Dr. R. Scheuermann

Discussion with panelists and audience
Moderator: Dr. D. Petkovic

3. Panelists' Abstracts

In this section we list panelist' abstracts reelecting some of their initial thoughts and ideas to be discussed at the workshop.

AI in healthcare: a case study in explainability

Andre Esteva, Stanford University

In a recent paper we demonstrate classification of skin lesions using a single deep convolutional neural networks (CNN), trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets — consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: malignant carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. An algorithm known as t-SNE is effective at visualization high-dimensional data - we employ it to understand how the algorithm clusters images into disease categories based on visual and clinical similarity. Additionally, we render saliency maps of several example images in order to demonstrate the individual pixels that most influence a trained model's prediction - this is done by backpropagating the gradient to the input layer. Finally, we calculate confusion matrices for the CNN and two board-certified dermatologists on our validation set categories, demonstrating that the CNN misclassifies lesions in a manner similar to experts.

Machine Learning to Machine Understanding, the Need for Explainable AI

Dr. Reza Ghanadan, Google

Machine learning has shown dramatic success across many Artificial Intelligence application areas in recent years, leveraging advances in computing power and the availability of large sets of training data. As an engine for the 4th industrial revolution, AI provides a tremendous opportunity to deploy autonomous systems in many complex and interactive tasks, such as personalized medicine and healthcare, to analyze, learn, decide, and act in complex situations. However, it is essential for the users to be able to understand and trust the decisions of emerging generation of artificial intelligence systems. Productization and wide acceptance of current systems are limited because of our inability to validate and verify their performance when they act in new situations, due in turn to machine's current inability to explain its decisions and actions to human users. To gain our trust, machine-learning systems will need to have the ability to explain their rationale in meaningful ways, characterize trade-offs, and convey an understanding of how they will behave in the future in new situations. Such intrinsic capability would help users and developer community with increasingly more powerful tools and applications. In this talk, we will describe a user's perspective and needs for such capability in emerging ML/AI systems, and highlight a few examples and tools for healthcare and biocomputing at Google.

Machine Learning, Deep Analytics, and support for a Learning Health System

Dr. Warren A. Kibbe, Ph.D, Duke University

There has been a lot of progress in applying MLDA techniques, especially in the field of imaging. Image analysis using MLDA approaches are hitting mainstream computing, as evidenced by the ability of the Photos app in the Mac OS High Sierra to classify pictures containing a 'beach', or 'trees' or 'flowers' and of course face recognition. In general, MLDA techniques have two phases – a feature identification phase, where features are extracted/identified, and an associative learning component, where various statistical techniques are used to identify features that correlate with attributes available in the training set. Using techniques that are 'explainable' for associative learning is highly desired in healthcare, especially when applying these techniques to complex biological data such as whole exome sequencing and RNAseq. Coupling MLDA with Natural Language Processing for classification and decision support processes will increase the value of data in healthcare. For these techniques to reach their potential, explainability is a key need.

Explainability Tales from the Health Entrepreneur Partners Program

Dr. Bethany Percha, Icahn School of Medicine at Mount Sinai; HD2i

At HD2i, we believe many of the technologies that will fuel next-generation healthcare will come from outside the traditional healthcare ecosystem. Often with the right application of machine learning and data science, products geared toward consumers or other industries, such as fitness, can be reoriented to address powerful clinical goals. We engage in data science partnerships with early-stage companies to help bring these fresh ideas and products into the clinic faster. One result of this is that we have had to think hard about how best to explain technical concepts from machine learning and statistics to folks with little previous exposure. In my talk, I'll share some stories from our first partnerships and discuss how what we've learned

through our interactions with industry can help inform broader concepts of explainability in machine learning.

