

## PACIFIC SYMPOSIUM ON BIOCOMPUTING 2014

The Pacific Symposium on Biocomputing (PSB) 2014 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 2014 will be held on January 3 – 7, 2014 in Kohala Coast, Hawaii. Tutorials and workshops will be offered prior to the start of the conference.

PSB 2014 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  
9078 eb

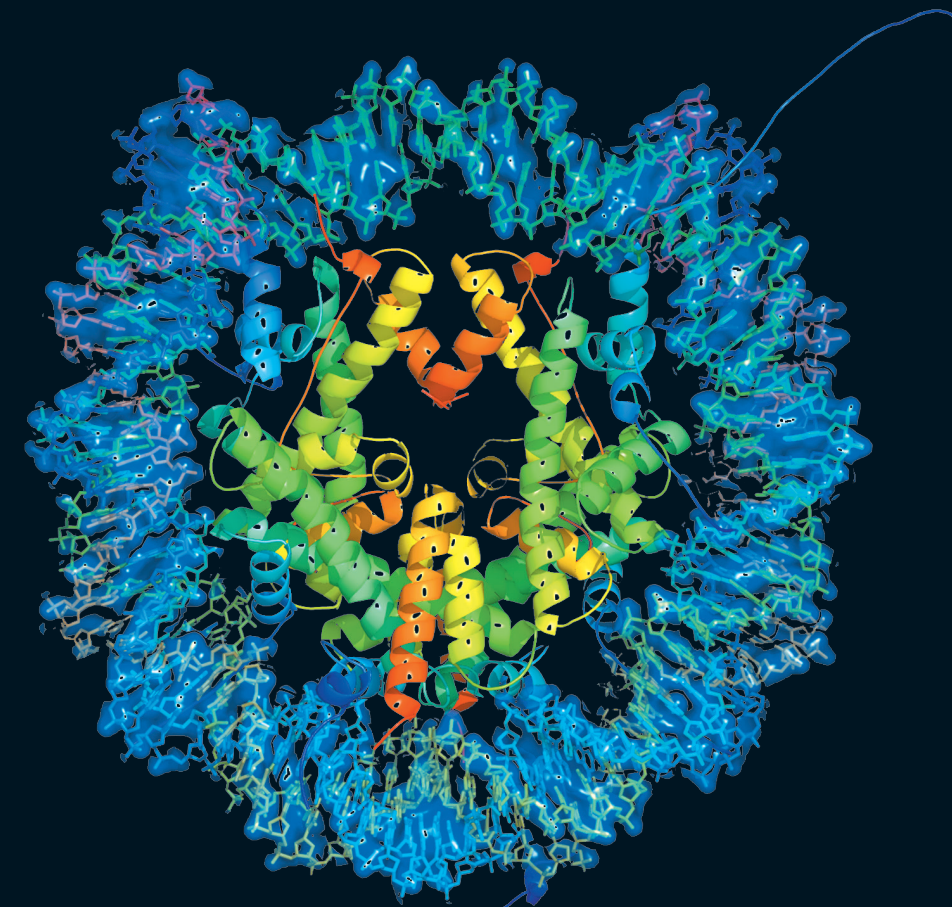


R. B. Altman  
A. K. Dunker  
L. Hunter  
M. D. Ritchie  
T. Murray  
T. E. Klein

## PACIFIC SYMPOSIUM ON BIOCOMPUTING 2014



# PACIFIC SYMPOSIUM ON BIOCOMPUTING 2014



*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.

Copyright © 2004 Pacific Symposium on  
Biocomputing.

PACIFIC SYMPOSIUM ON  

---

BIOCOMPUTING 2014



# PACIFIC SYMPOSIUM ON BIOCOMPUTING 2014

Kohala Coast, Hawaii, USA  
3–7 January 2014

*Edited by*

Russ B. Altman  
Stanford University, USA

A. Keith Dunker  
Indiana University, USA

Lawrence Hunter  
University of Colorado Health Sciences Center, USA

Marylyn D. Ritchie  
Pennsylvania State University, USA

Tiffany Murray  
Stanford University, USA

Teri E. Klein  
Stanford University, USA



NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Preface .....	v
---------------	---

## **CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT “OMICS” DATA**

<i>Session Introduction</i> .....	1
Søren Brunak, Francisco M. De La Vega, Gunnar Rätsch, Joshua M. Stuart	
<i>Tumor Haplotype Assembly Algorithms for Cancer Genomics</i> .....	3
Derek Aguiar, Wendy S.W. Wong, Sorin Istrail	
<i>Extracting Significant Sample-Specific Cancer Mutations Using Their Protein Interactions</i> .....	15
Liviu Badea	
<i>The Stream Algorithm: Computationally Efficient Ridge-Regression via Bayesian Model Averaging, and Applications to Pharmacogenomic Prediction of Cancer Cell Line Sensitivity</i> .....	27
Elias Chaibub Neto, In Sock Jang, Stephen H. Friend, Adam A. Margolin	
<i>Sharing Information to Reconstruct Patient-Specific Pathways in Heterogeneous Diseases</i> .....	39
Anthony Gitter, Alfredo Braunstein, Andrea Pagnani, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Riccardo Zecchina, Ernest Fraenkel	
<i>Detecting Statistical Interaction Between Somatic Mutational Events and Germline Variation from Next-Generation Sequence Data</i> .....	51
Hao Hu, Chad D. Huff	
<i>Systematic Assessment of Analytical Methods for Drug Sensitivity Prediction from Cancer Cell Line Data</i> .....	63
In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H. Friend, Adam A. Margolin	
<i>Integrative Analysis of Two Cell Lines Derived from a Non-Small-Lung Cancer Patient – A Panomics Approach</i> .....	75
Oleg Mayba, Florian Gnad, Michael Peyton, Fan Zhang, Kimberly Walter, Pan Du, Melanie A. Huntley, Zhaoshi Jiang, Jinfeng Liu, Peter M. Haverty, Robert C. Gentleman, Ruiqiang Li, John D. Minna, Yingrui Li, David S. Shames, Zemin Zhang	
<i>An Integrated Approach To Blood-Based Cancer Diagnosis And Biomarker Discovery</i> .....	87
Martin Renqiang Min, Salim Chowdhury, Yanjun Qi, Alex Stewart, Rachel Ostroff	
<i>Multiplex Meta-Analysis of Medulloblastoma Expression Studies with External Controls</i> .....	99
Alexander A. Morgan, Matthew D. Li, Achal S. Achrol, Purvesh J. Khatri, Samuel H. Cheshier	

## **COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY**

<i>Session Introduction</i> .....	110
S. Joshua Swamidass, Zhiyong Lu, Pankaj Agarwal, Atul Butte	
<i>Challenges in Secondary Analysis of High Throughput Screening Data</i> .....	114
Aurora S. Blucher, Shannon K. McWeeney	

<i>Drug Intervention Response Predictions with Paradigm (DIRPP) Identifies Drug Resistant Cancer Cell Lines and Pathway Mechanisms of Resistance</i> .....	125
Douglas Brubaker, Analisa Difeo, Yanwen Chen, Taylor Pearl, Kaide Zhai, Gurkan Bebek, Mark Chance, Jill Barnholtz-Sloan	

<i>Anti-Infectious Drug Repurposing Using an Integrated Chemical Genomics and Structural Systems Biology Approach</i> .....	136
Clara Ng, Ruth Hauptman, Yinliang Zhang, Philip E. Bourne, Lei Xie	

<i>Drug-Target Interaction Prediction by Integrating Chemical, Genomic, Functional and Pharmacological Data</i> .....	148
Fan Yang, Jinbo Xu, Jianyang Zeng	

<i>Prediction of Off-Target Drug Effects Through Data Fusion</i> .....	160
Emmanuel R. Yera, Ann E. Cleves, Ajay N. Jain	

<i>Exploring the Pharmacogenomics Knowledge Base (PharmGKB) for Repositioning Breast Cancer Drugs by Leveraging Web Ontology Language (OWL) and Cheminformatics Approaches</i> .....	172
Qian Zhu, Cui Tao, Feichen Shen, Christopher Chute	

## **DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES**

<i>Session Introduction</i> .....	183
Anna L. Tyler, Dana C. Crawford, Sarah A. Pendergrass	

<i>Using the Bipartite Human Phenotype Network to Reveal Pleiotropy and Epistasis Beyond the Gene</i> .....	188
Christian Darabos, Samantha H. Harmon, Jason H. Moore	

<i>Environment-Wide Association Study (EWAS) for Type 2 Diabetes in the Marshfield Personalized Medicine Research Project Biobank</i> .....	200
Molly A. Hall, Scott M. Dudek, Robert Goodloe, Dana C. Crawford, Sarah A. Pendergrass, Peggy Peissig, Murray Brilliant, Catherine A. McCarty, Marylyn D. Ritchie	

<i>Dissection of Complex Gene Expression Using the Combined Analysis of Pleiotropy and Epistasis</i> .....	212
Vivek M. Philip, Anna L. Tyler, Gregory W. Carter	

## **PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY**

<i>Session Introduction</i> .....	224
Jennifer Listgarten, Oliver Stegle, Quaid Morris, Steven E. Brenner, Leopold Parts	

<i>PATH-SCAN: A Reporting Tool for Identifying Clinically Actionable Variants</i> .....	229
Roxana Daneshjou, Zachary Zappala, Kim Kukurba, Sean M Boyle, Kelly E Ormond, Teri E Klein, Michael Snyder, Carlos D Bustamante, Russ B Altman, Stephen B Montgomery	

<i>Imputation-based Assessment of Next Generation Rare Exome Variant Arrays</i> .....	241
Alicia R. Martin, Gerard Tse, Carlos D. Bustamante, Eimear E. Kenny	
<i>Utilization of an EMR-Biorepository to Identify the Genetic Predictors of Calcineurin-Inhibitor Toxicity in Heart Transplant Recipients</i> .....	253
Matthew Oetjens, William S. Bush, Kelly A. Birdwell, Holli H. Dilks, Erica A. Bowton, Joshua C. Denny, Russell A. Wilke, Dan M. Roden, Dana C. Crawford	
<i>Robust Reverse Engineering of Dynamic Gene Networks Under Sample Size Heterogeneity</i> .....	265
Ankur P. Parikh, Wei Wu, Eric P. Xing	
<i>Variant Priorization and Analysis Incorporating Problematic Regions of the Genome</i> .....	277
Anil Patwardhan, Michael Clark, Alex Morgan, Stephen Chervitz, Mark Pratt, Gabor Bartha, Gemma Chandratillake, Sarah Garcia, Nan Leng, Richard Chen	
<i>Bags of Words Models of Epitope Sets: HIV Viral Load Regression with Counting Grids</i> .....	288
Alessandro Perina, Pietro Lovato, Nebojsa Jojic	
<i>Joint Association Discovery and Diagnosis of Alzheimer's Disease by Supervised Heterogeneous Multiview Learning</i> .....	300
Shandian Zhe, Zenglin Xu, Yuan Qi, Peng Yu	
<b>TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY</b>	
<i>Session Introduction</i> .....	312
Graciela H. González, Kevin Bretonnel Cohen, Robert Leaman, Casey Greene, Nigam Shah, Maricel G. Kann, Jieping Ye	
<i>Vector Quantization Kernels for the Classification of Protein Sequences and Structures</i> .....	316
Wyatt T. Clark, Predrag Radivojac	
<i>Combining Heterogenous Data for Prediction of Disease Related and Pharmacogenes</i> .....	328
Christopher S. Funk, Lawrence E. Hunter, K. Bretonnel Cohen	
<i>A Novel Profile Biomarker Diagnosis for Mass Spectral Proteomics</i> .....	340
Henry Han	
<i>Towards Pathway Curation Through Literature Mining – A Case Study Using PharmGKB</i> .....	352
Ravikumar K.E., Kavishwar B. Waghlikar, Hongfang Liu	
<i>Sparse Generalized Functional Linear Model for Predicting Remission Status of Depression Patients</i> ..	364
Yashu Liu, Zhi Nie, Jiayu Zhou, Michael Farnum, Vaibhav A Narayan, Gayle Wiittenberg, Jieping Ye	
<i>Development of a Data-Mining Algorithm to Identify Ages at Reproductive Milestones in Electronic Medical Records</i> .....	376
Jennifer Malinowski, Eric Farber-Eger, Dana C. Crawford	
<i>An Efficient Algorithm to Integrate Network and Attribute Data for Gene Function Prediction</i> .....	388
Shankar Vembu, Quaid Morris	



<i>Matrix Factorization-Based Data Fusion for Gene Function Prediction in Baker's Yeast and Slime Mold</i> .....	400
Marinka Zitnik, Blaz Zupan	

## WORKSHOPS

<i>Applications of Bioinformatics to Non-Coding RNAs in the Era of Next-Generation Sequencing</i> .....	412
Chao Cheng, Jason Moore, Casey Greene	
<i>Building the Next Generation of Quantitative Biologists</i> .....	417
Kristine A. Pattin, Anna C. Greene, Russ B. Altman, Lawrence E. Hunter, David A. Ross, James A. Foster, Jason H. Moore	
<i>Uncovering the Etiology of Autism Spectrum Disorders: Genomics, Bioinformatics, Environment, Data Collection and Exploration, and Future Possibilities</i> .....	422
Sarah A. Pendergrass, Santhosh Girirajan, Scott Selleck	

## PACIFIC SYMPOSIUM ON BIOCOMPUTING 2014

2014 marks the 19th Pacific Symposium on Biocomputing. The past year has been a magnificent one for computational biology. Dr. Michael Levitt won the Nobel Prize in Chemistry in October, 2013 for his work modeling and simulating biological macromolecular structure. The demand for bioinformatics and computational biology skills in the marketplace has exploded, and there is a robust industry emerging providing “Big Data” services in support of genomics, pharmaceuticals and health care. The NIH has announced a \$100M program in “Big Data to Knowledge” or BD2K, to support a broad portfolio of work across all biomedical disciplines.

In addition to being published by World Scientific and indexed in PubMed, the proceedings from all previous meetings are available online at <http://psb.stanford.edu/psb-online/>. PSB provides sessions focusing on emerging areas in biomedical computation. These sessions are typically conceived at the previous PSB meeting as people discuss the opportunities for new sessions. Once again, we have a very exciting set of areas that build on previous sessions and introduce new topics. There are revolutions occurring in cancer, drug repurposing, personalized medicine, the characterization of complex phenotypes, and our ability to extract knowledge from text. Each of these is associated with a session at this year’s meeting. In addition, we see increasing interest in non-coding RNA, quantitative training in the biosciences, and applications in autism, which form the basis for our workshops. The efforts of a dedicated group of leaders has produced an outstanding set of sessions, with associated introductory tutorials. These organizers provide the scientific core of PSB and their sessions are as follows:

### **Cancer Panomics: Computational Methods and Infrastructure for Integrative Analysis of Cancer High-Throughput “Omics” Data**

Søren Brunak, Francisco M. De La Vega, Gunnar Rätsch, Joshua M. Stuart

### **Computational Approaches to Drug Repurposing and Pharmacology**

Zhiyong Lu, Pankaj Agarwal, Atul Butte, S. Joshua Swamidass

### **Detecting and Characterizing Pleiotropy: New Methods for Uncovering the Connection Between the Complexity of Genomic Architecture and Multiple Phenotypes**

Sarah A. Pendergrass, Dana C. Crawford, Anna L. Tyler

### **Personalized Medicine: From Genotypes and Molecular Phenotypes Towards Therapy**

Oliver Stegle, Jennifer Listgarten, Steven Brenner, Leopold Parts, Quaid Morris

### **Text and Data Mining for Biomedical Discovery**

Graciela H. González, Kevin Bretonnel Cohen, Maricel G. Kann, Casey Greene, Robert Leaman, Nigam Shah, Jieping Ye

We are also pleased to present three 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:

### **Applications of Bioinformatics to Non-Coding RNAs in the Era of Next-Generation Sequencing**

Chao Cheng, Jason Moore, Casey Greene

### **Building the Next Generation of Quantitative Biologists**

Kristine A. Pattin, Anna C. Greene, James A. Foster, Jason Moore

### **Uncovering the Etiology of Autism Spectrum Disorders: Genomics, Bioinformatics, Environment, Data Collection and Exploration, and Future Possibilities**

Sarah A. Pendergrass & Scott Selleck

We thank our keynote speakers Atul Butte (Science keynote) and Douglas Fridsma (Ethical, Legal and Social Implications keynote).

Tiffany Murray manages the peer review process and assembly of the proceedings, and also plays a key role in many other aspects of the meeting. We also thank the National Institutes of Health and the International Society for Computational Biology (ISCB) for travel grant support. We are particularly grateful to the onsite PSB staff Al Conde, Brant Hansen, Georgia Hansen, BJ McKay-Morrison, 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.

We look forward to a great meeting once again.

Aloha!

Pacific Symposium on Biocomputing Co-Chairs,  
October 14, 2013

**Russ B. Altman**

*Departments of Bioengineering, Genetics & 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*

**Teri E. Klein**

*Department of Genetics, Stanford University*

**Marylyn D. Ritchie**

*Department of Biochemistry and Molecular Biology, Pennsylvania State University*

## Thanks to the reviewers...

Finally, we wish to thank the scores of reviewers. PSB aims for every paper in this volume 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.

Zaky Adam	Julien Gagneur	Yen Yi Lin	Oliver Stegle
Pankag Agarwal	Eugenia	Christoph Lippert	Josh Stuart
Uri David Akavia	Giannopoulou	Chunyu Liu	S. Joshua Swamidass
Gary Bader	Jesse Gillis	Haiguang Liu	Tasnia Tahsin
Joel Bader	Anna Goldenberg	Yi Liu	Haixu Tang
Amol Bhalla	Graciela Gonzalez	Xinghua Lou	Nick Tatonetti
Ziv Bar-Joseph	Assaf Gottlieb	Zhiyong Lu	Lisa Tucker-Kellogg
Alexis Battle	Jonathan Goya	Yves Lussier	Tamir Tuller
Gurkan Bebek	Gregory Grant	Florian Markowetz	Igor Ulitsky
Gill Bejerano	Casey Greene	Brett McKinney	Ryan Urbanowicz
Riccardo Bellazzi	Yuanfang Guan	Michael Menden	Vlado Uzunangelov
Stephen Benz	Rachael Hageman	Jill Mesirov	Alfonso Valencia
Andreas Beyer	David Haussler	Sara Mostafavi	Charles Vaske
Sourav Bhowmick	Xin He	Sayan Mukherjee	Shankar Vembu
Inanc Birol	Wenlian Hsu	Chad Myers	Karin Verspoor
Josh Bittker	Jing Hu	Dvir Netanel	Bjarni Vilhjalmsson
Sébastien Boisvert	Ting Hu	William Stafford	Tomas Vinar
Karsten Borgwardt	Jun Huan	Noble	Julia Vogt
Andrew Brown	Hailiang Huang	Chris O'Donnell	Kavishwar
Søren Brunak	Heng Huang	Hatice Osmanbeyoglu	Wagholikar
Atul Butte	Curtis Huttenhower	Princy Parsana	Peter Waltman
Colin Campbell	Janina Jeff	Leopold Parts	Fei Wang
Elias Chaibub Neto	Shuiwang Ji	Chirag Patel	Chih-Hsuan Wei
Sreenivas Chavali	Antonio Jimeno	Bethany Percha	John Witte
Keira Cheetham	Vladimir Jojic	Fernando Perez-Cruz	Chris Wong
Bin Chen	Siddharta Jonnalagada	Jason Piper	Jinbo Xu
Su-Shing Chen	Kenneth Jung	Feng Qi	Julie Yang
Jie Cheng	Irene Kaplow	Francis Quetier	Jieping Ye
John Cleary	Theofanis Karaletsos	Barbara Rakitsch	Lana Yeganova
Greg Cooper	Ritu Khare	Huzefa Rangwala	Xiaoqing You
Eivind Coward	Robert Kincaid	Ben Raphael	Habil Zare
Francisco De La Vega	Smita Krishnaswamy	Gunnar Ratsch	Jasmine Zhou
Jeroen De Ridder	Semyon Kruglyak	Ben Readhead	Jian Zhou
Emek Demir	Rui Kuang	Mireille Regnier	
Josh Denny	Natsuhiko Kumasaka	Juri Reimand	
Jen Doherty	David Kuo	Susanna Repo	
Frank Dudbridge	Christian Lang	Anna Ritz	
Sean Ekins	Robert Leaman	Luis Rocha	
Kyle Ellrott	Hayan Lee	Sushmita Roy	
Barbara Engelhardt	Kjong Lehmann	Julio Saez-Rodriguez	
Elana Fertig	Paea LePendu	S. Cenk Sahinalp	
Iddo Freidberg	Christina Leslie	Matthew Scotch	
Johannes	Haiquan Li	Jun Sese	
Freundenberg	Hua Li	Ron Shamir	
Brooke Fridely	Jason Li	Artem Sokolov	
Nicolo Fusi	Jiao Li	Joe Song	
Daniel Gaffney	Wenyuan Li	Sriganesh Srihari	



## **CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT “OMICS” DATA**

SØREN BRUNAK

*Center for Biological Sequence Analysis Department of Systems Biology,  
Technical University of Denmark, Copenhagen, Denmark  
Email: brunak@cbs.dtu.dk*

FRANCISCO M. DE LA VEGA

*Real Time Genomics, Inc., San Bruno, CA, USA  
Email: francisco@realtimegenomics.com*

GUNNAR RÄTSCH

*Computational Biology Center,  
Memorial Sloan-Kettering Cancer Center, New York City, NY, USA  
Email: raetsch@cbio.mskcc.org*

JOSHUA M. STUART

*Center for Biomolecular Science and Engineering,  
University of California Santa Cruz, CA, USA  
Email: jstuart@soe.ucsc.edu*

Precision medicine promises to transform cancer treatment in the next decade through the use of high-throughput sequencing and other technologies to identify telltale molecular aberrations that reveal therapeutic vulnerabilities of each patient's tumor [1]. This session will address the "panomics" of cancer – the complex combination of patient-specific characteristics that drive the development of each person's tumor and response to therapy [2]. The realization of this vision will require novel infrastructure and computational methods to integrate large-scale data effectively and query it in real-time for therapy and/or clinical trial selection for each patient.

The session will explore the computational needs to enable precision oncology from both the academic, industrial, and healthcare viewpoints. New methods and infrastructure to integrate multiple "omics" datasets (e.g., proteome, genome, exome, transcriptome), as well as existing clinical data types to enable precision medicine (e.g., medical literature, electronic medical records, clinical trial data, histopathology) will be discussed. The session is particularly interested in discussing pathway disruption analysis by combining data from different "omics" sources in single patients; joint analysis of "omics" data, literature, clinical trial data, and medical records; data structures & systems to enable big-data integrative analysis in patients. A summary of the accepted papers in this volume is below.

One of the most successful bioinformatics applications to cancer diagnosis and prognosis has been the identification and development of biomarkers that can distinguish disease subtypes, predict mutation status, or predict outcomes or treatment responses. However, the field is still in need of strategies that develop robust signatures as current methodologies often fail to translate

across studies and platforms (e.g., microarray- to RNA-Sequencing-based signatures). Two methods for novel biomarker discovery will be presented including an integrative approach by Min *et al.* as well as a method by Morgan *et al.* that combines multiple expression studies to identify more reliable robust gene expression-based signatures. In addition to the biomarker studies, machine-learning models for predicting the sensitivity of a cell to a drug based on its omics profile will be discussed including a comparison of methods in a comprehensive cell line panel by Jang *et al.*, the description of a new ensemble-based methods called Stream described by Chaibub Neto *et al.*, and an integrative method introduced by Mayba *et al.*

Interpreting the role of specific mutations in somatic cells is a fundamental problem in the individualized treatment of cancer. Identifying driver from passenger events and the assessment of the gain- or loss-of-function of specific proteins may offer important clues for drug targeting. Two papers investigate omics-derived statistical patterns to assess the functional role of somatic variants including connecting such events to the germline by Hu *et al.* and one that leverages protein-protein interactions to identify possibly important driving events by Badea *et al.* A new method for assembling haplotypes that is key for the interpretation of the combined influence of multiple variant alleles on the cancer phenotype is described in Aquilar *et al.*

Finally, to maximize the benefit of the cancer panomics endeavor, findings in the n=1 setting must be distributed in a way to empower the next n=1 analysis. Approaches that can interlink the findings of patients, doctors, trials, and researchers in one system would enable a new era of integrative approaches. Gitter *et al.* in this session describe one such strategy for approximating the influence of genetic pathways in disease.

Cancer panomics as applied to the individual patient is an emerging area driven by the lowering in cost of sequencing a patient's tumor and germline tissues. There is every expectation that the costs will continue their downward spiral once the competitive landscape of the industry and the maturity of 3<sup>rd</sup> or 4<sup>th</sup> generation sequencing technologies improve. In the very near future it will be feasible to sequence the complete cancer tumor genome and transcriptome as a routine procedure rather than just a targeted set of genes. The result of all this sequencing will mean that the bottleneck for the treatment of patients will transition from data production to the computational analysis of these massive information troves. Thus, it is critical to continue the discussion of novel bioinformatics ideas and strategies to empower the development of new cancer panomics approaches in the near future.

## References

1. Craig, D.W. *et al.* *Mol. Cancer Ther.*, **12**, 104–116 (2012).
2. ASCO Board of Directors. *American Society of Clinical Oncology*, 1–16 (2012).

# TUMOR HAPLOTYPE ASSEMBLY ALGORITHMS FOR CANCER GENOMICS

DEREK AGUIAR<sup>†</sup>, WENDY S.W. WONG<sup>‡,\*</sup>, SORIN ISTRAIL<sup>†,\*</sup>

<sup>†</sup> *Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA*

<sup>‡</sup> *Inova Translational Medicine Institute, Inova Health Systems, Falls Church, VA 22042, USA*

*\*Please address correspondence to Sorin\_Istrail@brown.edu and Wendy S.W. Wong  
wendy.wong@inova.org*

The growing availability of inexpensive high-throughput sequence data is enabling researchers to sequence tumor populations within a single individual at high coverage. But, cancer genome sequence evolution and mutational phenomena like driver mutations and gene fusions are difficult to investigate without first reconstructing tumor haplotype sequences. Haplotype assembly of single individual tumor populations is an exceedingly difficult task complicated by tumor haplotype heterogeneity, tumor or normal cell sequence contamination, polyploidy, and complex patterns of variation. While computational and experimental haplotype phasing of diploid genomes has seen much progress in recent years, haplotype assembly in cancer genomes remains uncharted territory.

In this work, we describe HapCompass-Tumor a computational modeling and algorithmic framework for haplotype assembly of copy number variable cancer genomes containing haplotypes at different frequencies and complex variation. We extend our polyploid haplotype assembly model and present novel algorithms for (1) complex variations, including copy number changes, as varying numbers of disjoint paths in an associated graph, (2) variable haplotype frequencies and contamination, and (3) computation of tumor haplotypes using simple cycles of the *compass graph* which constrain the space of haplotype assembly solutions. The model and algorithm are implemented in the software package *HapCompass-Tumor* which is available for download from [http://www.brown.edu/Research/Istrail\\_Lab/](http://www.brown.edu/Research/Istrail_Lab/).

*Keywords:* haplotype assembly; haplotype phasing; tumor haplotypes.

## 1. Introduction

Cancer is the worldwide leading cause of death and the second leading cause of death in the United States. Despite the tremendous amount of effort and resources spent on cancer research, our knowledge of the disease pathology is limited and the outlooks for certain types of cancer are usually dire. The commercialization of high-throughput sequencing platforms in the last decade has accelerated the growth of cancer genomics research dramatically. Since the first whole genome tumor sample was sequenced in 2008,<sup>1</sup> there have been hundreds of studies on numerous cancer types.<sup>2-5</sup> One of the fundamental computational challenges common to many of these studies is to separate the true driver mutation signal from the biological noise (e.g. passenger mutations) and experimental noise (e.g. sequencing errors). While it is possible to map sequence reads from tumor samples to a reference genome and call genomic variants, it is exceedingly difficult to determine the parental chromosome of origin for each variant allele – that is, the variant’s *phase*. But, the chromosomal sequence of alleles, or *haplotype*, is important for elucidating genomic events critical to the understanding of cancer like gene fusions or driver mutations.

A theory for carcinogenesis formulated by Knudson in 1971 demonstrates the importance

of haplotype phase in cancer.<sup>6</sup> In the two-hit hypothesis, Knudson suggested that in order to cause cancer, at least two “hits” have to take place. The first “hit” is usually an inherited mutation, and the second “hit” is a somatic mutation in the same gene or a different gene in the same pathway occurring later in life and out of phase with the first mutation. Having the ability to reconstruct tumor haplotypes would enable the discovery of such compound heterozygous relationships between variants and enhance our ability to identify driver mutations.

The computational problem of *haplotype assembly* aims to compute the sequence of co-inherited variant alleles for each chromosome given a set of aligned sequence reads and variants.<sup>7,8</sup> Haplotype assembly of diploid genomes has been addressed by many researchers<sup>9,10</sup> and several haplotype assembly algorithms for diploid genomes are available for use.<sup>11,12</sup> However, the methodologies for diploid haplotype assembly are unable to model polyploid genomes or complex copy number aberrations (CNA). Recently, we developed HapCompass-Polyploidy, the first modeling and algorithm for haplotype assembly in genomes with more than two sets of homologous chromosomes (polyploidy).<sup>13</sup> The HapCompass-Polyploidy algorithm assembles pairs of variants in polyploid genomes and then produces a haplotype assembly consistent with the pairwise variant phasings.

Cancer genomes have many similarities with polyploid genomes but present additional complexities that current methodologies do not model. Sequencing reads sampled from cancer patients exhibit a mixture of normal diploid cells and heavily rearranged, aneuploid cells. This introduces two major complexities into the haplotype assembly model: (1) heavily rearranged or translocated chromosomes will exhibit changes in copy number and (2) the heterogeneous nature of tumor samples requires reconstruction of more than two haplotypes each with a sample frequency which biases sequence read coverage.

Before these complexities can be modeled, the spectrum of variation must be inferred. While early cancer research was focused on small variants such as single nucleotide variants (SNV) and indels in a single gene or a small set of genes, advances in technology have enabled us to study large structural variants such as CNAs and large chromosomal rearrangements in tumor genomes. Several recent studies on multiple tumor genomes have found the important role of these large structural variants in tumor development.<sup>3,4,14,15</sup> In general, detection of cancer variation with sequencing data involves detecting those variants that are supported in the tumor genome but not found in the normal genome. The algorithms can be largely divided into three categories determined by the variant type they are trying to detect, i.e. small variants (SNVs and indels), CNAs and complex structural variants (translocations, duplications, and inversions).

Strelka jointly models the normal sample as a mixture of germline variation with noise, and the tumor sample as a mixture of the normal sample with somatic mutations, in a Bayesian framework.<sup>16</sup> VarScan 2 also uses the sequence reads from tumor and normal cells simultaneously, but uses a one tailed Fisher’s exact test to determine whether the variants are somatic, normal, or loss of heterozygosity.<sup>17</sup> Control-FREEC not only uses the coverage information but also the read count frequencies to estimate CNAs in tumor samples.<sup>18</sup> Control-FREEC also normalizes the tumor read depths by GC content and mappability and hence a normal genome is not required, although it could also be used for normalization.



Detection of large structural variations is often made possible by exploiting the properties of paired-end sequence reads. For example, the insert sizes of reads that are mapped to both sides of a large deletion would appear to have much larger insert sizes than the rest of the population. CREST first looks for a cluster of soft-clipped reads that exhibit evidence of a break point for a structural variant, and then locates the other break point by scanning the location neighboring the paired read.<sup>19</sup> However, the accuracy of these methods can be seriously affected when there is contamination in the samples. Cibulskis *et al* developed a Bayesian model to estimate the level of cross-individual contamination in each sample.<sup>20</sup> Contamination may also exist within an individual; tumor tissue can be contaminated with normal DNA and vice versa. Both incorrect variant calling as well as sequence contamination represent sources of complexity and errors for haplotype assembly.

In this work, we leverage the existing literature and tools for cancer genome variant inference and build on the polyploid HapCompass model to construct the first methodology for cancer genome haplotype assembly. In Section 2 we provide the necessary details of the HapCompass polyploid model and extensions for cancer genome haplotype assembly. The modeling section is followed by Section 3 which describes the HapCompass-Tumor algorithm and Section 4 which evaluates the implementation of the algorithm on cancer genome data. Finally, Sections 5 and 6 present a discussion of alternative models of cancer genome haplotype assembly, limitations and extensions to our model, future work, and conclusions.

## 2. Modeling

Let  $k$  be an integer representing the number of unique tumor haplotypes in a sample of tumor tissue. Because the tumor is actively evolving, this  $k$  may vary for independent samples of the same tumor. We assume that each sequence read is sampled from a single haploid fragment generated from one of the  $k$  haplotypes; this property enables the building of haplotype phase relationships between alleles in sequence reads that contain two or more *heterozygous* variants (homozygous variants do not provide phase information for assembly). The *phase-informative* sequence reads and variants are modeled with two graph structures termed the compass graph,  $G_C$ , and chain graph,  $G_h$ . These data structures are described in Aguiar *et al.* 2013 but their definitions are repeated here in order to present the novel aspects of the model for tumor genomes.<sup>13</sup>

The compass graph  $G_C(V_C, E_C)$  has  $v \in V_C$  for each variant and  $(v_i, v_j) \in E_C$  if variants  $v_i$  and  $v_j$  are contained within a sequence read. Edges  $(v_i, v_j)$  are annotated with the most likely haplotype phasing between variants  $v_i$  and  $v_j$  given the set of reads that contain both  $v_i$  and  $v_j$  (Figure 1).

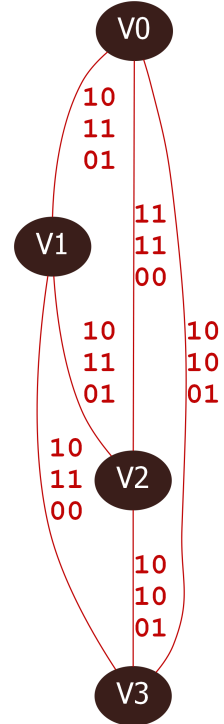


Fig. 1. An example tumor sample  $G_C$  with three unique haplotypes. Vertices are variants and edges show pairwise haplotype assemblies.

### 2.1. Phasing edges of $G_C$

Given the probability of sequencing error,  $s_e$  and a set of reads overlapping the two variants,  $r_1, \dots, r_n$ , the likelihood of a particular phasing,  $p_p$  from the set of all phasings between the two variants  $P$  can be computed as in Equation 1. Edges of  $G_C$  are phased by choosing the  $p_p$  that maximizes this likelihood.

$$L(p_p | s_e, r_1, r_2, \dots, r_n) = \frac{P(r_1 | s_e, p_p) \cdots P(r_n | s_e, p_p)}{\sum_{i=1}^{|P|} P(r_1, r_2, \dots, r_n | s_e, p_i)} \quad (1)$$

Equation 1 models haplotypes that are in equal proportion which may not be true for heterogeneous tumor samples. Thus the likelihood must be modified to accommodate the different frequencies of haplotypes. Consider the normal haplotype contamination that is often present in tumor sequence samples. Contamination may be modeled by jointly assembling the  $k$  tumor haplotypes with two low frequency normal haplotypes. Therefore, the probability of a haplotype  $h$  with frequency  $f_h$  in the phased haplotypes of a pair of variants can be expressed as  $p(h | s_e, p) = \sum_{h \in P} f_h F(s_e, p, h)$  where  $F$  is a function that takes the sequencing error probability  $s_e$ , the set of all haplotypes for the two variant phasing  $p$  and the particular haplotype  $h$  and computes the probability of generating a read containing haplotype  $h$ .

For example, assume the three haplotypes 00, 00, and 11 exist between two variants and one of the 00 haplotypes was considered contamination at frequency 10%. If the other two haplotypes were in equal proportions, then

$$P(00 | s_e, \{00, 00, 11\}) F(s_e, \{00, 00, 11\}, 00) = (1 - s_e)^2 \cdot 0.1 + (1 - s_e)^2 \cdot 0.45 + (s_e)^2 \cdot 0.45 \quad (2)$$

The number of unique phasings of an edge depends on the number of unique tumor haplotypes  $k$  and the allele content of the variant pair. Let the number of 1 alleles for variants  $v_i$  and  $v_j$  be  $l(v_i)$  and  $l(v_j)$  respectively. Then, the number of phasings of an edge is upper bounded by  $\min \left( \binom{k}{l(v_i)}, \binom{k}{l(v_j)} \right)$ . This is a bound and not equality because some phasings may be repeated in this enumeration.

### 2.2. Chain graph

Haplotype phasings of the edges of  $G_C$  can be extended to paths. Because two adjacent edges share a variant, haplotypes with the same allele can be merged on the shared vertex. If two paths in  $G_C$  of length  $i$  and  $j$  vertices are merged, the new phasing will have  $i + j - 1$  variants.

For paths or trees in  $G_C$ , there is exactly (at least) one consistent haplotype phasing, with respect to the edge phasings along the path or tree, for genomes with  $k = 2$  ( $k > 2$ ). In contrast, simple cycles in  $G_C$  may be either *conflicting* or *non-conflicting* depending on how many phasings are consistent with the cycle. A *conflicting* cycle does *not* have a consistent phasing while a *non-conflicting* cycle has at least one. The *chain graph*  $G_h$  is constructed for each simple cycle to determine its conflicting state.<sup>13</sup>

The chain graph  $G_h(V_h, E_h)$  is constructed for a path or simple cycle  $c = ((v_1, v_2), \dots, (v_{s-1}, v_s), (v_s, v_1))$  in the compass graph  $G_C$ . We introduce  $k$  haplotype vertices corresponding to the phasing for each edge  $(v_i, v_j)$  in the path or cycle. Vertices in  $G_h$  created

from adjacent edges of  $G_C$  share a variant; edges connect vertices in  $G_h$  if they share a variant and allele. Then, source nodes  $s_1, \dots, s_k$  are arbitrarily assigned to vertices at level 1 and sink nodes  $t_1, \dots, t_k$  are assigned to vertices at level  $s$  if the level  $s$  vertex shares an allele with the level 1 vertex. Vertices are annotated with  $t_i$  if there exists at least one  $s_i$  to  $t_i$  path which is computed by a depth first from each source.  $G_h$  can be described as a *trellis graph* in which the vertices can be divided into levels; each level in this case corresponds to an edge of  $G_C$ . Trellis graphs have a wide range of applications including communication network topology and survivability, encryption, encoding and decoding, and are a central data structure in Markov models.

### 2.3. Disjoint $s_i t_i$ paths in the trellis graph $G_h$

We now present new results on the theoretical properties of this graph and extensions to phasing the entire compass graph. A *valid phasing of a path* of compass graph edges  $e_{1,2}, \dots, e_{s-1,s}$  is defined as  $k$  vertex-disjoint paths from level 1 to level  $s$  in the corresponding  $G_h$ . A *valid phasing of a cycle* of compass graph edges  $e_{1,2}, \dots, e_{s,1}$  is defined as  $k$  vertex-disjoint paths from each source  $s_i$  to its corresponding sink  $t_i$  in the corresponding  $G_h$ . There always exists at least one phasing for paths of  $G_C$  by definition of  $G_h$ ; cycles may not exhibit a valid phasing (Lemma 2.1).

**Lemma 2.1.** *There exists at least one valid phasing of  $k$  haplotypes for a cycle  $c$  if and only if there exists a valid matching between sink node annotation and chain graph nodes at each level of  $G_c$ .*

**Proof.** If: Adjacent edges share a variant and thus the number of  $x$  alleles at level  $i$  must equal the number of  $x$  alleles at level  $i + 1$  where  $x$  is any allele of the shared variant. If there is a matching at level  $i$  and  $i + 1$ , then there must exist an edge between valid haplotype phase nodes because they share a common allele (adjacent levels). One can extend a valid haplotype phasing path from level  $i$  to  $i + 1$  using the edge generated by the shared allele. Only-if: Assume one level does not have a valid matching; then, either (1) at least two haplotypes share a phased haplotype node or (2) at least one phased haplotype node contain no sink node annotation. Case (1): multiple haplotype paths must share a phased haplotype node which breaks the vertex disjointness condition. Case (2): each level has exactly  $k$  nodes each of which must be taken once. If one or more phased haplotype nodes contain no sink annotation, then at least one phased haplotype node must be shared by 2 or more haplotype paths which breaks vertex disjointness.  $\square$

We will use this property of  $G_h$  later in the computation of the tumor haplotype phasing.

### 2.4. Copy number aberrations and translocations in $G_h$

The chain graph and disjoint paths framework accommodates modeling the types of variation typical of tumor genomes (Figure 2). CNAs insert or remove large genomic regions. Genomic deletions are modeled as an edge connecting the variants flanking the deletion breakpoint. In this case, the model still expects the computation of  $k$  disjoint paths spanning the deletion.

Large insertions of genetic material can be modeled as the addition of a temporary path in between or potentially overlapping vertices of  $G_h$ . The number of disjoint paths in this case changes to  $k + 1$ . Translocations may be modeled in  $G_h$  by combining deletions and insertions.

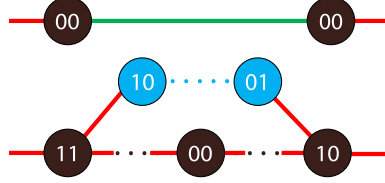


Fig. 2. Deletions and insertions are modeled with disjoint paths. The green edge models a deletion which effectively removes the deleted variants in the chain graph. The blue node insertion adds an extra path in  $G_h$ .

### 2.5. Disjoint subgraphs in the general chain graph

The general chain graph  $G_g$  is our final graph structure for representing the overall phasing of tumor genomes. Because there may be many matchings at each level of  $G_h$ , haplotype assembly of non-conflicting cycles in  $G_C$  will yield a set of potential phasings. The haplotype phasings of  $G_h$  constrain the haplotype assembly to include one of the  $k$  disjoint path solutions.

$G_g$  is built from the conflict-free spanning tree cycle basis of  $G_C$ . The vertices of  $G_g$  are constructed in a similar manner as  $G_h$ ; each edge  $(v_i, v_j)$  of  $G_C$  generates a vertex for each haplotype in the phasing of  $(v_i, v_j)$ . Each  $G_h$  constructed from a non-conflicting cycle of  $G_C$  defines a set of edge adjacencies; these adjacencies are represented in  $G_g$ . Therefore, if two edges are adjacent in a  $G_h$ , then they are also adjacent in  $G_g$ . Because of Lemma 2.1, we can determine the number of disjoint path solutions passing through adjacent levels  $i$  and  $j$  by simply computing the valid extensions of matchings from level  $i$  to  $j$ . We assume each of the  $l$  valid extensions of the sets of matchings at adjacent levels  $e_i$  and  $e_j$  are equally likely. Then, the weight of a particular extension  $\frac{w(e_i)w(e_j)}{l}$  where  $w(e_i)$  is the score or likelihood of edge  $e_i$ , is added to the edges of  $G_h$  (and  $G_g$ ).