On the Explainability of Clustering Results

Dr. Richard Roettger, University of Southern Denmark, Odense

In the recent years we have seen a tremendous growth in the amount, the complexity, and the diversity of biological data. Often, the first line of defense when facing this amount of data is clustering. But how reliable are clustering results when it is unclear whether the inherit model of the clustering tool fits the data? With ClustEval, we have automatized most steps of a cluster analysis which allowed us to provide a better overview of the existing clustering tools and their respective performances and we could demonstrate how sensitive and erratic some algorithms behave under certain conditions. Furthermore, we present the tool TiCoNE (time course network enrichment) which interactively involves the user in the machine learning process to create a time series clustering. In order to gain explainability and assess the validity of the clustering, the results are enriched with biological networks in order to extract connected biological components behaving consistent over time for a certain condition. This talk should serve as one example of gaining information and explanatory power by means of the integration of independent evidence instead of creating ever more complicated computational models. All resources are available at https://clusteval.compbio.sdu.dk and https://ticone.compbio.sdu.dk

Use of machine learning-derived gene expression features to explain the unique characteristics of cell types defined using single cell RNA sequencing

Dr. Richard H. Scheuermann, J. Craig Venter Institute

Machine learning has become an important instrument in the bioinformatics tool kit, with many different applications associated with various omics technologies. In our single cell transcriptomics program, we use machine learning for sample classification to identify poor quality samples and to help partition cell types using gene expression data from complex cell mixtures. In addition to these classification objectives, we are also finding that methods, like Random Forest, that provide quantitative information about features that are most useful for classification are equally useful for explaining important relationships between features and classes. In the case of sample quality classification, features that are useful in identifying poor quality samples also point to potential problematic steps in the experimental workflow that can then be targeted for process improvement. In the case of cell type classification, gene expression features that are useful for cell type partitioning are proving informative for identifying the necessary and sufficient characteristics for defining discrete cell types, and for illuminating the important biological distinctions that characterize unique cellular phenotypes. These examples highlight the value of capturing explainability information during the machine learning process. This work is supported by the Allen Institute for Brain Science, the JCVI Innovation Fund, and the U.S. National Institutes of Health 1R21AI122100.

Methods for examining data quality in healthcare integrated data repositories

Vojtech Huser †
National Library of Medicine, National Institutes of Health
8600 Rockville Pk, Bld 38a
Bethesda, MD, 20852, USA
Email: vojtech.huser@nih.gov

Michael G. Kahn
Department of Pediatrics, University of Colorado
13001 East 17th Place MS-F563
Aurora, CO 80045 USA
Email: Michael.Kahn@ucdenver.edu

Jeffrey S. Brown
Department of Population Medicine,
Harvard Medical School and Harvard Pilgrim Health Care Institute
401 Park Drive, Suite 401 East
Boston, MA 02215USA
Email: jeff_brown@hphc.org

Ramkiran Gouripeddi
University of Utah, School of Medicine
Salt Lake City, 84102, Utah, USA
Email: ram.gouripeddi@utah.edu

This paper summarizes content of the workshop focused on data quality. The first speaker (VH) described data quality infrastructure and data quality evaluation methods currently in place within the Observational Data Science and Informatics (OHDSI) consortium. The speaker described in detail a data quality tool called Achilles Heel and latest development for extending this tool. Interim results of an ongoing Data Quality study within the OHDSI consortium were also presented. The second speaker (MK) described lessons learned and new data quality checks developed by the PEDsNet pediatric research network. The last two speakers (JB, RG) described tools developed by the Sentinel Initiative and University of Utah's service oriented framework. The workshop discussed at the end and throughout how data quality assessment can be advanced by combining best features of each network.

Keywords: Data Quality, Evaluation Methods, Visualization, Observational Research.

† VH work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC)

1. Introduction

Large Integrated Data Repositories (IDRs) have become indispensable for clinical research. Recent emergence of Common Data Models (CDMs) facilitated creation of tools that provide syntactic integration (shared information model) and in some cases also semantic integration (shared set of target terminologies used by structured data). Retrospective data analyses are increasingly being executed on multiple datasets, and distributed research networks are creating reusable tools that streamline data wrangling, data repository maintenance, and data analytics. Examples of large, well-coordinated IDRs developed using a CDM and distributed network approach include the Health Care Systems Research Network (multi-purpose research network of 18 sites), the FDA Sentinel Initiative (17 sites representing billions of medical encounters to support medical product safety surveillance), and PCORnet (over 70 sites with millions of encounters to support clinical research). Each of these distributed networks has a unique approach to addressing data quality, including some shared approaches, and each has developed tools to facilitate data quality querying.

In 2016, a group of data quality researchers called Data Quality Collaborative (DQC)[1] published a milestone article that introduced a harmonized terminology and framework for data quality assessment (DQA).[2] Another recent study, published in 2017, described experience with regular assessment of data quality within a large pediatric data research network.[3,4] A similar summary exist for the Sentinel network.[5]

This paper provides a summary of the current state of the art and future trends presented at a conference workshop focused on examining current and novel methods in assessing data quality.

2. Data Quality within the context of the OHDSI consortium (presented by V. Huser)

2.1. *Achilles Heel Data Quality Tool*

Formulation and refinement of the Observational Medical Outcomes Partnership (OMOP) data model since 2009 provided a data standardization that opened the possibility of creating data quality assessment (DQA) tools that could work unmodified across multiple datasets or multiple healthcare institutions. The OHDSI consortium created in since 2014 a tool, called Achilles Heel that included several data quality checks focused on OMOP data model conformance and some DQA checks. It was sub-component of a tool, called Achilles that also provided data characterization functionality. This tool provides general level data quality assessment as well as model conformance checking and is being actively extended with new functionality.

In September 2017, OMOP CDM workgroup agreed on extending the model with a METADATA table that would allow capturing unstructured (as free text) and structured description of data. The new METADATA table, being part of core data model, opens new feature possibilities for data quality tools (for example recognizing clearly general population datasets from clinical trials datasets and running appropriate data quality checks depending on the dataset type).

2.2. *OHDSI network study evaluating Data Quality*

To advance the analysis of data quality of sites within the OHDSI network, in 2016, OHDSI community initiated a new study focused on comparing data quality measures within the network.[6] This study builds on previous study comparing Achilles Heel outputs at several OHDSI sites.[7] The study introduced a smaller subset of dataset measures (compared to full set of measures generated by the Achilles tool) that contain less detailed data about site and thus possibly encouraging site participation in measure comparisons. The study also deals with quantifying the amount of data that has not been fully mapped to standard concepts (target terminologies for a given data domain). See Figure 1. Such data is typically present in the dataset but mapped to concept with a concept_id of zero. The Procedure and Observation domains were found to have most unmapped data across 10 OMOP datasets compared in the study to date.

2.2.1. *Empirical Thresholds*

Reaching consensus around what constitutes a high quality dataset (in general context, not for a given study) is difficult. Given lack of consensus, the approach chosen by the DataQuality study was to use empirical threshold, where datasets that are below 10th percentile receive a data quality notification (the lowest level of data quality error). For example, for a measure of 'percentage of outpatient visits', within the network, the measure varies from 33% to 92%. Using this approach, datesets below 43% of outpatient visits (below 10th percentile) are marked as likely not containing visit details about all spectrum of care (possibly come from a delivery network that manages mostly hospitals and not a full spectrum of outpatient care).

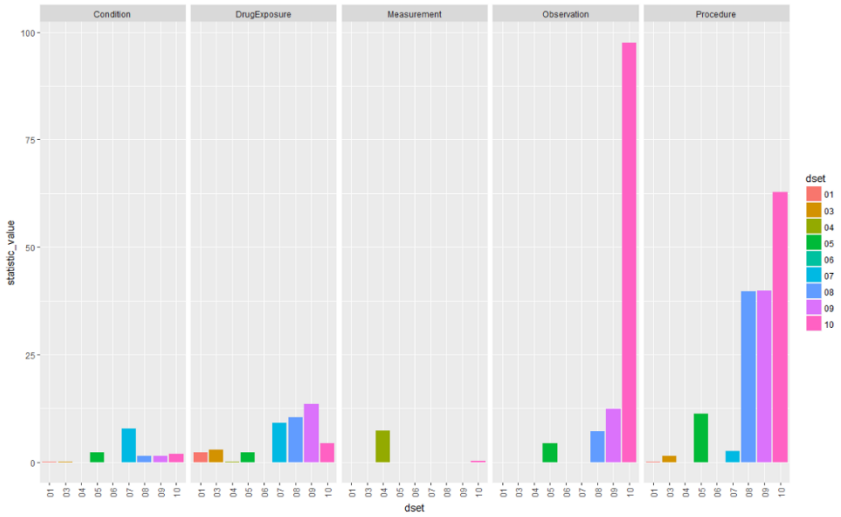


Fig. 1. Percentage of unmapped data (shown on x-axis) by data domain across 10 OMOP datasets in the Data Quality study.