However unlike  $G_h$ ,  $G_g$  is not necessarily a trellis graph if the cycles in the basis do not agree on the ordering of edge adjacencies (Figure 3). If  $G_g$  were a tree, finding a phasing could be modeled as packing disjoint Steiner trees or disjoint spanning trees. Instead, we model the computation of the tumor haplotype assembly as the  $k$ -maximum weight node-disjoint spanning tree problem. That is, we compute a set of  $k$  node-disjoint (within levels) spanning trees in  $G_g$  whose total weight is maximum over all  $k$  node-disjoint spanning trees and includes every vertex in  $G_g$ .

## 3. HapCompass-Tumor Algorithm

HapCompass-Tumor optimizes the minimum weighted edge removal (MWER) problem. MWER aims to compute a set of edges  $L$  of minimum weight, whose removal resolves all conflicting cycles of  $G_C$ . After all conflicting cycles have been removed, each non-conflicting cycle's  $G_h$  is added to  $G_g$ .  $G_g$  represents the constrained solution space by incorporating the





Fig. 3. (Left) An example tumor genome  $G_C$  with three non-conflicting cycles. Dashed lines represent edges not in the spanning tree of  $G_C$ . The inclusion of each non-tree edge creates a cycle in the cycle basis of  $G_C$ . The two inner cycles  $((v_0, v_1), (v_1, v_3), (v_3, v_0))$  and  $((v_0, v_2), (v_2, v_3), (v_3, v_0))$  create the red-edge adjacencies in  $G_g$  (right). Computing the haplotype assembly of a tree ( $G_g$  with just the red edges) is simple. However, if the blue non-tree edge is added, the edge adjacency  $((v_0, v_1), (v_0, v_2))$  is included in  $G_g$  creating a cycle.

valid haplotype assemblies on subsets of variants computed from each non-conflicting  $G_h$  (Algorithm 1).

**input** : Sequence reads, variant calls, and number of distinct haplotypes  $k$   
**output**:  $k$  haplotypes

$G_C \leftarrow$  spanning tree cycle basis  
 $C_C \leftarrow$  set of conflicting simple cycles with respect to  $G_C$   
**for**  $c_C \in C_C$  **do**  
    Remove edge with smallest likelihood in  $c_C$   
    Reconstruct  $G_C$   
**end**  
Compute  $G_g$   
 $C_N \leftarrow$  set of non-conflicting simple cycles with respect to  $G_C$   
**for**  $c_N \in C_N$  **do**  
    Compute  $G_h$  with respect to  $c_N$   
    Compute matchings at each level of  $G_h$   
    Compute disjoint paths of  $G_h$   
    Increase the weight of each edge  $e$  between levels shared by  $G_h$  and  $G_g$  in  $G_g$  proportional to the number of disjoint paths using edge  $e$  and the likelihood of each edge (Equations 1 and 2)  
**end**  
Compute a maximum weight spanning tree of the adjacencies in  $G_g$   
Output the haplotype assembly computed from the spanning tree of  $G_g$

**Algorithm 1:** HapCompass-Tumor

The final step involves computing  $k$  spanning trees in  $G_g$  which are node disjoint in respect to haplotype level vertices. Adjacencies between levels in  $G_g$  correspond to matchings between the haplotype nodes (Figure 4 right). So, HapCompass-Tumor computes  $k$  disjoint spanning trees corresponding to the  $k$  tumor haplotypes. We have implemented two algo-

gorithms inspired by Kruskal's and Prim's algorithms for computing maximum spanning trees. The principle difference between the two algorithms in the context of HapCompass is the Kruskal-like algorithm focuses on constructing disjoint trees by including strong phasings on the same haplotype (edges of  $G_g$ ) while the Prim-like algorithm phases all haplotypes between two levels at a time (vertices of  $G_g$ ).

We illustrate the modeling and algorithm with a series of examples. Let the compass graph  $G_C$  of a tumor sample with three unique haplotypes be shown in Figure 1. Then, if  $(v_0, v_3)$ ,  $(v_2, v_1)$ , and  $(v_3, v_2)$  are the non-tree edges of  $G_C$ , the chain graphs in Figure 4 (left) are constructed. Figure 4 (right) shows the  $G_g$  updated after the disjoint paths and weights of edges in  $G_h$  are computed and distributed to  $G_g$ .

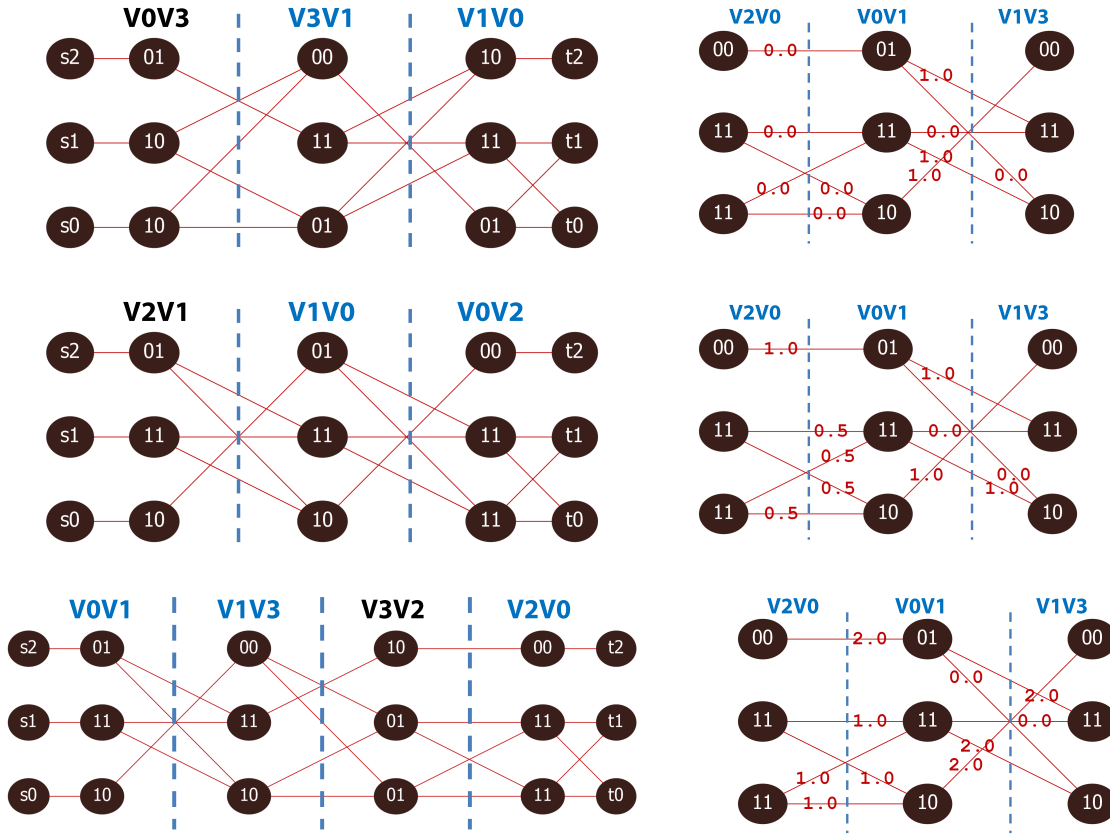


Fig. 4. (Left) chain graphs ( $G_h$ ) from the compass graph in Figure 1. The level corresponding to edges in  $G_C$  are denoted by black (non-tree edges) and blue (spanning tree edges) lettering above the vertices. In this example, the edge phasing probabilities in  $G_C$  are all 1. So, an edge connecting level  $i$  to level  $j$  which is in  $b$  disjoint path solutions will receive a weight of  $b/d$  if there are  $d$  unique disjoint path solutions from level  $i$  to level  $j$ . The weights of edges calculated from disjoint  $s_i t_i$  paths in each  $G_h$  are added to the  $G_g$  (right).

#### 4. Results

We implemented HapCompass-Tumor and evaluated its performance on simulated tumor haplotypes. In these experiments we use insert size as a proxy for the computed haplotype length. It has been shown that the dominant factor in producing long haplotype assemblies is the

length between the read pairs.<sup>13,21</sup> Briefly, if the length between two variants is  $x$  and the insert size is  $y$ , then a sequence read can never span the two variants if  $x > y$ .

#### 4.1. *Dependence on insert size and error rates*

Using the sequence for the *BRCA1* breast cancer susceptibility gene, we simulated three hyper variable tumor haplotypes. Distance between variants were distributed normally  $\sim N(500, 50)$ . The following procedure was repeated 250 times for each data point in Figure 5. Given the set of variants which remained fixed for each experiment, a random phasing is computed that is consistent with the allele distributions. We then sampled 10000 phase-informative simulated reads from the true haplotypes and computed the average edit distance between assembled and true haplotypes. We compared the distance of haplotype assemblies for the randomly generated triploid *BRCA1* genes while varying sequence read insert size, standard deviation of insert size, and single base substitution error rate.

Figure 5 (left) demonstrates several interesting trends. First, as the insert size is increased the haplotype assemblies become more accurate. Second, the more variable the insert length, the more accurate the haplotype assembly. A hyper variable insert length appears to have a similar effect as increasing the insert size. These findings confirm patterns observed in conventional diploid haplotype assembly. Finally, while the error rate does affect haplotype assembly accuracy, as long as the error rate is less than 0.2%, the haplotype assemblies are similar in quality. This phenomenon is likely caused by the constant coverage coupled with uncertainty in phasing the edges of  $G_C$ . When the coverage is fixed and the insert sizes are short, haplotype assemblies are smaller but more accurate. Conversely, when error rates reach a threshold where edge phasings are no longer accurately called, the haplotype assembly quality suffers.

#### 4.2. *Cancer genome heterogeneity*

We also compared the accuracy of haplotype assembly in terms of tumor genome heterogeneity (Figure 5 right). Sequencing parameters were fixed to produce insert sizes between 500 and 2500, short insert size standard deviations, 10000 sequence reads, and no errors. Each data point contains the average of 250 haplotype assembly edit distances. The more unique tumor haplotypes in the sample the less accurate the solution. The increasing edit distance with 5 unique haplotypes between insert sizes 2000 and 2500 is likely an effect of the rising uncertainty of edge phasings when coverage is kept fixed and more edges are being generated in  $G_C$ .

#### 4.3. *NA12878*

We simulated paired tumor sequence reads and their mappings with Enhanced Artificial Genome Engine (EAGLE) developed by Illumina Cambridge Ltd (personal communications). The sequencing parameters were set to model paired-end Illumina data with 101bp read lengths and a mixture of long (length= $N(60000, 141^2)$ ) and short (empirical distribution from  $2 \times 101$  runs, with median size  $\sim 300bp$ ) fragment sizes. The variants simulated include SNV and indels called in NA12878 by the Genome in a Bottle Consortium<sup>22</sup> and the HCC1187 tumor sample

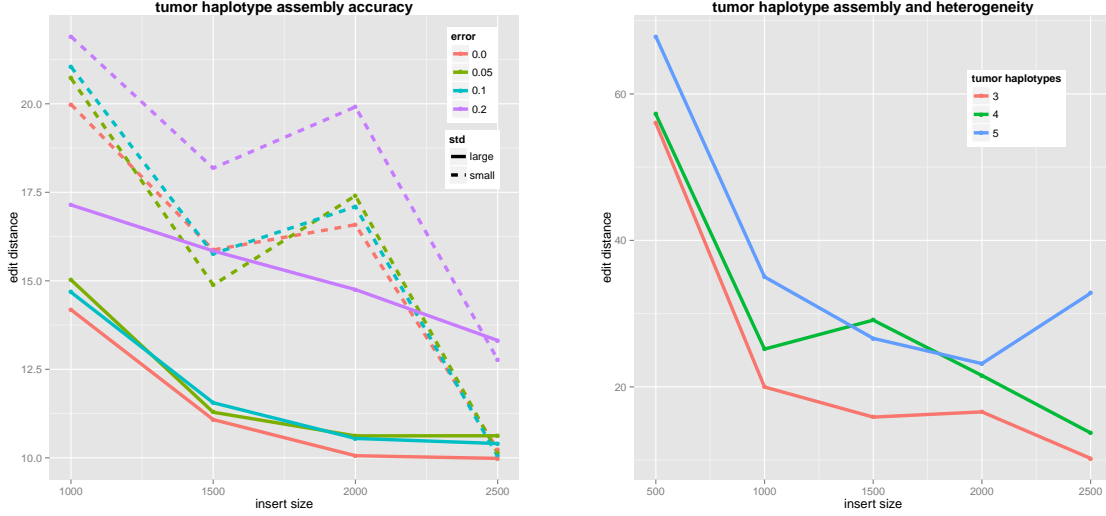


Fig. 5. (Left) The average edit distance between haplotypes and the simulated true haplotypes is calculated with a fixed coverage and varying insert sizes, error rates (error), and standard deviations (std). (Right) Haplotype assembly accuracy is plotted as a function of the number of tumor haplotypes in the sample.

(downloaded from Illumina’s Basespace<sup>23</sup>). Variants were combined then randomly divided into two sets for each homologous chromosome, with 30X coverage for the first chromosome and 15X coverage for the second to simulate tumor genome amplification. Sequence reads were mapped to their simulated location after single base mismatches were introduced according to empirical error rates.

We evaluated HapCompass-Tumor on all autosomes of the EAGLE simulated data and longer reads simulated using HapCompass. The reads simulated from HapCompass include medium (200bp) and long (2000bp) read lengths with error rates of 2% and 5% respectively to model the higher error rates associated with long-read high-throughput sequence technologies. We used the number of allele bit flips required to map the sequence reads to the assembled haplotypes as the evaluation metric. Table 1 shows the results for HapCompass-Tumor using the Kruskal-like and Prim-like algorithms for resolving  $G_g$ . Additionally, we implemented a scoring scheme that scores pairs of vertices with more diversity in haplotype sequence higher (termed *Diverse* in Table 1). This scheme is designed to limit uninformative pairs of vertices in the spanning tree of the compass graph  $G_C$ .

Table 1 demonstrates that the accuracy of the haplotype assembly depends minimally on the selection of algorithm when using Illumina-like sequencing parameters. However, as the read length increases, the Kruskal-like algorithm becomes favorable.

## 5. Discussion

Opportunities exist to extend HapCompass-Tumor to address some of the limitations in the current model. First, HapCompass-Tumor only computes a single solution when the compass graph model allows computation of suboptimal solutions. Phase extension in  $G_g$  is deterministic but many highly probable suboptimal solutions may exist. As long as the number of

Table 1. The proportion of incorrectly mapped alleles (error) by  $G_g$  resolution algorithm. Sequence data was simulated for 1000 Genomes Project individual NA12878 using EAGLE to simulate Illumina reads and HapCompass to simulate reads with medium (200bp, 2% error rate) and long (2000bp, 5% error rate) read lengths.

$G_g$ resolution	error (autosomes, EAGLE)	error (chr20, 200bp)	error (chr20, 2000bp)
Kruskal	0.002658	0.02079	0.04626
Kruskal Diverse	0.002659	0.02071	0.04679
Prim	0.002659	0.02789	0.05639
Prim Diverse	0.002659	0.02631	0.05867

alternative disjoint paths is bounded by a low degree polynomial, we can carry these partial solutions to the assembly step and report multiple haplotype assemblies.

Second, incorporating *a priori* knowledge of haplotype distributions from population samples or long read lengths would improve the assembly. For example, we assumed each valid haplotype phasing for a cycle in  $G_C$  is equally likely. However, this assumption can be easily modified to accommodate known haplotype likelihoods in the area (e.g. linkage disequilibrium). Consider a collection of valid disjoint paths for a cycle in  $G_C$ ; if the probability of both phasings is 1 and the edge extension has  $i$  distinct matchings, then each matching is given a weight  $\frac{1}{i}$ . If, however, one of the haplotypes in an extension is never observed in the population, HapCompass-Tumor could penalize the extension.

A related application of HapCompass-Tumor is in cancer panomics. Much attention in cancer research has been focused on allelic specific expression (ASE). Studies have shown that germline ASE is associated with cancer risk,<sup>24,25</sup> and somatic ASE is associated with tumor development.<sup>26</sup> ASE in cancer was found not only correlated with CNAs,<sup>26</sup> but also with allelic specific methylation (ASM).<sup>27</sup> Existing algorithms for detecting ASE with RNA-seq and detecting ASM with Bisulfite-Seq do not usually make use of phased genotype information.<sup>26,28</sup> We therefore propose using the phased haplotypes from whole genome sequencing of tumor samples as a reference for RNA-seq and Bisulfite-seq alignment when such data is available.

Finally, the viral quasispecies reconstruction (VQR) problem aims to compute the spectrum of viral quasispecies haplotypes from the sequence reads of a heterogeneous viral sample. The problems of haplotype assembly and VQR are similar but the research literature is largely independent due to the inability of haplotype assembly algorithms to model more than two sets of homologous haplotypes. However, it is possible to model VQR with HapCompass-Tumor by leaving the number of haplotypes in the sample ( $k$ ) as an unknown parameter. Two possible approaches include inferring the number of quasispecies *a priori* and then performing haplotype assembly with  $k$  unique haplotypes or computing assemblies for a number of different  $k$  and comparing the quasispecies solutions. But, using a general haplotype assembly tool for VQR does not take advantage of two critical properties of most viral genomes: (1) knowledge of the phylogenetic relationships between mutations is known for well-studied viral genomes especially those under selective pressures from treatment and (2) the genomes are many orders of magnitude smaller than eukaryotes.

## 6. Conclusions

In this work, we developed algorithms and models for tumor genome assembly building on our existing haplotype assembly framework HapCompass. We demonstrated how to model tumor haplotype heterogeneity and haplotypes containing CNAs and translocations. The HapCompass-Tumor algorithm was presented using the combined evidence of cycles in  $G_C$  and disjoint paths in  $G_h$  to inform which haplotype assemblies in  $G_g$  are probable. Finally, we evaluated the HapCompass-Tumor algorithm on simulated cancer data showing that, while the accuracy is a function of many parameters including the level of cancer genome heterogeneity, we are still able to produce accurate haplotype assemblies. HapCompass-Tumor is available for download from [http://www.brown.edu/Research/Istrail\\_Lab/](http://www.brown.edu/Research/Istrail_Lab/).

## 7. Acknowledgements

We thank Lilian Janin and Anthony Cox at Illumina Cambridge Ltd for sharing and helping us with the EAGLE simulator. This work was supported by the National Science Foundation [1048831 and 1321000 to S.I.].

## References

1. T. J. Ley, E. R. Mardis *et al.*, *Nature* **456**, 66 (November 2008).
2. E. D. Pleasance, R. K. Cheetham *et al.*, *Nature* **463**, 191 (2009).
3. M. Meyerson, S. Gabriel and G. Getz, *Nature Reviews Genetics* **11**, 685 (October 2010).
4. The Cancer Genome Atlas, *Nature* **490**, 61 (September 2012).
5. E. R. Mardis, *Curr. Opin. Genet. Dev.* **22**, 245 (Jun 2012).
6. A. G. Knudson, *Proceedings of the National Academy of Sciences* **68**, 820 (1971).
7. R. Lippert, R. Schwartz, G. Lancia and S. Istrail, *Brief Bioinform* **3**, 23 (March 2002).
8. S. Istrail, *The Haplotype Phasing Problem*, tech. rep., Celera Genomics (2002).
9. R. Schwartz, *Commun. Inf. Syst.* **10**, 23 (2010).
10. F. Geraci, *Bioinformatics (Oxford, England)* **26**, 2217 (September 2010).
11. D. Aguiar and S. Istrail, *J. Comput. Biol.* **19** (2012).
12. V. Bansal and V. Bafna, *Bioinformatics* **24**, i153 (August 2008).
13. D. Aguiar and S. Istrail, *Bioinformatics* **29**, i352 (2013).
14. L. Ding, M. J. Ellis *et al.*, *Nature* **464**, 999 (May 2010).
15. W. Lee, Z. Jiang *et al.*, *Nature* **465**, 473 (May 2010).
16. C. T. Saunders, W. S. W. Wong *et al.*, *Bioinformatics (Oxford, England)* **28**, 1811 (July 2012).
17. D. C. Koboldt, Q. Zhang *et al.*, *Genome research* **22**, 568 (March 2012).
18. V. Boeva, A. Zinovyev *et al.*, *Bioinformatics (Oxford, England)* **27**, 268 (January 2011).
19. J. Wang, C. G. Mullighan *et al.*, *Nature Methods* **8**, 652 (2011).
20. K. Cibulskis, A. McKenna *et al.*, *Bioinformatics (Oxford, England)* **27**, 2601 (September 2011).
21. B. V. Halldorsson, D. Aguiar and S. Istrail, *Pacific Symposium of Biocomputing*, 88 (2011).
22. Genome in a Bottle Consortium, NIST NA12878 Highly Confident integrated genotype (April 2013), [ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant\\_calls/NIST/](ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/).
23. Illumina Inc., BaseSpace G.C.C. (April 2013), <https://basespace.illumina.com/home/index>.
24. L. Valle, T. Serena-acedo *et al.*, *Science* **321**, 1361 (2009).
25. C. Gao, K. Devarajan *et al.*, *BMC genomics* **13**, p. 570 (January 2012).
26. B. B. Tuch, R. R. Laborde *et al.*, *PloS one* **5**, p. e9317 (January 2010).
27. P.-C. Lin, E. G. Giannopoulou *et al.*, *Neoplasia* **15**, 373 (2013).
28. F. Fang, E. Hodges *et al.*, *Proceedings of the National Academy of Sciences* **109**, 7332 (2012).

# EXTRACTING SIGNIFICANT SAMPLE-SPECIFIC CANCER MUTATIONS USING THEIR PROTEIN INTERACTIONS

LIVIU BADEA<sup>†</sup>

*University Politehnica Bucharest and Bioinformatics Group, ICI  
8-10 Averescu Blvd, Bucharest, Romania  
Email: badea.liviu@gmail.com*

We present a joint analysis method for mutation and gene expression data employing information about proteins that are highly interconnected at the level of protein to protein (pp) interactions, which we apply to the TCGA Acute Myeloid Leukemia (AML) dataset. Given the low incidence of most mutations in virtually all cancer types, as well as the significant inter-patient heterogeneity of the mutation landscape, determining the true causal mutations in each individual patient remains one of the most important challenges for personalized cancer diagnostics and therapy. More automated methods are needed for determining these “driver” mutations in each individual patient. For this purpose, we are exploiting two types of contextual information: (1) the pp interactions of the mutated genes, as well as (2) their potential correlations with gene expression clusters. The use of pp interactions is based on our surprising finding that *most AML mutations tend to affect nontrivial protein to protein interaction cliques*.

## 1. Introduction and motivation

Although various aspects of the cancer genome, such as gene expression, mutations, DNA copy number changes, or DNA methylation profiles have been studied (mostly) in isolation for more than a decade, their multi-modal, combined analysis has only recently been possible due to large scale projects such as The Cancer Genome Atlas (TCGA), as well as to the dwindling costs of high-throughput sequencing.

Landmark studies of the TCGA have for the first time revealed the genomic changes and their consequences in several cancer types, such as glioblastoma [1,2,3], ovarian [4], breast [5], squamous cell lung cancer [6], colorectal cancer [7] and acute myeloid leukemia [8].

Most of these and other integrated studies of the cancer genome use state of the art methods for analyzing the *separate* data types (such as gene expression, mutation, DNA copy number changes and DNA methylation profiles), and then try to correlate the *separate* findings into a global integrated picture of the cancer genome (for example by searching for mutation enrichment in consensus gene expression clusters, or by comparing miRNA clusters with expression clusters [8]).

Despite numerous attempts at a *joint* analysis of the various data types (as opposed to separate analyses), currently there is no *universally accepted* approach available.

---

<sup>†</sup> Work partially supported by grant PN-II-ID-PCE-2011-3-0198.

In this paper, we present a joint analysis method for mutation and gene expression data that employs information about proteins that are highly interconnected at the level of protein to protein (pp) interactions, which we apply to the Acute Myeloid Leukemia (AML) dataset obtained by TCGA [8].

Given the low incidence of most mutations in virtually all cancer types, as well as the significant inter-patient heterogeneity of the mutation landscape, determining the true causal mutations in each individual patient remains one of the most important challenges for personalized cancer diagnostics and therapy [18].

For example, since in AML only 3 genes have been found mutated with a frequency above 10% (FLT3, NPM1, and DNMT3A), the state of the art AML study of the TCGA group [8] has used the known gene annotations to determine the genes relevant for pathogenesis (based on a few categories deemed biologically significant by human investigators).

Still, annotations are imperfect and many genes have surprisingly heterogeneous functions. Moreover, annotations reveal nothing about gene interactions (except maybe pathway annotations, which are currently hopelessly incomplete). For example, the NPM1 gene is placed by the TCGA study in a category of its own, solely based on its high mutation rate in AML.

More automated methods are therefore needed for determining the mutations that have caused the disease in each individual patient, the so called “driver” mutations. For this purpose, we are trying to exploit two types of contextual information:

- (1) the protein-to-protein (pp) interactions of the mutated genes in question, as well as
- (2) their potential correlations with gene expression clusters.

These two types of contextual information are used in a synergistic manner.

The use of pp interactions is based on our surprising finding that *most AML mutations tend to affect complete pp interaction cliques*. More precisely, the protein-to-protein interaction network between AML mutated genes contains a large number of nontrivial maximal cliques (of size  $\geq 3$ ).<sup>\*</sup>

This is highly surprising given the very low number of somatic mutations in AML, much lower than in all other solid cancers analyzed to date [8]. The fact that mutations tend to affect cliques in the pp interaction network suggests the disruption of biological processes or protein complexes involving the corresponding protein cliques. It is as if such biological processes or complexes can be perturbed by mutations in any of their components. This is important since only very few mutations in AML (or other cancer types for that matter) have an incidence larger than 10%. Grouping mutations based on their pp interactions thereby enhances the statistical power of detecting correlations between mutations (the causal factors) and their transcriptional consequences, such as gene expression subtypes of the disease.

---

<sup>\*</sup> The nontrivial complete maximal cliques of mutated genes have an average size of  $\sim 3$ .



## 2. Data and preprocessing

### 2.1. *The TCGA AML dataset*

The TCGA Acute Myeloid Leukemia (AML) dataset was downloaded from the TCGA data portal<sup>†</sup>, as well as from the supplementary data of the TCGA landmark publication [8] (in preprocessed form). More specifically, we downloaded:

- *gene expression* data (RNASeqV2 UNC Illumina HiSeq, level 3 RSEM normalized data),
- *copy number variation* data (profiled using Affymetrix SNP6 arrays, level 4 data obtained using Gistic2),
- *somatic mutation* data (obtained using either whole-genome sequencing, or whole-exome sequencing),
- data regarding *gene fusions* (obtained from de novo assembly of RNA-sequencing data), as well as
- *clinical annotations*.

We retained 163 samples with simultaneous gene expression, copy number, mutation, gene fusion and clinical data.

### 2.2. *Generalized mutations*

Since somatic mutations, copy number aberrations and gene fusions can all act as drivers of the disease in individual patients, we defined “*generalized mutations*” as either:

- (1) expressed somatic mutations,
- (2) expressed fusion genes, or
- (3) significant copy number aberration events.

A somatic mutation in a given gene was considered *expressed* if the expression of the corresponding gene exceeded the expression threshold of 6 (on the  $\log_2$  scale).

Since gene fusions have been determined from de novo assembly of RNA-seq data, they were all considered to be expressed.

Copy number aberrations were considered *significant* if

- the corresponding gene’s expression levels were not uniformly low (below the expression threshold of 6, mentioned above), and
- they were accompanied by *concordant* gene expression changes with  $|Z| > 2$  (i.e. amplifications accompanied by gene up-regulation and deletions accompanied by gene down-regulation), and
- the copy number profile had at least a slight correlation (exceeding 0.3) with the gene’s expression profile.

There were 2142 genes with generalized mutations in at least one sample, with a total number of 5865 events, of which 1050 expressed mutations and 202 gene fusions (for more details, see

---

<sup>†</sup> tcga-data.nci.nih.gov

Table 1). Gene fusions  $g_1$ - $g_2$  were recorded as separate generalized mutations in  $g_1$  and  $g_2$  respectively, to allow their mixing with other generalized mutations in those genes.

Table 1. Generalized mutations in 163 AML samples

Generalized mutation type	Number of generalized mutations
CN deletions	3008
CN amplifications	1605
Somatic mutations	1041
Somatic mutations + CN deletions	8
Somatic mutations + CN amplifications	1
Gene fusions	193
Gene fusions + CN deletions	7
Gene fusions + CN amplifications	2
Total	5865

### 2.3. Protein-to-protein interaction data

We used the BioGRID protein interaction database<sup>†</sup> (version 3.2.101 for Homo Sapiens), which we restricted to the physical interactions. This resulted in 136201 interaction pairs involving 14791 unique human genes.

### 3. Proteins mutated in AML form pp interaction cliques

Compared to solid cancers, AML genomes have much lower numbers of mutations [8]. This is to be expected, as leukemias do not have to evade the source tissue and metastasize, as solid cancers do. (Along these lines, a two-hit model of leukemogenesis has been proposed by Gilliland [9].)

Interestingly however, restricting the BioGRID pp interaction network to the set of genes mutated in AML, we obtain a large number of pp interaction cliques. More precisely, we obtain 4160 maximal cliques<sup>§</sup> involving the 2142 genes with generalized mutations (of which 3564 *nontrivial* cliques involving more than one gene). The average nontrivial clique size was 2.96. Figure 1 depicts the corresponding distribution of nontrivial maximal clique sizes, showing that mutated genes form many large cliques.

Compared to the complete BioGRID interaction network, the edge density<sup>\*\*</sup> of the network of mutated genes is significantly larger ( $2.3 \cdot 10^{-3}$  versus  $5.6 \cdot 10^{-4}$ ), although the average clustering coefficient is smaller (0.1002 versus 0.1758).

<sup>†</sup> <http://thebiogrid.org/>

<sup>§</sup> Although the clique decision problem (testing whether a graph contains a clique larger than a given size) is NP-complete, while listing all maximal cliques may require exponential time (as there exist graphs with exponentially many maximal cliques [16]), finding all maximal cliques in our setting is reasonably fast (running times of the order of minutes on a 3GHz machine using a Matlab implementation of the Bron–Kerbosch algorithm [17]).

<sup>\*\*</sup> i.e. the number of edges divided by the maximal possible number of edges, i.e.  $n(n-1)/2$ , where  $n$  is the number of nodes of the network.

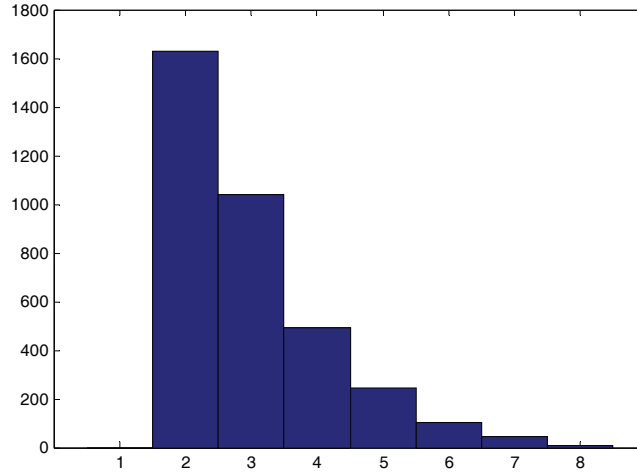


Fig. 1. The distribution of nontrivial maximal clique sizes of mutated proteins in the BioGRID pp interaction network.

Mutations affecting nontrivial protein interaction cliques suggest different ways of perturbing certain key biological processes or protein complexes involving the corresponding cliques. Therefore, although most individual mutations have a low incidence in the AML patient population (thereby masking their possible role in the pathogenesis of the disease), cliques tend to be mutated<sup>††</sup> in a higher number of patients and thus could be used to order mutations in individual patients. The supplementary table ‘*sample mutations clique cover.xls*’ (online at [www.ai.ici.ro/PSB2014/](http://www.ai.ici.ro/PSB2014/)) shows for each patient sample its mutations sorted in descending order of the number of samples in which the largest maximal clique containing the corresponding gene is mutated.

More precisely, in the following, by ‘clique’ we always mean ‘maximal clique’. We denote by  $M_{ms}$  the binary mutation matrix ( $M_{ms}=1$  iff sample  $s$  has mutation  $m$ ) and by  $C_{mc}$  the clique membership matrix ( $C_{mc}=1$  iff mutated protein  $m$  is involved in clique  $c$ ). We define the cover of a clique  $c$  to be the number of samples with mutations in at least a gene  $m$  of that clique:

$$\text{clique-cover}(c) = | \{ s \mid \exists m. M_{ms}=1 \text{ and } C_{mc}=1 \} |.$$

We can also define the largest clique associated to a given mutation  $m$  as a clique containing  $m$  having the largest clique cover<sup>††</sup>:

$$\text{largest-clique}(m) = c \text{ iff } C_{mc}=1 \text{ and } \forall c' \text{ such that } C_{mc'}=1, \text{ clique-cover}(c') \leq \text{clique-cover}(c).$$

Now, for each sample  $s$ , we can sort the mutations  $m$  in descending order of the cover of the largest clique associated to  $m$ :  $\text{clique-cover}(\text{largest-clique}(m))$ . The top mutations are likely causal,

<sup>††</sup> A clique is said to be mutated in a given sample iff at least one of its genes is mutated in that particular sample.

<sup>††</sup> in case there are several such largest cliques, we arbitrarily choose one.

as they or their interactors are mutated in large numbers of samples. For example, all acute promyelocytic leukemia samples (FAB code ‘M3’) have the PML and RARA fusion proteins at the top of the list.

#### 4. Joint analysis of gene expression data and mutations using pp interaction data

Although by using protein-to-protein interaction data we have obtained a reasonable ordering of (generalized) mutations w.r.t. their potential causal role in the disease, we still have not made use of all available data to the fullest. For example, we have only employed gene expression data for filtering out mutations in genes that are not expressed, but we have completely ignored any potential similarities in the transcriptomes of samples with different mutations.

In the following, we describe an approach that simultaneously looks for similarities among mutation and gene expression data and, most importantly, is able to extract potentially causal sample-specific mutations, despite their low frequency in the dataset.

By a *direct* joint clustering of gene expression and mutation data, we may only pick up the mutations with the highest incidence. To avoid this, instead of directly clustering mutation data, we cluster the pp interactions of the observed mutations with other mutated proteins. Mutated proteins with similar interactor sets (among the set of mutated proteins) will likely affect the same pathways or protein complexes and produce similar expression changes.

For example, assume sample  $s_1$  is affected by mutation  $m_1$ , while sample  $s_2$  is affected by a different mutation,  $m_2$ . Even with similar gene expression profiles,  $s_1$  and  $s_2$  may not be grouped into a common cluster  $k$ , since we wouldn’t know which of the mutations  $m_1$  and  $m_2$  to associate to  $k$ . If however,  $m_1$  and  $m_2$  have similar sets of interactors among the other mutated genes  $p_1, p_2, p_3, \dots$ , we could *cluster the interactor sets of the mutations* instead of the individual mutations, thereby merging  $s_1$  and  $s_2$  despite their different mutations.

##### 4.1. The joint clustering of expression and mutation interactor data

More formally, let  $s$  denote samples,  $g$  genes,  $m$  mutations,  $k$  clusters,  $X_{gs}$  the gene expression matrix,  $M_{ms}$  the binary (generalized) mutations matrix and  $P_{pm}$  the binary protein-to-protein interaction matrix involving mutated genes (although the matrix is symmetric, we use distinct  $p$  and  $m$  indices to distinguish the mutations  $m$  from their interactors  $p$ ).

Now, instead of jointly clustering the gene expression  $X_{gs}$  and mutation data  $M_{ms}$ , we cluster the gene expression data and the *mutation interactor data*  $\sum_m P_{pm} \cdot M_{ms}$  :

$$X_{gs} \approx \sum_k G_{gk} \cdot S_{sk} \quad (1)$$

$$\sum_m P_{pm} \cdot M_{ms} \approx \sum_k A_{pk} \cdot S_{sk} \quad (2)$$

where  $G_{gk}$ ,  $S_{sk}$  and  $A_{pk}$  are *gene*, *sample* and respectively *mutation interactor cluster matrices*. Note that the sample cluster matrix  $S_{sk}$  is common to the gene expression and mutation interactor data factorizations (1) and (2).

Running the nonnegative multirelational decomposition system MNMF<sup>§§</sup> [10,11] with a relative weight  $w=0.001$  for the mutation interactors (to enable the gene expression data to dominate the factorization), we obtain the cluster matrices  $S_{sk}$ ,  $G_{gk}$  and  $A_{pk}$  for samples, genes and mutation interactors respectively.

The mutation interactor clusters  $A_{pk}$  encode the frequently co-occurring mutation interactors  $p$  in the various clusters  $k$ , but do not tell us anything directly about the mutations proper. To obtain the *sample-specific mutations*  $m$  that lie behind these cluster-specific mutation interactors  $p$ , we solve the following *nonnegative least squares problem* (with  $M'_{ms}$  as unknown):

$$\sum_k A_{pk} \cdot S_{sk} \approx \sum_m P_{pm} \cdot M'_{ms} \quad (3)$$

using a multiplicative update algorithm that randomly initializes  $M'$  and then iteratively applies the following update rule until convergence:

$$M'_{ms} \leftarrow M'_{ms} \frac{(P^T \cdot Y)_{ms}}{(P^T \cdot P \cdot M')_{ms}} \quad (4)$$

where  $Y_{ps} = \sum_k A_{pk} \cdot S_{sk}$ .

Finally, we can use  $MM_{ms} = M'_{ms} \cdot M_{ms}$  as a measure of the significance of mutation  $m$  for given clustering. Mutations with higher  $MM_{ms}$  are deemed more causally relevant, as they better match the given gene expression clustering. Note that frequently occurring mutations tend to have higher  $MM$  scores, especially if they are not at odds with the gene expression clustering.

Figure 2 below is a graphical depiction of the decomposition (1)-(3). The system was implemented in Matlab.

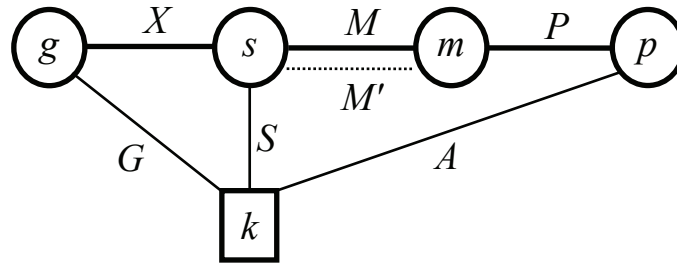


Fig. 2. The relational diagram corresponding to the decomposition (1)-(3). Circles correspond to entities ( $g$  genes,  $s$  samples,  $m$  mutations,  $p$  mutation interactors), while the boxed  $k$  represents the unknown clusters. Bold edges correspond to the original relations  $X, M, P$ , normal edges to inferred entity clusters  $G, S, A$ , and the dotted edge to the significant sample-specific mutations  $M'$ .

<sup>§§</sup> MNMF is a multirelational generalization of Nonnegative Matrix Factorization (NMF) [13,14] and of simultaneous NMF [15].

#### 4.2. The dimensionality of the factorization

Determining the optimal dimensionality  $n_c$  of the factorization (1)-(2) is tricky. Similar to Kim and Tidor [12], we performed a series of MNMF runs with progressively larger  $n_c$ , ranging from 2 to 50. To avoid overfitting, we performed a similar set of runs on randomized entity matrices and estimated the signal to noise ratio (SNR) as follows:

$$SNR(n) = \frac{\varepsilon_r(n)^2 - \varepsilon(n)^2}{1 - \varepsilon_r(n)^2}$$

where  $\varepsilon(n)$  and  $\varepsilon_r(n)$  are the relative factorization reconstruction errors for the original and respectively the randomized data. The dimensionality  $n_c=22$  was chosen to maximize the SNR (see Figure 3). Note that the clusters obtained by our nonnegative decompositions should not be confused with partitions of the samples into *disjoint* subgroups. They are rather biclusters corresponding to biological processes that may overlap in the various samples (as well as for certain genes).

We also tried the smaller dimensionality  $n_c=7$  obtained by optimizing NMF consensus sample clustering (a partitional method), as in [8].

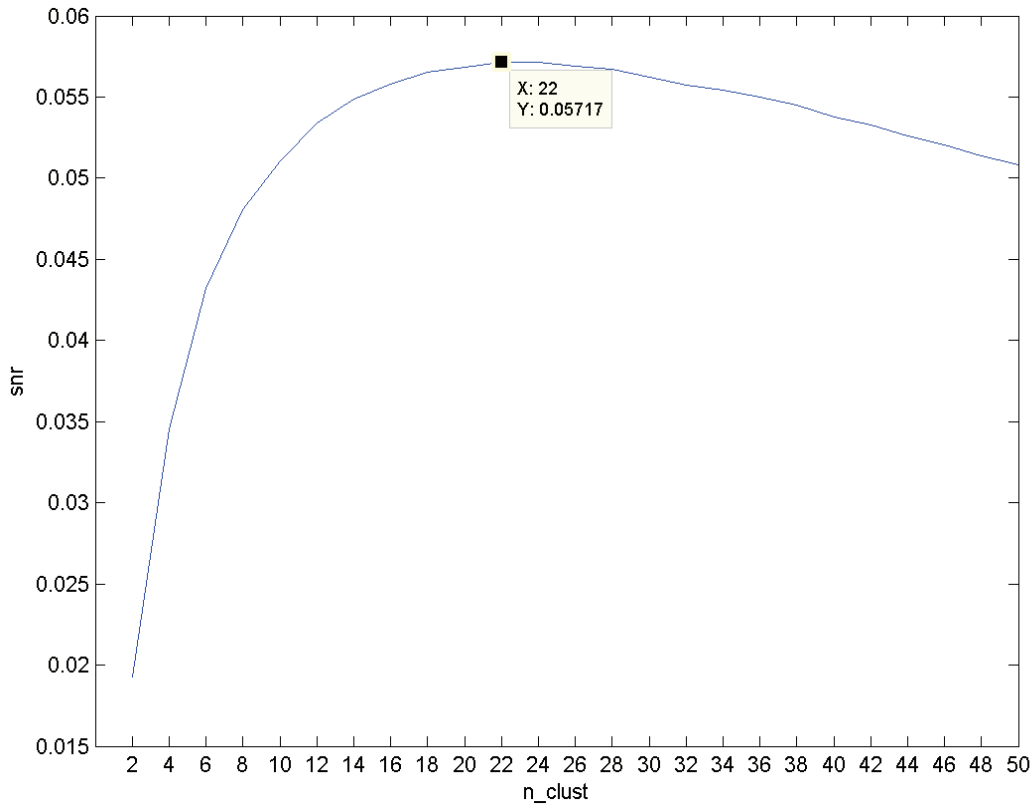


Fig. 3. The estimated SNR for the factorizations ranging from  $n_c=2$  to 50 clusters.