3. Data Quality within the context of the PEDSNet (presented by M. Kahn)

PEDSnet is one of eleven national clinical data research networks sponsored by the Patient Centered Outcomes Research Institute that combines electronic health record (EHR) data from eight large

free-standing pediatric research hospitals.[8,9] All eight institutions implemented the OHDSI OMOP common data model for harmonizing data across different clinical environments and EHR systems. In the process of creating PEDSnet, an extensive data quality assessment program was constructed that examines data submitted by each organization and across all PEDSnet data partners. As the data quality program expanded in scope, findings related to different interpretations of how data were to be harmonized started to appear, resulting in data that were not comparable across institutions, greatly reducing the key objective for investing in a common data model. Different system configurations, workflows, and business processes often were the source of these descripancies. The data quality findings triggered the need to establish detailed data conventions with explicit rules or conventions describing how to transform site-specific data into the common data model to ensure all eight data partners are making the same transformation decisions, especially in unusual or unique situations. This modest document has grown to over 60 pages and continues to grow as new "edge cases" appear or as new data domains introduce previously unknown differences in site-level ETL proceesing. And the types of data quality issues have evolved over time as the ETL processes and network has matured.[4]

Like PEDSnet, other data networks have developed data quality processes. We examined over 11,000 data quality rules used in six large data networks of varying size and maturity.[10] We show the vast differences in data quality rules across these networks. One highly desired goal of the data quality community is to develop sharable/reusable data quality tools. We also describe recent work in developing a common data model for data quality measures that may enable networks to share data quality methods and tools.

4. Data Quality methods used by Sentinel Network (presented by J. Brown)

The US FDA Sentinel System is a medical product safety surveillance network supported by the US FDA.[11] The collaboration includes 17 data partners – 7 health insurers, 9 integrated delivery systems (i.e., insurance and care delivery) and one national hospital system - that have transformed their data into the Sentinel Common Data Model (SCDM).[12] The system includes billions of medical encounters and over 425 million person-years of data. The Sentinel Operations Center (SOC) manages the SCDM and the data quality assurance review[13,14] process. The data quality assurance team has developed and posted online a distributed program to assess network data quality across four levels of data quality checks. The first two levels of checks focus on data model compliance issues such as conformance with data model formats, completeness, validity, and cross-table and cross-variable integrity. For example, the process will check known "always" relationships such as a person with a recorded medical encounter or outpatient pharmacy dispensing must also be in the demographic table. The other checks focus on less concrete metrics focusing on trends and logical plausibility. These include metrics such as the proportion of care visits that are inpatient stays, the number of visits per person per year, the number of medication dispensing per person per year, and overall monthly trends in patients and visits. The SOC developed a set of tools to support the data quality review process; two reviewers conduct each review. The SOC conducts approximately 50 reviews per year, each review involves up to 1,500 individual data checks. The data quality review process has changed substantially since the initiation of the Sentinel pilot (Mini-Sentinel) in 2009.

The data quality approach is now more comprehensive, includes more checks, and now has explicit errors rules built into the distributed quality assurance program that will halt the program and reject a refresh is certain metrics are not met.

5. A Service Oriented Architecture for Assessing Quality of Heterogeneous Health Data (presented by R. Gouripeddi)

Translational research is dependent on secondary use of electronic health data for selection of participant cohorts and assessing real-world effectiveness of different interventions. It is therefore important to assess the quality of data used for these studies for often present data quality (DQ) issues and differentiate between natural and extraneous variations in data[15]. Assessing quality of health data needs to account for semantic and syntactic heterogeneity in health data and the diverse needs of practitioners of DQA. Several data quality conceptual frameworks (DQF) have been proposed across DQ and Information Quality domains for DQA. These DQF are diverse, varying from simple lists of concepts to complex ontological and taxonomical representations of data quality concepts (DQC) for different domains of application. There is a lack of consensus on using these DQF for a comprehensive DQA of a given dataset as well as absence of a "one-framework-fits-all" solution for DQA.[16]