#### 4.3. Significant sample-specific mutations

For both  $n_c=7$  and 22, the expression clusters were highly associated (using Fisher's exact test) with the French-American-British (FAB) AML subtypes, as noticed in previous studies (see Table 2 below). In particular, the clustering perfectly distinguishes the Acute Promyelocytic Leukemia (M3) samples from the rest (cluster 3). FAB types M6 and M7 are too weakly represented in our 163 samples (just 1 and respectively 3 samples) to influence the factorization much.

Table 2. Association of clusters with FAB subtypes ( $n_c=7$ )

FAB subtype	FAB samples	Best associated cluster ( $n_c=7$ )	Cluster samples ( $n_c=7$ )	$\log_2(p)$ ( $n_c=7$ )	Best associated cluster ( $n_c=22$ )	Cluster samples ( $n_c=22$ )	$\log_2(p)$ ( $n_c=22$ )
M0	15	7	33	-8.44	12	23	-24.74
M1	38	5	30	-13.94	14	16	-6.25
M2	39	6	32	-6.06	17	14	-13.18
M3	16	3	16	-41.97	8	16	-41.97
M4	32	1	28	-11.58	22	27	-12.29
M5	17	2	25	-16.36	19	11	-27.13
M6	1	2	25	-2.70	13	14	-3.54
M7	3	7	33	-7.02	13	14	-5.67

The tables '*sample-specific mutations 7 clusters.xls*' and '*sample-specific mutations 22 clusters.xls*' (online at [www.ai.ici.ro/PSB2014](http://www.ai.ici.ro/PSB2014)) list the sample-specific mutations (in descending order of their significance  $MM_{ms}$  for each sample  $s$ ) obtained by our approach based on joint clustering of expression and mutation interactor data.

To estimate the concordance of the three mutation significance lists ('*sample mutations clique cover.xls*' from section 3, as well the two tables mentioned in this section), we have computed the average overlap of the top 5 mutations in each sample for all pairs of lists and depicted the results in Figure 4.

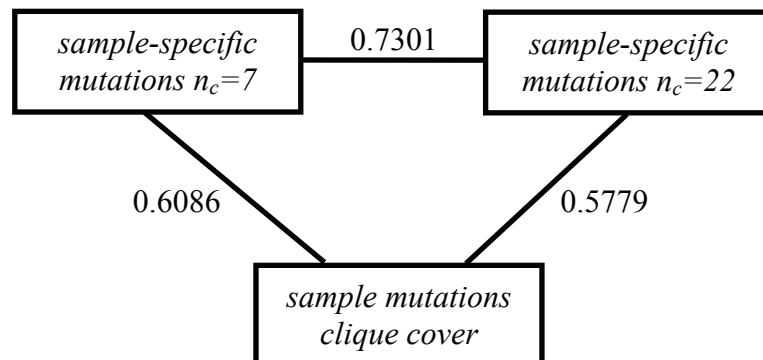


Fig 4. Average overlap between sample-specific mutation significance lists discussed in this paper.

Note that the two sample-specific mutation lists overlap best, as expected, and that the list for  $n_c=7$  is slightly closer to ‘*sample mutations clique cover*’ than the  $n_c=22$  list. The overlap is typically lower for the samples with large numbers of mutations (also as expected).

Careful inspection of the 3 mutation lists shows that we have been able to pick up at least a large fraction (if not most) of the mutations with a causal role in the disease. Virtually all mutations (such as those in NPM1, FLT3, TP53, DNMT3A, etc.), as well as all gene fusions (PML-RARA, MYH11-CBFB, RUNX1-RUNX1T1, etc.) with a known involvement in AML are at the top of the lists of the samples harboring them.

Besides these obvious true positives however, it is difficult to objectively compute accuracy figures for the lists, given the fact that rare individual patient mutations are still largely *terra incognita*. Still, we selected from ‘*sample-specific mutations 7 clusters.xls*’ the sample entries whose top first mutation is *not* among the known AML mutations – we obtained 18 such samples (out of a total of 163), which we list in ‘*NOT EXPLAINED sample-specific mutations.xls*’ (also online). A careful inspection of these samples places them in one of the following 3 categories:

1. Samples with very few detected mutations.
2. A known mutation/fusion is not at the top, but close to it (having significance coefficients close to the top ones).
3. Samples with very many mutations for which known mutations/fusions are far from the top.

Subcategory 3.1. The first few top entries may contain generalized mutations mentioned in the literature in connection with leukemia.

Obviously, our algorithm does not err too much in categories 1 and 2. Only category 3 (including a few outlier samples with very many mutations) could in principle be improved on – we suspect that they misbehave because those samples do not fit very well in any expression cluster, due to the large numbers of defects accumulated. Table 3 below shows the corresponding samples and their category assignments.

Table 3. Samples with rare mutations

Category	Sample	Comments
1	2946	Only two mutations of unknown role.
1	2995	Only 3 mutations. DDX41(mut) observed by others mutated in AML.
1	3000	Only 3 mutations of unknown role.
1	3008	Only two mutations. Possible role of KAT2B.
2	2832	MLL-MLLT10 fusion close to top.
2	2855	MLLT10-PICALM fusion close to top.
2	2874	IDH2(mut) close to top.
2	2911	MLL-ELL fusion, with MLL significance $3.6 \cdot 10^{-4}$ (close to top significance $5.9 \cdot 10^{-4}$ ).
2	2940	MLL3(mut) close to top.
2	2955	DNMT3A(mut) with significance $10^{-3}$ (top $1.2 \cdot 10^{-3}$ ).



2	3005	MLL-MLLT10 fusion close to top.
3/3.1(?)	2817	CBFB(mut), EZH2(mut), BCR-ABL fusion are far from the top, but LUC7L2(del) at the top (LUC7L2 mutations mentioned in AML).
3.1	2849	MLLT10-PICALM fusion far from top, but at the top, KDM3B (a H3K9 demethylase) is a tumor suppressor linked to leukemia.
3	2882	U2AF1(mut) far from top (significance $1.1 \cdot 10^{-3}$ , top $2.3 \cdot 10^{-3}$ ).
3	2917,2929	KRAS(mut), SETBP1(mut) far from top.
3/3.1(?)	2920	NF1(mut) far from top, but LUC7L2(del) at the top.
3/3.1(?)	2939	MTOR-CDH1 fusion far from top, but LUC7L2(del) at the top.

## 5. Conclusions

AML, like other cancer types is a heterogeneous disease. But even with multi-genomic data available (related to gene expression, mutations, copy number changes, etc.), finding well-defined *sub-classifications with prognostic and therapeutic value* is still an elusive objective for many cancers (although partial encouraging results have been obtained). This is probably due to the complexity of the biological processes that are perturbed in the disease and which can be affected by a large number of (generalized) mutations. Some of these mutations have a high enough incidence for us to be sure of their causal role in the disease, but many (if not the majority) of the causal genomic events are rare and patient-specific.

In this paper we have shown that we can exploit protein-to-protein interaction data to relate these possibly rare mutations to one another, thereby enabling a better automated detection of the driver mutations in each individual patient. An original feature of our approach is the use of pp interactors of the mutations to enable clustering and especially the back-reconstruction of the significant mutations from the interactor clusters.

HotNet [19], used in the original TCGA publication [8], identified only 4 significantly mutated subnetworks (which are similar to some of our maximal mutation cliques). However, HotNet does not take into consideration the gene expression data, whereas we expect *driver mutations affecting the same pathway* to produce *similar expression changes*.

Future work will address the much more difficult problem of finding *clinically* useful prognostic markers. This will likely require looking at the precise mutations and possibly larger sample sizes, as different mutations in the same pathway or even in the same gene can have significantly different clinical outcomes.

## 6. Acknowledgments

We are deeply grateful for the invaluable resources put together and made publicly available by the TCGA project. We would also like to thank Andrei Halanay, Daniel Coriu and Jardan Dumitru for discussions, as well as the reviewers for their comments, which helped improve the paper.

## References

1. McLendon, Roger, et al. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature* 455.7216 (2008): 1061-1068.
2. Noushmehr, Houtan, et al. "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma." *Cancer cell* 17.5 (2010): 510-522.
3. Verhaak, Roel GW, et al. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer cell* 17.1 (2010): 98-110.
4. Bell, D., et al. "Integrated genomic analyses of ovarian carcinoma." *Nature* 474.7353(2011):609-615.
5. Koboldt Daniel C. et al. "Comprehensive molecular portraits of human breast tumours." *Nature* 490.7418 (2012):61-70.
6. Hammerman, Peter S., et al. "Comprehensive genomic characterization of squamous cell lung cancers." *Nature* 489.7417 (2012): 519-525.
7. Muzny, Donna M., et al. "Comprehensive molecular characterization of human colon and rectal cancer." *Nature* 487 (2012): 330-337.
8. Ley, T. J., et al. "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia." *N. Engl. J. Med* 368.22 (2013): 2059-2074.
9. Gilliland, D. Gary. "Hematologic malignancies." *Current opinion in hematology* 8.4 (2001): 189-191.
10. Badea, Liviu. "Multirelational Consensus Clustering with Nonnegative Decompositions." *Proc. of the 20th European Conference on Artificial Intelligence ECAI 2012* (2012): 97-102.
11. Badea, Liviu. "Unsupervised analysis of leukemia and normal hematopoiesis by joint clustering of gene expression data." *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, (2012): 338-343.
12. Kim, Philip M., and Bruce Tidor. "Subsystem identification through dimensionality reduction of large-scale gene expression data." *Genome research* 13.7 (2003): 1706-1718.
13. Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791.
14. Seung, D., and L. Lee. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems* 13 (2001): 556-562.
15. Badea, Liviu. "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization." In *Pacific Symposium on Biocomputing*, vol. 290, pp. 279-290. 2008.
16. Moon, J.W., Moser L. "On cliques in graphs." *Israel journal of Mathematics* 3.1 (1965): 23-28.
17. Bron, Coen, and Joep Kerbosch. "Algorithm 457: finding all cliques of an undirected graph." *Communications of the ACM* 16.9 (1973): 575-577.
18. Eifert, C., Powers, R.S. "From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets." *Nature Reviews Cancer* 12, 572-578 (2012).
19. Vandin, F et al. "Algorithms for detecting significantly mutated pathways in cancer." *Journal of Computational Biology* 18.3 (2011): 507-522.

# THE STREAM ALGORITHM: COMPUTATIONALLY EFFICIENT RIDGE-REGRESSION VIA BAYESIAN MODEL AVERAGING, AND APPLICATIONS TO PHARMACOGENOMIC PREDICTION OF CANCER CELL LINE SENSITIVITY

ELIAS CHAIBUB NETO\*, IN SOCK JANG, STEPHEN H. FRIEND, ADAM A. MARGOLIN\*

*Sage Bionetworks, 1100 Fairview Avenue North, Seattle, Washington 98109, USA*

*E-mail: elias.chaibub.neto@sagebase.org*

*E-mail: in.sock.jang@sagebase.org*

*E-mail: friend@sagebase.org*

*E-mail: margolin@sagebase.org*

*www.sagebase.org*

Computational efficiency is important for learning algorithms operating in the “large  $p$ , small  $n$ ” setting. In computational biology, the analysis of data sets containing tens of thousands of features (“large  $p$ ”), but only a few hundred samples (“small  $n$ ”), is nowadays routine, and regularized regression approaches such as ridge-regression, lasso, and elastic-net are popular choices. In this paper we propose a novel and highly efficient Bayesian inference method for fitting ridge-regression. Our method is fully analytical, and bypasses the need for expensive tuning parameter optimization, via cross-validation, by employing Bayesian model averaging over the grid of tuning parameters. Additional computational efficiency is achieved by adopting the singular value decomposition reparametrization of the ridge-regression model, replacing computationally expensive inversions of large  $p \times p$  matrices by efficient inversions of small and diagonal  $n \times n$  matrices. We show in simulation studies and in the analysis of two large cancer cell line data panels that our algorithm achieves slightly better predictive performance than cross-validated ridge-regression while requiring only a fraction of the computation time. Furthermore, in comparisons based on the cell line data sets, our algorithm systematically out-performs the lasso in both predictive performance and computation time, and shows equivalent predictive performance, but considerably smaller computation time, than the elastic-net.

*Keywords:* ridge-regression, Bayesian model averaging, predictive modeling, machine learning, cancer cell lines, pharmacogenomic screens

## 1. Introduction

Analysis of high-throughput “omics” data sets to infer molecular predictors of cancer phenotypes is a common type of problem in modern computational biology research. The use of genomic features such as from gene expression, copy number variation, and sequence data, in the predictive modeling of anticancer drug response is a particularly relevant example, which holds the potential to speed up the emergence of “personalized” cancer therapies.<sup>1,5</sup> A common theme of such high-dimensional prediction problems is that the number of genomic features,  $p$ , is usually much larger than the number of available samples,  $n$ , and regularized regression approaches such ridge-regression,<sup>2</sup> lasso,<sup>3</sup> and elastic-net<sup>4</sup> are popular methodological choices in this context.<sup>1,5</sup> Computational efficiency is of key importance for any learning algorithm operating in this “large  $p$ , small  $n$ ” setting; a method that improves computational efficiency without sacrificing prediction accuracy could enable such models to be readily applied across

---

\*corresponding authors

a large number of phenotype prediction problems, such as inferring genomic predictors for large panels of anticancer compounds.

In this paper we propose a novel Bayesian formulation of ridge-regression, which executes in a fraction of the time required by the most efficient current implementations of regularized regression methods, while achieving comparable prediction accuracy. We refer to our approach as Stream (Scalable-Time Ridge Estimator by Averaging of Models). First, Stream replaces cross-validation by Bayesian model averaging<sup>6</sup> (BMA) over the grid of tuning parameters. For each tuning parameter in the grid, we interpret the corresponding ridge-regression fit as a distinct model, and average all models, weighted by how well each model fits the data. Second, it replaces the computation of large  $p \times p$  matrix inversions by efficient inversions of small and diagonal  $n \times n$  matrices derived from the singular value decomposition<sup>7</sup> (SVD) of the feature matrix. Note that the use of SVD re-parameterization is a practice to improve the computational efficiency of ridge-regression model fit.<sup>8</sup>

We point out that both improvements are allowed by the analytical tractability of the Bayesian hierarchical formulation of ridge-regression, where the marginal posterior distribution of the regression coefficients and the prior predictive distribution of the data are readily available, leading to a fully analytical expression for the BMA estimate of the regression coefficients. Furthermore, the quantities that need to be evaluated, namely, model specific posterior expectations and marginal likelihoods, can be efficiently computed under the SVD re-parametrization.

The rest of the paper is organized as follows. In Section 2.1 we present the Stream algorithm, and, in Section 2.2, we present its re-parametrization in terms of the singular value decomposition of the feature data matrix. Section 3.1 presents a simulation study comparing the predictive performance and computation time of Stream against the standard cross-validated ridge-regression model. Section 3.2 presents real data illustrations using two compound screening data sets performed on large panels of cancer cell lines. Finally, in Section 4 we discuss our results, and point out strengths and weaknesses of our proposed algorithm.

## 2. Statistical model

In the next subsections we present the Stream-regression model and its re-parametrization in terms of the SVD of the feature data matrix. First, we introduce some notation. Throughout the text we consider the regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{y}$  represents the  $n \times 1$  vector of responses,  $\mathbf{X}$  corresponds to the  $n \times p$  matrix of features,  $\boldsymbol{\beta}$  corresponds to the  $p \times 1$  vector of regression coefficients, and  $\boldsymbol{\epsilon}$  represents a  $n \times 1$  vector of independent and identically distributed gaussian error terms with expectation 0 and precision  $\tau$ . The notation  $\text{Ga}(a, b)$  represents a gamma distribution with shape and rate parameters  $a$  and  $b$ , respectively;  $\text{U}(a, b)$  stands for the uniform distribution on the interval  $[a, b]$ ;  $\text{DU}(a, b)$  represents the discrete uniform distribution with support in the range  $\{a, \dots, b\}$ ;  $\text{Ber}(\phi)$  corresponds to the Bernoulli distribution with success probability  $\phi$ ;  $\text{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  represents a  $k$ -dimensional multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ ; and  $\text{St}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$  corresponds to a  $k$ -dimensional multivariate t-distribution with mean vector  $\boldsymbol{\mu}$ , scale matrix  $\boldsymbol{\Sigma}$ , and  $v$  degrees of freedom. We represent the  $k$ -dimensional identity matrix by  $\mathbf{I}_k$ , the indicator function assuming values 0

or 1 by  $\mathbb{1}$ , and the determinant of a matrix  $\mathbf{A}$  by  $\det(\mathbf{A})$ .

### 2.1. Stream regression model

Consider the Bayesian hierarchical form representation of ridge-regression (a special case of the Bayesian formulation for the linear regression model with a normal-gamma prior<sup>9</sup>):

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \tau \sim N_n(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n),$$

$$\boldsymbol{\beta} \mid \tau, \lambda \sim N_p(\mathbf{0}, \tau^{-1}\lambda^{-1}\mathbf{I}_p),$$

$$\tau \sim \text{Ga}(a_\tau, b_\tau),$$

where the precision parameter  $\lambda$  plays the role of the tuning parameter in ridge-regression. Under this analytically tractable model we have that the marginal posterior distribution of the regression coefficients is

$$\pi(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \text{St}_p \left( \boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \frac{2b_\tau + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t \mathbf{y}}{2a_\tau + n} (\mathbf{X}^t \mathbf{X} + \lambda \mathbf{I}_p)^{-1}, 2a_\tau + n \right),$$

where the expectation,  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^t \mathbf{y}$ , corresponds to the usual (frequentist) ridge-regression estimator, and the prior predictive distribution is given by

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{X}) &= \int_{\tau} \int_{\boldsymbol{\beta}} N_n(\mathbf{y}; \mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n) N_p(\boldsymbol{\beta}; \mathbf{0}, \lambda^{-1}\tau^{-1}\mathbf{I}_p) \text{Ga}(\tau; a_\tau, b_\tau) d\boldsymbol{\beta} d\tau \\ &= \text{St}_n \left( \mathbf{y}; \mathbf{0}, \frac{b_\tau}{a_\tau} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^t)^{-1}, 2a_\tau \right), \end{aligned} \quad (1)$$

Now let  $\lambda_k$ ,  $k = 1, \dots, K$  represent the grid of ridge-regression tuning parameters, and let  $\mathcal{M}_k$  represent a ridge-regression model that uses  $\lambda = \lambda_k$ . The BMA estimate of  $\boldsymbol{\beta}$  is then

$$E[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}] = \sum_{k=1}^K E[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \mathcal{M}_k] pr(\mathcal{M}_k \mid \mathbf{X}, \mathbf{y}), \quad (2)$$

where

$$E[\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \mathcal{M}_k] = (\mathbf{X}^t \mathbf{X} + \lambda_k \mathbf{I}_p)^{-1} \mathbf{X}^t \mathbf{y}$$

and the posterior distribution of model  $\mathcal{M}_k$ , given the data, is computed as

$$pr(\mathcal{M}_k \mid \mathbf{X}, \mathbf{y}) = \frac{f(\mathbf{y} \mid \mathbf{X}, \mathcal{M}_k) pr(\mathcal{M}_k)}{\sum_{k=1}^K f(\mathbf{y} \mid \mathbf{X}, \mathcal{M}_k) pr(\mathcal{M}_k)}.$$

Here,  $f(\mathbf{y} \mid \mathbf{X}, \mathcal{M}_k)$  corresponds to the prior predictive distribution in (1) with  $\lambda$  replaced by  $\lambda_k$ , and we adopt a discrete uniform prior for the models, so that  $pr(\mathcal{M}_k) = K^{-1}$ ,  $k = 1, \dots, K$ .

In regression based predictive modeling, one is generally interested in making a prediction,  $\hat{\mathbf{y}} = \mathbf{X}_{test} \hat{\boldsymbol{\beta}}_{train}$ , of the response vector  $\mathbf{y}_{test}$ , where  $\mathbf{X}_{test}$  represents the feature data on the testing set, and  $\hat{\boldsymbol{\beta}}_{train}$  represents the regression coefficients estimate learned from the training set. In our Bayesian model, we are interested on the the expectation of the response's posterior predictive distribution,

$$E[\mathbf{y}_{test} \mid \mathbf{X}_{test}, \mathbf{X}_{train}, \mathbf{y}_{train}] = \mathbf{X}_{test} E[\boldsymbol{\beta} \mid \mathbf{X}_{train}, \mathbf{y}_{train}],$$

where  $E[\boldsymbol{\beta} \mid \mathbf{X}_{train}, \mathbf{y}_{train}]$  is given by equation (2).

## 2.2. SVD re-parametrization

Consider the SVD of the  $n \times p$  feature data matrix  $\mathbf{X}$  of rank  $n$ . One possible representation of  $\mathbf{X}$  is given by  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$ , where  $\mathbf{U}$  is a  $n \times n$  orthogonal matrix of left singular vectors;  $\mathbf{D}$  is a  $n \times n$  diagonal matrix of singular values  $d_j$ ; and  $\mathbf{V}$  is a  $p \times n$  matrix of right singular vectors. An alternative representation is  $\mathbf{X} = \mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^t$  where  $\mathbf{U}_*$  is a  $n \times p$  matrix obtained by augmenting  $\mathbf{U}$  with  $p - n$  extra columns of zeros,  $\mathbf{U}_* = (\mathbf{U}, \mathbf{0})$ ;  $\mathbf{D}_*$  is a  $p \times p$  diagonal matrix with the first  $n$  diagonal entries given by the singular values and the remaining  $p - n$  diagonal entries set to zero; and  $\mathbf{V}_*$  is a  $p \times p$  orthogonal matrix obtained by augmenting  $\mathbf{V}$  with  $p - n$  additional right singular vectors. Exploring these re-parametrizations we can, after some algebra, re-express  $\hat{\beta}$  in the computationally more efficient form,

$$E[\beta \mid \mathbf{X}, \mathbf{y}, \mathcal{M}_k] = \mathbf{V}_* (\mathbf{D}_*^2 + \lambda_k \mathbf{I}_p)^{-1} \mathbf{D}_* \mathbf{U}_*^t \mathbf{y} = \mathbf{V} (\mathbf{D}^2 + \lambda_k \mathbf{I}_n)^{-1} \mathbf{D} \mathbf{U}^t \mathbf{y} .$$

In addition to the efficient computation of  $\hat{\beta}$ , we can also explore the SVD reparametrization for efficient computation of the prior predictive distribution, which involves two computationally expensive steps; namely, evaluation of the quadratic form  $\mathbf{y}^t (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X} + \lambda_k \mathbf{I}_p)^{-1} \mathbf{X}^t) \mathbf{y}$ , and of  $\det((\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X} + \lambda_k \mathbf{I}_p)^{-1} \mathbf{X}^t)^{-1})$ . Starting with the quadratic form, observe that

$$\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X} + \lambda_k \mathbf{I}_p)^{-1} \mathbf{X}^t = \mathbf{I}_n - \mathbf{U}_* \mathbf{D}_* (\mathbf{D}_*^2 + \lambda_k \mathbf{I}_p)^{-1} \mathbf{D}_* \mathbf{U}_*^t = \mathbf{I}_n - \mathbf{U} (\mathbf{I}_n + \lambda_k \mathbf{D}^{-2})^{-1} \mathbf{U}^t ,$$

so that we replace a  $p \times p$  matrix inversion by a  $n \times n$  diagonal matrix inversion in the computation of the quadratic form. Next, consider the determinant. From the application of the Woodbury matrix inversion formula<sup>10</sup> we have that

$$\mathbf{I}_n - \mathbf{U} (\mathbf{I}_n + \lambda_k \mathbf{D}^{-2})^{-1} \mathbf{U}^t = (\mathbf{I}_n + \lambda_k^{-1} \mathbf{U} \mathbf{D}^2 \mathbf{U}^t)^{-1} ,$$

and from standard properties of the determinant and the orthogonality of the  $\mathbf{U}$  matrix we have that

$$\det((\mathbf{I}_n - \mathbf{X}(\mathbf{X}^t \mathbf{X} + \lambda_k \mathbf{I}_p)^{-1} \mathbf{X}^t)^{-1}) = \det(\mathbf{I}_n + \lambda_k^{-1} \mathbf{U} \mathbf{D}^2 \mathbf{U}^t) = \prod_{j=1}^n \left(1 + \frac{d_j^2}{\lambda_k}\right) .$$

Hence, the prior predictive distribution can be efficiently computed as

$$f(\mathbf{y} \mid \mathbf{U}, \mathbf{D}, \mathcal{M}_k) = C \left(1 + \frac{\mathbf{y}^t (\mathbf{I}_n - \mathbf{U} (\mathbf{I}_n + \lambda_k \mathbf{D}^{-2})^{-1} \mathbf{U}^t) \mathbf{y}}{2 b_\tau}\right)^{-\frac{2 a_\tau + n}{2}}$$

with the normalization constant,  $C$ , given by

$$C = \frac{\Gamma\left(\frac{2 a_\tau + n}{2}\right)}{\Gamma\left(\frac{2 a_\tau}{2}\right) (2 a_\tau \pi)^{\frac{n}{2}}} \left(\frac{b_\tau}{a_\tau}\right)^{-\frac{n}{2}} \left(\prod_{j=1}^n \left(1 + \frac{d_j^2}{\lambda_k}\right)\right)^{-\frac{1}{2}} .$$

### 3. Illustrations

Before we present our simulation study and real data illustrations, we provide a few model fitting details relevant to the next subsections. Throughout this paper we evaluate predictive performance using the RMSE statistic,  $\sqrt{(\mathbf{y}_{test} - \hat{\mathbf{y}})^t(\mathbf{y}_{test} - \hat{\mathbf{y}})/n_{test}}$ , where  $\hat{\mathbf{y}} = \mathbf{X}_{test}\hat{\boldsymbol{\beta}}_{train}$ . For ridge-regression, lasso, and elastic-net, we adopted 10 fold cross validation. We adopted a data-driven approach, described in detail in the appendix, for the determination of the tuning parameter grid for ridge-regression and Stream. Each simulated or real data set used a different grid, composed of  $K = 100$  values. For each data set we used the same grid in the ridge-regression and Stream model fits. For the lasso and elastic-net algorithms, we adopted the tuning parameter grids generated by default by the `glmnet` R package.<sup>11</sup> Both response and feature data are scaled prior to analysis. In both simulation studies and real data analysis illustrations we tested whether the difference in RMSE between two methods is statistically significant using the Wilcoxon paired-sample test.<sup>12</sup>

#### 3.1. Simulation study

We performed a simulation study illustrating how Stream achieves slightly better predictive performance than ridge-regression (when we adopt non-informative priors for the residual precision parameter  $\tau$ ), while requiring only a fraction of the computation time.

In order to evaluate the method's performance under widely heterogenous conditions, we simulated 5,000 distinct data sets, each one generated with a unique and random combination of sample size ( $n$ ), number of features ( $p$ ), model sparsity ( $\phi$ ), residual noise ( $\sigma$ ), and strength of feature correlation ( $\rho$ ), sampled according to:  $n \sim \text{DU}(100, 500)$ ;  $p \sim \text{DU}(501, 10000)$ ;  $\phi \sim \text{U}(0.1, 0.9)$ ,  $\sigma \sim \text{U}(0.1, 5)$ ; and  $\rho \sim \text{U}(0.1, 0.9)$ .

Each simulated data set was generated as follows: (i) we first draw a single value of  $n$ ,  $p$ ,  $\phi$ ,  $\sigma$ , and  $\rho$ , from their respective uniform distributions; (ii) given the sampled values of  $n$ ,  $p$ , and  $\rho$ , we simulate the feature data matrix,  $\mathbf{X}_{n \times p} = (\mathbf{X}_{n \times p_1}, \dots, \mathbf{X}_{n \times p_L})$ , as  $L$  separate matrices,  $\mathbf{X}_{n \times p_l}$ , generated independently from  $N_{p_l}(\mathbf{0}, \boldsymbol{\Sigma}_l)$  distributions, where  $\Sigma_{ij,l} = 1$ , for  $i = j$ , and  $\Sigma_{ij,l} = \rho^{|i-j|}$ , for  $i \neq j$ . The number of features,  $p_l$ , in each of these matrices were randomly chosen between 20 and 100 under the constraint that  $p = \sum_{k=1}^L p_l$ ; (iii) given the sampled values of  $p$  and  $\phi$ , we computed each regression coefficient,  $\beta_j$ ,  $j = 1, \dots, p$ , as  $\beta_j = \beta_j^* \mathbb{1}_{\beta_j}$ , where  $\beta_j^* \sim N(0, 1)$ , and  $\mathbb{1}_{\beta_j} \sim \text{Ber}(\phi)$  (note that, by defining  $\beta_j$  as above, we have that, on average,  $\phi p$  regression coefficients will be non-zero); and (iv) given the sampled value of  $\sigma$  and the computed feature matrix and regression coefficients vector, we computed the response vector as  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$ , where  $\boldsymbol{\epsilon}$  is a vector of standard gaussian error variables. We note that for each simulated data set we actually generated  $2n$  data samples, and used the first  $n$  samples as the training set, and the second half as the test set. Figure 1 present the results.

Panel (a) in Figure 1 shows that the predictive performance of Stream is quite similar to ridge-regression when the RMSE values are small, but Stream tends to slightly outperform ridge when RMSE values are larger, as suggested by the increased number of points below the diagonal for RMSE values closer to 1. Application of the Wilcoxon paired-sample test shows that, overall, Stream achieves statistically significant increased performance over ridge (p-value =  $2.501 \times 10^{-5}$ ). We note that the results in Figure 1 were computed using an uninformative

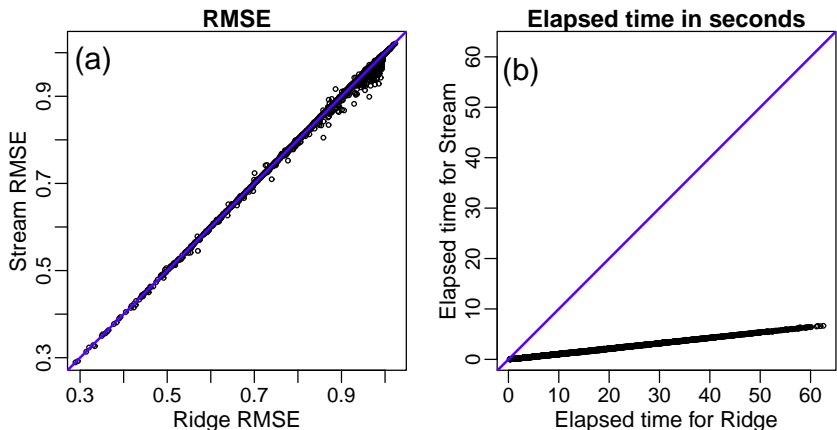


Fig. 1. Simulation results: comparison of Stream and ridge-regression in terms of predictive performance and computation time.

gamma prior distribution (hyper-parameters  $a_\tau = b_\tau = 0.001$ ). As expected, the adoption of informative gamma priors led to decreased predictive accuracy (results not shown).

Panel (b) in Figure 1 shows the comparison of computation times between Stream and ridge-regression. Overall, Stream was approximately 10 times faster than ridge. In general, Stream is approximately  $f$  times faster than ridge-regression, where  $f$  represents the number of cross-validation folds used by ridge.

### 3.2. Cancer cell line panels

In this section we compare the predictive performance and computation time of the Stream, ridge-regression, lasso, and elastic-net algorithms in inferring molecular predictors of compound sensitivity based on the Sanger<sup>5</sup> and CCLE<sup>1</sup> data sets, which contain compound screening data performed on large panels of molecularly characterized cancer cell lines.

In Sanger we have 535 cell lines and a total of 30,765 features comprised of 4 distinct feature data types, including gene expression measurements on 12,024 genes, copy number variation measurements on 18,601 genes, cell line tumor type classifications according to 93 distinct tumor lineages, and mutation profiling on 47 genes. In CCLE we have 411 cell lines and 41,911 features comprised of 5 distinct feature types, including gene expression measurements on 18,897 genes, copy number measurements on 21,217 genes, cell line tumor type classifications on 97 tumor lineages, mutation profiling on 33 genes using the oncomap 3.0 platform,<sup>13</sup> and mutation profiling of 1,667 genes using hybrid capture sequencing. Mutation data was summarized to gene-level binary calls, with 1 representing a somatic mutation observed at any base pair within the gene. Gene expression, copy number, and mutation data were summarized to gene-level features. The Sanger dataset tested 138 compounds and summarized the sensitivity of each cell line based on IC50 values. The CCLE dataset tested 24 compounds and summarized the sensitivity of each cell line based on the area over the dose response curve (where response values at each compound dose are scaled with -100 representing complete growth inhibition and 0 representing no growth inhibition).

In the present analysis we discarded samples and features with missing data, and we filtered



out genomic features with variance smaller than 0.01, and with non-significant correlation with the response ( $p$ -value  $> 0.1$ ). After filtering we obtained, on average,  $5,588.80 \pm 2,046.48$  genomic features in Sanger, and  $12,512.12 \pm 3,345.16$  in CCLE. We evaluated predictive performance by splitting the data into five parts, using four parts as the training set and the left out part as the testing set. In each of the 5 splits, we trained the ridge, lasso, and elastic-net models using 10 fold cross validation and adopted  $a_\tau = b_\tau = 0.01$  for the Stream model. At each split we obtained a prediction vector  $\hat{\mathbf{y}}_j$ ,  $j = 1, \dots, 5$ , and we computed a single RMSE using the concatenated vector of predictions,  $\hat{\mathbf{y}}^t = (\hat{\mathbf{y}}_1^t, \dots, \hat{\mathbf{y}}_5^t)$ , and the full observed response data,  $\mathbf{y}$ , as  $\sqrt{(\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}})/n}$ .

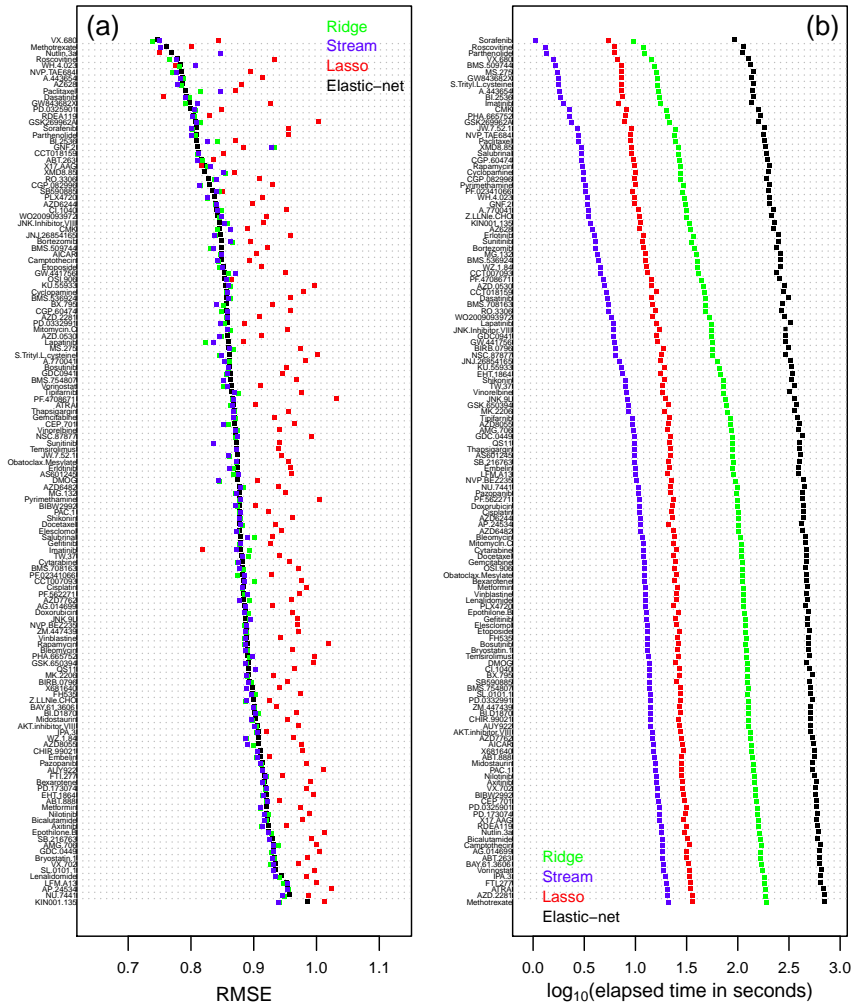


Fig. 2. Predictive performance and computation time for the Sanger cell line panel. Results for each compound.

Figures 2 and 3 depict the results for the Sanger data. Figure 2(a) shows the RMSE scores across the 138 drugs sorted according to the elastic-net RMSE. Overall, Stream seems to perform slightly better than ridge and elastic-net, especially for compounds with high RMSE, consistent with results on the simulated data. Figure 3(a) confirms this result, showing that

the median RMSE of Stream (horizontal blue line) is in fact slightly smaller than those of ridge and elastic-net. Furthermore, application of the Wilcoxon paired-sample test shows that the slight advantage of Stream is statistically significant (p-values equal to 0.004611 and 0.02147 for the comparisons of Stream against elastic-net and ridge, respectively). The lasso performance, on the other hand, is considerably worse than all other methods. Figures 2(b) and 3(b) show considerably smaller computation times for Stream than the other methods. The elastic-net is the most expensive, followed by ridge and then the lasso. Note that the results are shown in the  $\log_{10}$  scale. In the original scale, Stream was, on average,  $2.46 \pm 0.74$ ,  $9.06 \pm 0.09$ , and  $47.30 \pm 14.23$  times faster than the lasso, ridge, and elastic-net, respectively.

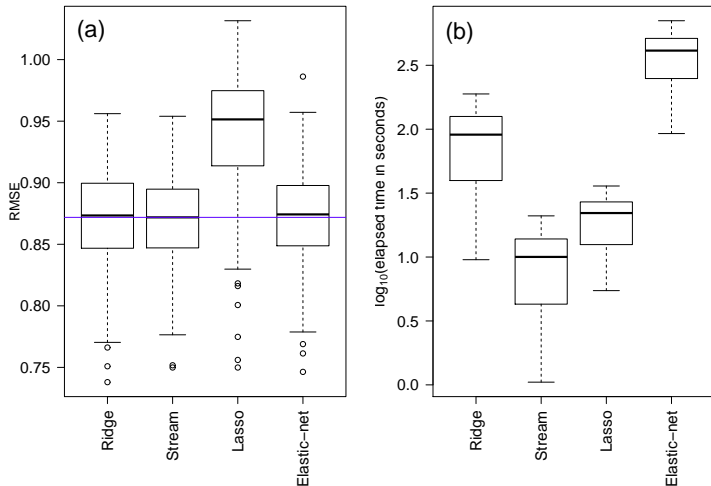


Fig. 3. Predictive performance and computation time for the Sanger cell line panel. Overall results.

Figure 4 depicts the results for the CCLE data. Panels (a) and (c) show that Stream performs slightly better than ridge and similarly to elastic-net, although, in both cases, the differences are not statistically significant (p-values equal to 0.16 and 0.1074, respectively). Once again, the lasso performance is considerably worse than the other methods. Panels (b) and (d) show, again, smaller computation times for Stream than the other methods. Stream was, on average,  $1.9 \pm 0.2$ ,  $9.16 \pm 0.08$ , and  $38.04 \pm 4.54$  times faster than the lasso, ridge, and elastic-net, respectively.

A particularly attractive feature of Stream is the ability to perform feature selection by estimating the posterior distribution of regression coefficients. In the context of compound sensitivity prediction, previous studies have demonstrated that such feature selection may provide the basis for identifying functional biomarkers underlying compound sensitivity or resistance.<sup>1,5</sup> We compiled a list of known biomarkers of sensitivity (gold standards) for 8 compounds represented in both the Sanger and CCLE panels (first column of Table 1) and evaluated the rank of each biomarker (based on the absolute value of the regression coefficients) in the model generated by Stream, ridge, lasso, and elastic-net for the corresponding compound.

Table 1 present the results. Overall, the relative performance of all four methods tended to be similar in the sense that the gold standard biomarkers tended to be either well ranked

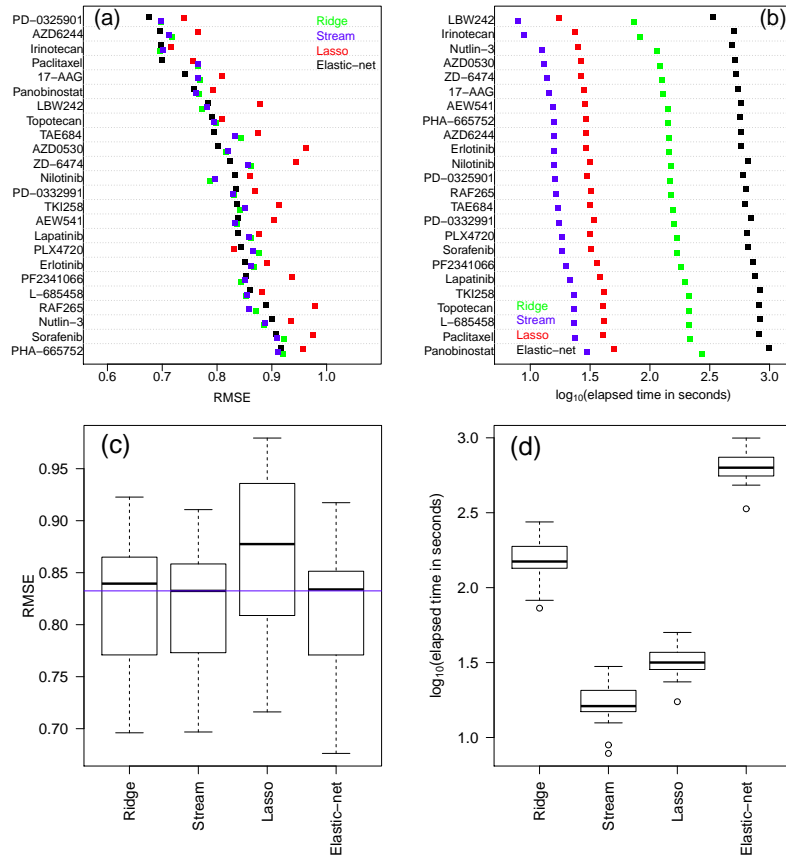


Fig. 4. Predictive performance and computation time for the CCLE cell line panel. Results for each compound (top panels) and overall (bottom panels).

Table 1. Genomic feature ranks. Entries present the rank position followed by the feature type: copy number variation (C), expression (E), mutation from the oncomap platform ( $M_o$ ), and mutation from hybrid capture sequencing ( $M_h$ ). Missing entries (-) represent features that had ranks higher than 1,000, were not represented in the data, or were filtered-out from the analysis.