In order to meet these requirements we developed a service-oriented architecture-based (SOA) DQA platform, Open Quality and Analytics Framework[17] (OQAF) consisting of three components:

1. Quality Knowledge Repository (QKR): We extracted DQC, their definitions and applicable measures, their relationships, and the computability of DQC in existing DQF (e.g. Kahn[15], Weiskopf[18] from literature. We identified primitives existing in different DQF to develop a DQ metamodel and implemented it as the QKR. The QKR is based on OpenFurther's Metadata Repository[19] (MDR) which is a standard-based, in-house developed repository that stores metadata artifacts and their relationships.
2. Federated Data Integration Platform: We use the OpenFurther platform (OF) that supports semantically consistent metadata-centric querying of heterogeneous data sources for translational research using an MDR, a terminology/ontology server and various software services (SS) for orchestrating execution of queries.[19]
3. Visualization Meta-Framework (VMF): This provides a repository for storing visualizations for different DQC and their measures. We are currently generating content for the VMF from literature and using crowd-sourced methods.

An end-user characterizes heterogeneous datasets or sources that are distributed or aggregated, by selecting one or more DQC and their measures from any DQF stored in the QKR. At the DQ query execution layer, OF's SS use messaging obtained from the QKR and the VMF along with reusing model and terminology mappings generated for performing federated data queries to perform DQA and visualizations. The content stored in the VMF informs users in the selection of appropriate visualizations for their DQA.

We present the architectural design, implementation, and evaluation of OQAF and its components - open-source, agnostic approach for standardizing DQA.

References

1. Health A. Data Quality Collaborative. 2015; http://repository.academyhealth.org/dqc/. Accessed May 15, 2015.
2. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4(1):1244.
3. CHOP. PEDSNet data quality assesment tool. 2017; https://github.com/PEDSnet/Data-Quality-Analysis. Accessed June 5, 2017.
4. Khare R, Utidjian L, Ruth BJ, et al. A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc*. 2017.
5. Brown J, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Medical care*. 2013;51(8 0 3):S22-S29.
6. Huser V. OHDSI Data Quality study. 2016; http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy. Accessed Jan 2, 2017.
7. Huser V, DeFalco F, Schuemie M, et al. Multi-site Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets. *eGEMs (Wash DC)*. 2016.
8. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578-582.
9. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc*. 2014;21(4):602-606.
10. Callahan TJ, Bauck AE, Bertoch D, et al. A Comparison Of Data Quality Assessment Checks In Six Data Sharing Networks. *Generating Evidence & Methods to improve patient outcomes (eGEMS)*. 2017;5(1).
11. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clinical pharmacology and therapeutics*. 2016;99(3):265-268.
12. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:23-31.
13. Sentinel. Data Quality Review and Characterization. 2016; https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/data-quality-review-and-characterization.
14. Raebel MA, Haynes K, Woodworth TS, et al. Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf*. 2014;23(6):609-618.
15. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50 Suppl:S21-29.
16. Kahn MG, Brown JS. Transparent Reporting of Data Quality in Distributed Data Networks. *EGEMS (Wash DC)*. 2015;3(1).
17. Rajan NS, Gouripeddi R, Facelli JC. A service oriented framework to assess the quality of electronic health data for clinical research. Paper presented at: Healthcare Informatics (ICHI), 2013 IEEE International Conference on2013.
18. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*. 2013;46(5):830-836.
19. Gouripeddi R. FURTHeR: An Infrastructure for Clinical, Translational and Comparative Effectiveness Research. . ***Proc AMIA Symp***. 2013.

ERRATUM

Identifying Mutation Specific Cancer Pathways Using a Structurally Resolved Protein Interaction Network

by H. Billur Engin, Matan Hofree & Hannah Carter

In the above PSB article published in *Biocomputing 2015: Proceedings of the Pacific Symposium,* pp. 84-95, doi: 10.1142/9789814644730_0010 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299875/), the following Acknowledgments section is missing:

Acknowledgements

The results published in this article are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/."