CCLE						Sanger			
Drug	Gold	Stream	Ridge	Lasso	Elastic-net	Stream	Ridge	Lasso	Elastic-net
17-AGG	NQO1	2-E	2-E	1-E	2-E	1-E	1-E	1-E	1-E
AZD-0530	EGFR	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	-	-	-	-
Erlotinib	EGFR	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	-	-	863-E	-
		10-M <sub>h</sub>	9-M <sub>h</sub>	73-M <sub>h</sub>	38-M <sub>h</sub>	-	-	-	-
		36-C	39-C	347-E	258-C	-	-	-	-
Lapatinib	ERBB2	8-E	8-E	1-E	5-E	6-E	2-E	1-E	2-E
		1-C	1-C	2-C	2-C	-	-	-	-
PD-325901	BRAF	94-M <sub>o</sub>	67-M <sub>o</sub>	191-M <sub>o</sub>	235-M <sub>o</sub>	3-M <sub>o</sub>	1-M <sub>o</sub>	2-M <sub>o</sub>	1-M <sub>o</sub>
		2-M <sub>h</sub>	2-M <sub>h</sub>	3-M <sub>h</sub>	3-M <sub>h</sub>	180-E	181-E	252-E	214-E
PF-2341066	MET HGF	9-C	3-C	19-C	45-C	-	-	-	-
		3-E	1-E	1-E	7-E	-	-	-	-
PHA-665752	MET HGF	-	-	-	-	-	-	-	-
		104-E	212-E	540-E	480-E	-	-	-	-
PLX4720	BRAF	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>	1-M <sub>o</sub>
		2-M <sub>h</sub>	2-M <sub>h</sub>	6-M <sub>h</sub>	3-M <sub>h</sub>	-	-	-	-

by all methods or poorly ranked by all methods. For instance, in the CCLE panel, the gold standard features usually showed up among the top ranking features for all methods for most of the drugs. In the Sanger panel, on the other hand, we see that for several of the drugs, all algorithms failed to rank the gold standards among their top ranking features.

#### 4. Discussion

In this paper, we proposed a novel and highly efficient Bayesian version of ridge-regression, which explores Bayesian model averaging and the singular value decomposition re-parametrization for computational efficiency. Our analysis of two large cancer cell line panels showed that the predictive performance of the Stream algorithm tends to be slightly better than ridge-regression in terms of RMSE, suggesting that BMA might be slightly more effective than cross-validation in noisy data sets. This finding was corroborated by our large-scale simulation study, where Stream tended to slightly outperform ridge-regression in the cases where both methods produced high RMSE scores, and showed quite similar performance otherwise. This competitive predictive performance, combined with the considerably higher computational efficiency of the Stream algorithm, suggests that this novel method should be the preferred choice, over standard ridge-regression, in high-dimensional regularized regression applications.

Furthermore, the analysis of cell line panels showed that the predictive performance of the Stream algorithm is also competitive with the elastic-net algorithm, showing slightly better average performance in the Sanger data, and similar performance on the CCLE data. In terms of feature selection ability, Stream showed similar performance to the elastic-net, the current state-of-the-art algorithm employed for the identification of functional biomarkers underlying compound sensitivity or resistance in cancer cell lines.<sup>1</sup> Most importantly, this competitive performance of the Stream algorithm is achieved while requiring only a small fraction of the computation time required by the elastic-net.

Even though, the application of elastic-net, the most time consuming algorithm in this study, is still computationally feasible for the two data sets investigated in this work, we point out that increased computational efficiency opens possibilities for much broader exploration of pharmacogenomic modeling. For instance, in large scale exploration of modeling choices<sup>14,15</sup> such as type of input data (e.g. gene expression, copy number variation, mutation) or method of summarizing sensitivity values (e.g. IC50, ActArea), we need to build models for a large number of possible combinations of input/output data. Additionally, the pharmacogenomic data sets that computational biologists will need to analyze in the near future will only grow bigger, and the use of highly efficient algorithms will likely become a practical necessity in the near future. Efficient algorithms make it easier to infer models for much larger compound screening collections, or even infer models for each of over 10,000 genes from genome wide RNAi screens. The increased efficiency could even allow models to be applied both the whole data set and different subsets of data (e.g. tumors from different tissue types).

In the cancer cell line panels investigated in this work, the lasso performed significantly worse than the other methods. Possible explanations include: (i) that the drug sensitivity phenotype might behave as a complex trait, associated with a large number of predictors,

so that the sparseness assumption made by the lasso is violated in our applications; and (ii) because many features tend to be clustered into highly correlated groups of predictors, the lasso might be effectively selecting one feature randomly from each group, while methods using  $L_2$  regularization can select more than a single feature from a group of highly correlated predictors.

The feasibility of the Stream algorithm is due to the analytical tractability of the Bayesian hierarchical formulation of ridge-regression. Even though Bayesian hierarchical formulations for the lasso and elastic-net models have been proposed in the literature,<sup>16,17</sup> they do not lead to closed analytical forms for the marginal posterior distributions of the regression coefficients and for the marginal likelihoods, so that BMA-based versions of these models are not readily available. The development of model averaging approaches for these methods represents an interesting topic for future research.

We note that, compared to frequentist implementations of penalized regression models, the Bayesian formulation provides several advantages and opportunities for future extensions. For instance, Bayesian approaches provide valid quantifications of the uncertainty associated with the estimates of regression coefficients in the form of probability intervals, whereas even the estimation of standard errors associated with the frequentist versions of penalized regression models is a non-trivial and often problematic task.<sup>17</sup> Furthermore, Bayesian approaches represent a natural framework for the incorporation of additional sources of prior information, such as pathway-based relationships between genes, or prior knowledge of the functional importance of a given gene. Such extensions are topics of active research.

In summary, Stream provides a Bayesian ridge-regression framework with significantly increased computational efficiency without a trade-off of prediction accuracy or feature selection ability. Thus we believe that Stream advances current state-of-the art approaches for inferring molecular predictors of compound sensitivity, with natural extensions to other phenotype prediction problems or general predictive modeling applications in the “large p, small n” setting.

## 5. Availability of code and data

We implemented the Stream algorithm in R,<sup>18</sup> and the source code is available in GitHub (<https://gist.github.com/echaibub/6117763>). The data and code necessary to reproduce the simulation study and analysis of the cancer cell line panels presented in this paper are available in Synapse ([www.synapse.org](http://www.synapse.org)) under the Stream project (<https://www.synapse.org/#!/Synapse:syn2010337>).

## 6. Acknowledgements

This work was funded by NIH/NCI grant 5U54CA149237.

## Appendix A. Computation of the tuning parameter grid

In this section we describe the rationale behind the automatic/data-driven determination of the tuning parameter grid for ridge-regression. It is a simple adaptation of the approach adopted in the `glmnet` R package<sup>11</sup> for the default determination of the  $\lambda$  grid in the lasso

and elastic-net algorithms. The basic idea is to: (i) determine  $\lambda_{\max}$ , as the  $\lambda$  value such that the largest regression coefficient is equal in absolute value to a certain small constant  $\kappa$ ; (ii) determine the smallest  $\lambda$  value in the grid as  $\lambda_{\min} = \epsilon \lambda_{\max}$ , where  $\epsilon$  is another small constant; and (iii) determine the  $\lambda$  grid as a sequence of  $K$  values of  $\lambda$  decreasing from  $\lambda_{\max}$  to  $\lambda_{\min}$  on the log scale. Explicitly, we set the lambda grid as follows: (a) create a decreasing sequence of  $K$  equally spaced values in the interval  $[\log(\lambda_{\max}), \log(\lambda_{\min})]$ ; and (b) take the exponential of each of value in the sequence.

Next, we describe the derivation of  $\lambda_{\max}$ . Considering the singular value decomposition of  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$ , we can re-express the ridge estimator as  $\hat{\boldsymbol{\beta}} = \mathbf{V}(\mathbf{D}^2 + \lambda \mathbf{I}_n)^{-1} \mathbf{D}\mathbf{U}^t \mathbf{y}$ , or

$$\hat{\beta}_j = \frac{d_j}{d_j^2 + \lambda} \mathbf{V}_j \mathbf{U}^t \mathbf{y} ,$$

for  $j = 1, \dots, n$  (and zero for  $i = n+1, \dots, p$ ) where  $\mathbf{V}_j$  represents the  $j$ th row of  $\mathbf{V}$ . Our goal then is to find  $\lambda$  such that  $\max_j(|\hat{\beta}_j|) = \kappa$ . Since

$$|\hat{\beta}_j| = \frac{d_j}{d_j^2 + \lambda} |\mathbf{V}_j \mathbf{U}^t \mathbf{y}| \leq \frac{d_j}{\lambda} |\mathbf{V}_j \mathbf{U}^t \mathbf{y}|$$

for all  $d_j$ , it follows that

$$\kappa = \max_j \left( \frac{d_j}{d_j^2 + \lambda} |\mathbf{V}_j \mathbf{U}^t \mathbf{y}| \right) \leq \frac{1}{\lambda} \max_j (d_j |\mathbf{V}_j \mathbf{U}^t \mathbf{y}|)$$

so that  $\lambda \leq \max_j (d_j |\mathbf{V}_j \mathbf{U}^t \mathbf{y}|) / \kappa$  and we take  $\lambda_{\max} = \max_j (d_j |\mathbf{V}_j \mathbf{U}^t \mathbf{y}|) / \kappa$ . In our simulations and real data analysis we adopted  $\kappa = 10^{-3}$ ,  $\epsilon = 10^{-6}$ , and  $K = 100$ .

## References

1. J. Barretina, et al, *Nature* **483**, 603-607 (2012).
2. A. E. Hoerl, R. W. Kennard, *Technometrics* **42**, 80-86 (1970).
3. R. Tibshirani, *Journal of the Royal Statistical Association, Series B* **58**, 267-288 (1996).
4. H. Zou, T. Hastie, *Journal of the Royal Statistical Association, Series B* **67**, 301-320 (2005).
5. M. J. Garnett, et al, *Nature* **483**, 570-577 (2012).
6. J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky *Statistical Science* **14**, 382-417 (1999).
7. G. H. Golub, C. F. Van Loan, *Matrix Computations* 3rd edition (Johns Hopkins, 1996).
8. T. Hastie, R. Tibshirani, Friedman, *The Elements of Statistical learning: data mining, inference, and prediction* 2nd edition (Springer, 2009).
9. J. M. Bernardo, A. F. M. Smith, *Bayesian Theory* (Wiley, 1994).
10. M. A. Woodbury, *Inverting modified matrices, Memorandum Rept. 42, Statistical Research Group* (Princeton University, Princeton, NJ, 1950).
11. J. H. Friedman, T. Hastie, R. Tibshirani, *Journal of Statistical Software* **33** (1), (2010).
12. F. Wilcoxon, *Biometrics Bulletin* **1**, 80-83 (1945).
13. L. E. MacConaill, et al, *PloS One* **4**(11), e7887 (2009).
14. I. S. Jang, et al, *Pacific Symposium on Biocomputing* (accepted) (2014).
15. L. M. Shi, et al, *Nature Biotechnology* **28**, 827-838 (2010).
16. T. Park, G. Casella, *Journal of the American Statistical Association* **103**, 681-686, (2008).
17. M. Kyung, J. Gill, M. Ghosh, G. Casella, *Bayesian Analysis* **5**, 369-412 (2010).
18. R Core Team, *R Foundation for Statistical Computing* (Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/> 2012).

# SHARING INFORMATION TO RECONSTRUCT PATIENT-SPECIFIC PATHWAYS IN HETEROGENEOUS DISEASES

ANTHONY GITTER<sup>1,2</sup>, ALFREDO BRAUNSTEIN<sup>3,4</sup>, ANDREA PAGNANI<sup>3,4</sup>, CARLO BALDASSI<sup>3,4</sup>, CHRISTIAN BORGS<sup>1</sup>, JENNIFER CHAYES<sup>1</sup>, RICCARDO ZECCHINA<sup>3,4</sup>, ERNEST FRAENKEL<sup>2,\*</sup>

<sup>1</sup>*Microsoft Research, Cambridge, MA, USA*

<sup>2</sup>*Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

<sup>3</sup>*DISAT and Center for Computational Sciences, Politecnico di Torino, Turin, Italy*

<sup>4</sup>*Human Genetics Foundation, Turin, Italy*

\**E-mail: fraenkel-admin@mit.edu*

Advances in experimental techniques resulted in abundant genomic, transcriptomic, epigenomic, and proteomic data that have the potential to reveal critical drivers of human diseases. Complementary algorithmic developments enable researchers to map these data onto protein-protein interaction networks and infer which signaling pathways are perturbed by a disease. Despite this progress, integrating data across different biological samples or patients remains a substantial challenge because samples from the same disease can be extremely heterogeneous. Somatic mutations in cancer are an infamous example of this heterogeneity. Although the same signaling pathways may be disrupted in a cancer patient cohort, the distribution of mutations is long-tailed, and many driver mutations may only be detected in a small fraction of patients. We developed a computational approach to account for heterogeneous data when inferring signaling pathways by sharing information across the samples. Our technique builds upon the prize-collecting Steiner forest problem, a network optimization algorithm that extracts pathways from a protein-protein interaction network. We recover signaling pathways that are similar across all samples yet still reflect the unique characteristics of each biological sample. Leveraging data from related tumors improves our ability to recover the disrupted pathways and reveals patient-specific pathway perturbations in breast cancer.

*Keywords:* Prize-collecting Steiner forest, breast cancer, protein-protein interactions

## 1. Introduction

Cancer is caused by mutations or other alterations that perturb normal biological processes in a manner that confers a selective growth advantage to the mutated cell. Massive efforts to sequence the DNA of thousands of tumors have detected hundreds of thousands of mutations [1]. However, due to the heterogeneity of tumors, very few genes are mutated frequently enough to be identified as driver genes [1] — those that yield a growth advantage — and generally the significantly mutated genes are already known cancer genes [2]. Fortunately, although even tumors within a specific subtype of cancer may be genetically diverse, the perturbed pathways are similar [1]. A promising direction is therefore combining genomic data with complementary data types to focus on these signaling pathways [2] and computationally searching for ‘driver pathways’ instead of individual driver genes.

Existing algorithms for analyzing cancer are unable to learn patient-specific driver pathways. Many algorithms find modules or subnetworks of altered genes [3–8] but produce a single set of modules for all tumors instead of tumor-specific predictions, limiting the potential for individualized therapies. PARADIGM [9] addresses this issue by combining multiple types of data to learn protein and pathway activity for each individual tumor. However, it relies on

fixed collections of pathways from pathway databases, which are inconsistent and incomplete even in model organisms like yeast [10] and can be altered by gain-of-function mutations [11].

*De novo* pathway discovery has been successful in other biological settings [10, 12–18], but previous approaches are not suitable for analyzing genomic alterations in cancer patients. Most pathway inference algorithms operate on a single set of input. In the cancer setting, this input is data from a single tumor, which makes it very difficult to determine which meaningful genes should compose the driver pathway amid the more numerous passenger mutations.

To overcome the noisiness of the input, we propose to discover tumor-specific driver pathways by leveraging the wealth of data that is available for other tumors of the same cancer subtype. Instead of learning pathways independently for all tumor samples we study all tumors simultaneously, constraining the predicted pathways to be similar. This idea is similar to what is known as multitask learning in other domains [19]. As we demonstrate in simulated settings and with real data from basal-like breast cancer tumors, such an approach can recover individualized driver pathways that contain common core elements that are relevant to the disease even though they may not be mutated in each tumor.

## 2. Methods

### 2.1. Prize-collecting Steiner forest

The prize-collecting Steiner forest (PCSF) algorithm [16] is a computational technique for *de novo* signaling pathway discovery. Given a biological network, such as a protein-protein interaction (PPI) network, and a set of proteins in the network that are believed to be relevant to a disease or condition of interest, PCSF returns a small subnetwork that connects a subset of the disease-related proteins with high-confidence paths. These paths typically reveal additional proteins termed ‘Steiner nodes’ that were not initially implicated as disease proteins but are useful in forming concise, trusted connections among the disease proteins. The discovered subnetwork is a forest, a collection of trees.

Formally, the PPI network is represented as a weighted graph  $G(V, E)$  where  $V$  is the set of proteins and  $E$  is the set of interactions between those proteins. A cost function assigns a cost  $c(e) > 0 \quad \forall e \in E$  and a prize function  $P$  assigns prizes  $p(v) \in \mathbb{R} \quad \forall v \in V$ . Prizes are derived from biological data such as gene expression or quantitative proteomic data.  $p(v) > 0$  indicates that the protein is biologically altered and should be included in the Steiner forest, if possible, with the magnitude indicating the degree of relevance to the disease or condition.  $p(v) = 0$  denotes that there is no observed data for vertex  $v$  or no prior reason to believe it is relevant to the disease, and such vertices compose the potential Steiner nodes. The original PCSF optimization problem [16] is defined as  $\argmin_F o(F)$  where

$$o(F) = \beta \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) + \omega \kappa \quad (1)$$

where  $V_F$  and  $E_F$  are the vertices and edges of the forest  $F$  and  $\kappa$  is the number of trees in the forest.  $\beta$  is a parameter that controls the tradeoff between including prizes and avoiding expensive edges, and  $\omega$  is a parameter that controls how many distinct trees are in the forest.

A PCSF instance can be transformed into a prize-collecting Steiner tree (PCST) instance



by adding an artificial vertex  $v_0$  that must be included in the Steiner tree and artificial edges  $E_0 = V \times \{v_0\}$  with  $c(e) = \omega \quad \forall e \in E_0$  [16]. Without loss of generality we can instead connect  $v_0$  only to prize nodes, vertices for which  $p(v) > 0$ , because in an optimal solution any tree connected to  $v_0$  must contain at least one prize. PCST is NP-hard so we recover an approximate solution using an efficient message-passing algorithm [13] that performs very well on benchmarks [20] and has been shown to be optimal in certain cases [20]. From the approximate PCST solution, we solve the original PCSF instance by deleting  $v_0$  and its incident edges. In all analyses here, we set  $\omega = 1.0$  to bias toward solutions with few connected components.

## 2.2. Multi-sample prize-collecting Steiner forest

The original PCSF formulation is designed for a single set of prizes from a single sample, condition, or patient. However, in many settings there are multiple samples that are expected to have some common properties even though the prizes may be very heterogeneous across the samples. This is particularly the case when the data are derived from patients who suffer from the same disease. In these cases, we would like to find a middle ground between two extremes. On the one hand, treating each patient in isolation ignores valuable data that can more accurately identify the common disease pathway. On the other, if we merge all the patient data, we miss patient-specific aspects of the disease. To address this challenge, we introduce the multi-sample prize-collecting Steiner forest (Multi-PCSF) problem.

We define ‘artificial prizes’  $\phi$  (described below) that are derived from the frequency at which a node is included in forests for all the samples. By adding  $\phi$  to the sample-specific prizes, we introduce a link that constrains the forests to be similar but not identical. Below we introduce two alternative definitions for  $\phi$ , one that tends to increase precision and one that promotes recall, and provide an algorithm to solve the Multi-PCSF problem.

Without loss of generality we assume that PCSF instances are transformed to PCST instances as described above. We further assume that  $\beta$  does not change during the execution of the algorithm, which allows us to redefine  $p(v) = \beta \hat{p}(v)$  before execution, where  $\hat{p}(v)$  are the original prizes from the biological data. We can then simplify Equation 1 to

$$o(F) = \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) \quad (2)$$

which is a PCST instance whose solution can be transformed into a PCSF solution.

In the Multi-PCSF setting we have  $N$  samples and each sample  $i \in \{1, \dots, N\}$  has its own prize function  $P_i$ . The goal is to learn a collection of forests  $\mathbf{F} = \{F_1, \dots, F_N\}$  that are constrained to be similar to one another yet still reflect the diversity of the prizes in each sample. We expand the objective to create a joint objective function over the collection of forests  $\mathbf{F}$  and solve  $\argmin_{\mathbf{F}} o(\mathbf{F})$  where

$$o(\mathbf{F}) = \sum_{i=1}^N o(F_i) + \lambda \sum_{i=1}^N \sum_{v \notin V_{F_i}} \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}) \quad (3)$$

The term  $o(F_i)$  refers to the single forest objective function (Equation 2). The function  $\phi$  is a new term that promotes similarity among all  $F_i \in \mathbf{F}$  by introducing artificial prizes. The

parameter  $\lambda$  controls the tradeoff between  $F_i$  that is similar to the other forests versus  $F_i$  that concisely explains the observed data for tumor sample  $i$ . The role of  $\lambda$  is similar to how  $\beta$  controls the tradeoff between prizes and edge costs in the original PCST formulation.

The first of the two definitions of  $\phi$  uses positive artificial prizes

$$\phi(\alpha, v, p(v), \mathbf{F}) = \begin{cases} \left( \frac{\sum_{i=1}^{|\mathbf{F}|} \mathbb{1}(v \in V_{F_i})}{|\mathbf{F}|} \right)^\alpha, & \text{if } p(v) = 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The positive artificial prizes provide rewards for including nodes that are common to many other forests.  $\mathbb{1}(v \in V_{F_i})$  is an indicator function that has the value 1 if forest  $F_i$  contains vertex  $v$ . The artificial prize on  $v$  is therefore determined by the fraction of other forests that contain  $v$ . The parameter  $\alpha$  allows for non-linear relationships between the fraction and the artificial prize. As  $\alpha$  grows, the vertices that are in many other forests will have larger artificial prizes relative to the vertices in few other forests.

To optimize Equation 3 we iteratively refine our estimates of the optimal forest for each sample given all other samples' current forests for a fixed number of iterations (five here) or until  $\mathbf{F}$  converges. At the first iteration we set  $\lambda = 0$  so that each optimal  $F_i$  is independent of  $F_j \quad \forall i \neq j$  because there is no similarity constraint imposed. At all subsequent iterations, we update each  $F_i$  individually in a sequential random order using the fixed current estimate of all  $\mathbf{F} \setminus \{F_i\}$ . Below we show how to update  $F_i$  by formulating a new PCST instance with modified prizes. To derive the modified prizes we consider only the  $i$ th term of each summation in Equation 3 to approximately solve  $\argmin_{F_i} o_i(\mathbf{F})$ .

$$\begin{aligned} o_i(\mathbf{F}) &= o(F_i) + \lambda \sum_{v \notin V_{F_i}} \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}) \\ &= \sum_{v \notin V_{F_i}} p_i(v) + \sum_{e \in E_{F_i}} c(e) + \lambda \sum_{v \notin V_{F_i}} \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}) \\ &= \sum_{v \notin V_{F_i}} (p_i(v) + \lambda \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\})) + \sum_{e \in E_{F_i}} c(e) \end{aligned} \quad (5)$$

By substituting the definition of  $o(F_i)$  from Equation 2 into Equation 5 and rearranging the terms we can define a new prize function  $P'_i$  that adds artificial prizes to the original  $P_i$

$$\begin{aligned} p'_i(v) &= p_i(v) + \lambda \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}) \\ &= \begin{cases} \lambda \left( \frac{\sum_{i=1}^{|\mathbf{F} \setminus \{F_i\}|} \mathbb{1}(v \in V_{F_i})}{|\mathbf{F} \setminus \{F_i\}|} \right)^\alpha, & \text{if } p_i(v) = 0 \\ p_i(v), & \text{otherwise} \end{cases} \end{aligned} \quad (6)$$

We obtain the new PCST instance that can be solved as described in Section 2.1.

$$o_i(\mathbf{F}) = \sum_{v \notin V_F} p'_i(v) + \sum_{e \in E_F} c(e) \quad (7)$$

The alternative definition of  $\phi$  uses negative artificial prizes, which encourage the algorithm to exclude potential Steiner nodes that appear in few other forests. We define

$$\phi(\alpha, v, p(v), \mathbf{F}) = \begin{cases} -\left(\frac{\sum_{i=1}^{|\mathbf{F}|} \mathbb{1}(v \notin V_{F_i})}{|\mathbf{F}|}\right)^\alpha, & \text{if } p(v) = 0 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

The algorithm is otherwise identical except the updated prize function  $P'_i$  becomes

$$\begin{aligned} p'_i(v) &= p_i(v) + \lambda \phi(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}) \\ &= \begin{cases} -\lambda \left(\frac{\sum_{i=1}^{|\mathbf{F} \setminus \{F_i\}|} \mathbb{1}(v \notin V_{F_i})}{|\mathbf{F} \setminus \{F_i\}|}\right)^\alpha, & \text{if } p_i(v) = 0 \\ p_i(v), & \text{otherwise} \end{cases} \end{aligned} \quad (9)$$

### 2.3. Simulated data

In our first analysis, we generated a synthetic scale-free PPI network using the Barabási-Albert preferential attachment model [21] with 1000 total nodes, 10 initial nodes, and 10 edges per new node attached (9900 total edges). We created artificial pathways by initiating a depth-first search from a randomly selected root node in the graph. The search visited at most two children per node up to a maximum depth of five. Given a pathway with  $m$  nodes and parameters  $f$  (pathway fraction) and  $n$  (noise level), we simulated patients by selecting  $\lceil fm \rceil$  prizes from the pathway and  $\lceil n \lceil fm \rceil \rceil$  noisy prizes (nodes that are not on the pathway) as mutated genes. For example, if we have a 1000 node network,  $m = 25$ ,  $f = 0.25$ , and  $n = 2.0$ , we would randomly select 7 pathway members as true prizes and another 14 nodes from the 975 that are not pathway members as noisy prizes for each patient. All edges had a cost of 0.1, and we assigned a prize of 1.0 to all mutated genes.

We tested our Multi-PCSF algorithm under a variety of parameter configurations and for various  $f$  and  $n$  (Section 3.1). We varied one parameter at a time and set all others to their default value (Table 1). For all configurations we tested positive and negative artificial prizes. In each Multi-PCSF run, we simulated 25 patients per pathway and calculated the precision and recall (Equation 10) for each forest.

Table 1. Multi-PCSF parameters

Parameter	Values tested	Default
$\alpha$	1, 2, 3	2
$\beta$	0.25, 0.5, 1.0	0.5
$\lambda$	0.5, 1.0, 2.0	1.0
$f$	0.1, 0.25, 0.5, 1.0	0.25
$n$	0, 0.5, 1.0, 2.0	0.5

$$\text{precision} = \frac{\text{correct predictions}}{\text{total predictions}} \qquad \text{recall} = \frac{\text{correct predictions}}{\text{pathway members}} \quad (10)$$

#### 2.4. Human data

We evaluated Multi-PCSF using two types of human data: canonical pathways and breast cancer data from 98 patients. For both human analyses we used physical PPI from STRING (version 9.0) [22]. Using the edge scores  $s(e)$  from STRING, we removed low confidence interactions with  $s(e) < 0.5$  and defined edge costs as  $\max(0.01, 1 - s(e))$ . We downloaded the ‘Epidermal Growth Factor Receptor Pathway’ (EGFR) from the *Science Signaling* Database of Cell Signaling [23], translating all pathway node names into gene symbols. Three non-protein nodes could not be mapped and retained their original names. We selected only a single gene symbol per gene family to maintain the original pathway topology. We downloaded National Cancer Institute-Nature Pathway Interaction Database (PID) pathways [24] and mapped UniProt ids to gene symbols. To calculate  $P$ -values for PID pathway enrichment, we used the right-tailed Fisher’s exact test. All  $P$ -values were corrected for multiple hypothesis testing by multiplying them by the number of hypotheses tested (Bonferroni correction).

We obtained The Cancer Genome Atlas (TCGA) breast cancer data from the Broad Institute’s Genome Data Analysis Center Firehose (April 21, 2013 analysis run). We considered only the 98 basal-like tumors defined in [25]. For each tumor  $i$ , we defined the prize on a gene to be  $p_i(g) = p_i^m(g) + p_i^p(g)$  where  $p_i^m(g)$  is the number of non-silent mutations or indels in gene  $g$  and  $p_i^p(g)$  denotes proteomic changes in the reverse phase protein array data. If an antibody exhibited a  $\log_2$  scale fold change with magnitude of at least 1.0, we set  $p_i^p(g)$  to be that magnitude and took the maximum magnitude when multiple antibodies mapped to a single gene. To simulate 100 patients in the EGFR pathway, we set  $f = 0.25$  and  $n = 10.0$  and generated noisy prizes as described above. We used  $\alpha = 2$ ,  $\beta = 1.0$ , and  $\lambda \in \{0.5, 1.0, 2.0, 5.0\}$ . For the breast cancer analysis we set  $\alpha = 2$ ,  $\beta = 0.5$ , and  $\lambda = 1.0$ .

#### 2.5. HotNet analysis

We ran generalized HotNet (version 1.0.0) [5, 26], which takes a gene-gene influence matrix and a score on genes as input. We used the influence matrix packaged with HotNet, which is derived from the Human Protein Reference Database (HPRD) PPI network [27], and set the gene score to be  $\sum_{i=1}^N p_i(g)$  where  $N$  is the number of basal-like breast cancer tumors. We allowed HotNet to choose the optimal  $\delta$  parameter, which it selected as  $\delta = 0.05$ , and used all other default parameters (1000 permutations, smin of three, and smax of ten). We defined ‘HotNet PID pathways’ as the five PID pathways that most significantly overlapped a HotNet subnetwork, which happened to be the same 864-gene HotNet subnetwork for all five.

### 3. Results

We tested Multi-PCSF in three increasingly challenging settings to demonstrate how sharing information across samples improves pathway recovery for each individual sample. In the first two test cases, we generated prizes from a known reference pathway and quantified how well

the pathway was recovered. In the third, we analyzed data from 98 patients with basal-like breast cancer tumors and showed that Multi-PCSF produces individualized representations of the signaling pathways that are perturbed in this breast cancer subtype.

### 3.1. Recovering simulated pathways

In order to quantitatively evaluate whether Multi-PCSF improves pathway recovery, we first simulated prizes for cancer samples with a common driver pathway. We simulated a 1000 node scale-free network, which reflects the topology of real PPI networks [28] and allowed us to run Multi-PCSF under a wide range of parameter configurations (solving 32500 PCST instances) to ensure its advantages are not limited to specific settings. We generated a driver pathway that would be altered in each tumor. We then randomly assigned prizes in each synthetic tumor sample to a fraction of the pathway members as well as a fraction of other proteins that are not on the pathway, which represent noisy passenger mutations. We ran baseline PCSF (which does not share information across samples) and Multi-PCSF and calculated precision and recall (Equation 10) for the nodes and edges of each forest. We assessed the average performance over ten synthetic pathways (Figure 1).

With very few exceptions, Multi-PCSF improves both the precision and recall under all tested parameter configurations. The improvements in recall, how much of the reference path-

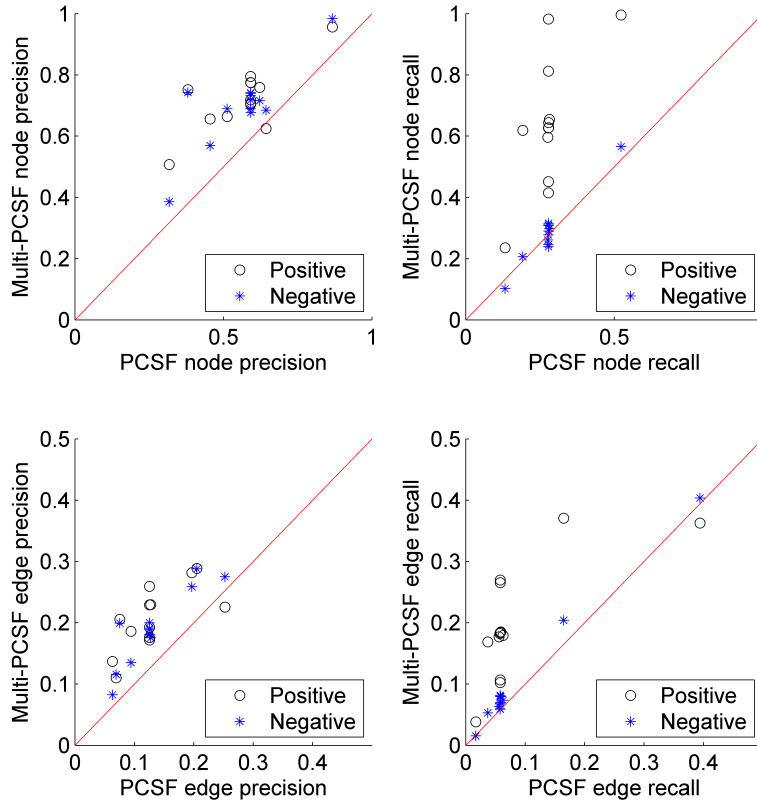


Fig. 1. Node and edge precision and recall for Multi-PCSF versus PCSF on simulated pathways. Positive and negative refer to the Multi-PCSF artificial prizes. Points above the red diagonal indicate instances where Multi-PCSF outperforms PCSF.

way is recovered, are especially notable. In the best case Multi-PCSF node recall is 3.5 times greater than PCSF and edge recall is 4.6 times greater. On this instance PCSF node recall is 0.28 signifying that for most synthetic tumors the prize nodes are the only pathway members that could be recovered. Multi-PCSF node recall is 0.98 — in most cases the entire pathway could be recovered. Positive artificial prizes yield greater improvements in recall than negative artificial prizes. With positive prizes, Multi-PCSF includes proteins that are shared by many other forests even if they are not needed to connect additional prize nodes. Conversely, with negative prizes Multi-PCSF is more likely to use such nodes as Steiner nodes but will not include them in a forest unless they help connect prize nodes.

### 3.2. Recovering the *EGFR* signaling pathway

Having established that Multi-PCSF can substantially improve pathway recovery in a simulated setting, we assessed its performance in a human PPI network. We selected the human EGFR pathway as the hypothetical driver pathway that was perturbed in a cohort of simulated tumors and applied both Steiner forest algorithms. The randomly generated prizes in this setting were much noisier than in the simulated pathway setting to better reflect the large number of passenger alterations per driver mutation in real cancer datasets. For every functional prize selected from the EGFR pathway, we added ten noisy prizes from elsewhere in the PPI network. We simulated 100 tumor samples, ran PCSF and Multi-PCSF, and calculated precision and recall (Figure 2). For Multi-PCSF we varied  $\lambda$ , which controls the strength of the constraint that requires forests to be similar to one another.

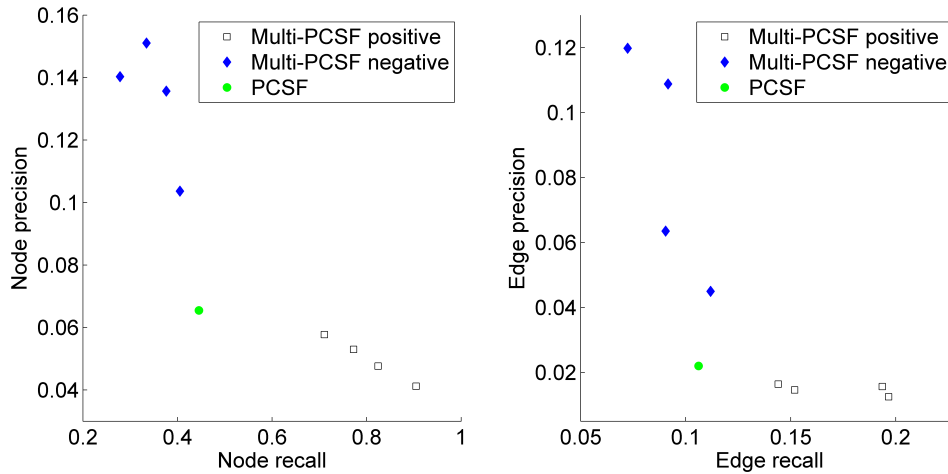


Fig. 2. Precision-recall graphs for Multi-PCSF with positive and negative artificial prizes and baseline PCSF on the human EGFR pathway. The four Multi-PCSF points correspond to different values of  $\lambda$ .

In the EGFR setting, PCSF node precision is only 0.065 and edge precision is 0.022 because even the noisy prizes could often be connected to the EGFR pathway members. By sharing information across samples, Multi-PCSF is better able to discern which prizes are spurious and which potential Steiner nodes are preferable because they are perturbed in other samples. With positive artificial prizes, proteins that are members of other forests (as either prize nodes

or Steiner nodes) are introduced as Steiner nodes. This enhances recall, which increases with  $\lambda$ , culminating in a 2.0 times improvement in node recall and 1.9-fold edge recall improvement when  $\lambda = 5.0$ . The maximum node recall attained is 0.90. Even in this difficult setting, nearly all proteins on the pathway can be extracted from the PPI network at the expense of a decrease in precision. Parallel paths in the EGFR pathway cannot be captured by our inferred forests, which suggests that edge recall could potentially be further improved by applying perturbation techniques that merge multiple forests and produce more general topologies [16].

With negative artificial prizes, Multi-PCSF excludes proteins that are not useful in other forests, which boosts precision. When  $\lambda = 5.0$  and negative prizes are used, Multi-PCSF node precision is 2.1 times greater than PCSF and edge precision is 5.4 times greater. In addition, when using a weaker similarity constraint ( $\lambda = 0.5$ ), Multi-PCSF exhibits superior precision as well as a small improvement in edge recall.

### 3.3. Pathways in breast cancer

To assess Multi-PCSF's ability to interpret and suggest mechanistic hypotheses about real clinical data we applied it to TCGA breast cancer data [25], inferring the pathways perturbed in these tumors and their common and unique components. Because cancer subtypes defined by mRNA expression similarity are likely to share common driver pathways, we focus on only the basal-like breast cancer subtype (98 tumors). We calculated prizes using the TCGA non-silent mutations and proteomic data. Other data types such as copy number alterations can easily be integrated into our analysis, and we have previously shown how to combine epigenomic features and mRNA expression to place prizes on transcription factors [17]. Some of the tumors had sparse prizes so we used positive artificial prizes in Multi-PCSF to leverage its ability to construct more complete pathways based on alterations in other tumors.

Multi-PCSF achieves our goal of discovering pathways that have a common core structure and many individual characteristics connected to the core that reflect the diverse manners in which the driver pathways are affected in each tumor (Figure 3). The shared core is composed of 198 nodes (8.30% of all nodes appearing in any forest) that are present in all 98 forests. This core likely contains pathways that are altered in all patients despite their heterogeneous mutations. For example, we recover basal-like breast cancer-related proteins such as ATM, BRCA1, BRCA2, MYC, RB1, and TP53 [25]. In addition, we find HIF1A in the common core, consistent with the fact that high HIF1A pathway activity is a key feature of basal-like breast cancers [25]. By jointly analyzing all patients we find potential therapeutic targets that would have been missed in individual analyses. Two genes, ARHGDIA and SMAD2, do not appear in any forests when PCSF is run independently on each sample but appear in the Multi-PCSF common core. ARHGDIA encodes the protein RhoGDI-1, which is overexpressed in breast cancer and blocks chemotherapy drug-induced apoptosis in cancer cells [29]. SMAD2 knockdowns in breast cancer cells suggest it is a tumor suppressor [30].

Although many nodes are identical across the forests, the edges used to connect those nodes to each other vary as only 39 edges (1.36%) are common to all forests. Beyond the shared core, 1411 nodes (59.14%) and 1712 edges (59.55%) appear in only one forest. 917 nodes are Steiner nodes in at least one forest, including all nodes in the common core and 435 nodes

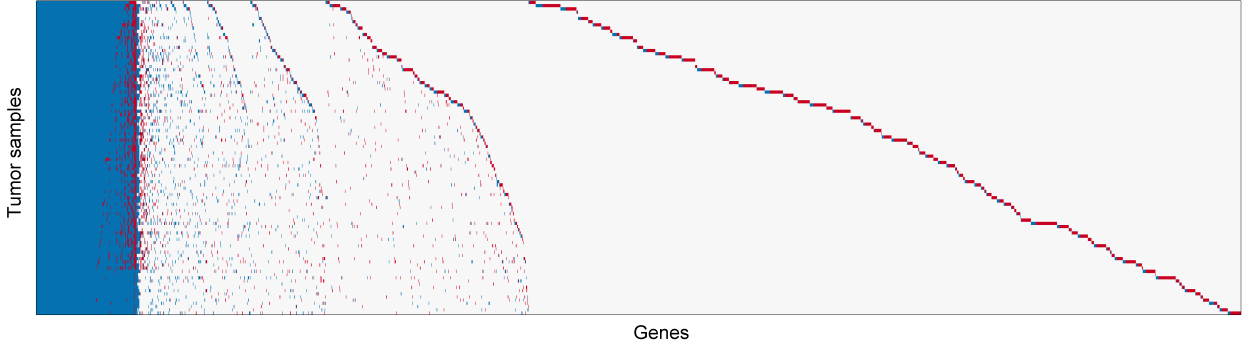


Fig. 3. The heat map summarizes all 98 Multi-PCSF forests. Each row represents the forest for a particular tumor sample and depicts which nodes are collected prizes (red), Steiner nodes (blue), and absent (white).

that are present in multiple but not all forests. The variation among the forests demonstrates that even tumors within a single subtype cannot be represented by a single pathway structure.

HotNet [5, 26] is an algorithm for discovering PPI subnetworks that are significantly affected in a cancer dataset. We applied generalized HotNet to the basal-like tumor data, providing HotNet’s HPRD-derived gene-gene influence matrix and the same mutation- and proteomic-based prizes as input. HotNet returned 109 subnetworks. One large subnetwork contained 864 proteins and the other subnetworks had two to seven members. HotNet’s subnetworks significantly overlap PID pathways (Table 2), which we refer to as HotNet PID pathways (Section 2.5), demonstrating that HotNet can reveal which reference pathways are relevant in a cancer subtype. However, because it produces a single list of subnetworks for the entire subtype and does not reveal hidden pathway members (the equivalent of Steiner nodes in PCSF), it is difficult to use HotNet to generate mechanistic hypotheses or guide individualized treatment. Although HotNet would produce different results if we tune its parameters to generate smaller subnetworks or use an influence matrix derived from the STRING PPI network, these fundamental differences between HotNet and Multi-PCSF would remain.

Multi-PCSF not only recovers forests that capture the same annotated pathways as HotNet, but it also presents custom versions of those pathways for each tumor, which better

Table 2. HotNet PID pathways and whether they significantly overlap Multi-PCSF forests, PCSF forests, or both (corrected  $P \leq 0.05$ ). If both, the table shows whether the overlap is better or worse for Multi-PCSF.

HotNet PID pathway	HotNet subnetwork overlap corrected $P$	Only Multi-PCSF	Better Multi-PCSF	Worse Multi-PCSF	Only PCSF
SHP2 signaling	9.36 E-10	65	33	0	0
IL2-mediated signaling events	2.97 E-9	36	62	0	0
Signaling events mediated by Stem cell factor receptor (c-Kit)	3.08 E-9	29	69	0	0
Integrins in angiogenesis	7.80 E-9	60	38	0	0
GMCSF-mediated signaling events	4.23 E-8	45	53	0	0



enables follow-up biological analysis. In many cases standard PCSF does not recover the reference pathways affected in the basal-like subtype because it does not leverage data from related tumors. For all tumors where the PCSF forest is significantly enriched with a PID pathway, the enrichment is stronger after sharing information with Multi-PCSF (Table 2). Individualized representations of the PID pathways, such as ‘Signaling events mediated by Stem cell factor receptor (c-Kit)’, could potentially lead to new therapeutic strategies for subsets of the basal-like breast cancer cases. KIT abnormalities have been implicated in several other cancers [31], and KIT-positive gastrointestinal stromal tumors have been approved for Gleevec (imatinib) treatment [32]. Post-processing procedures for prioritizing Steiner tree members have shown that highly-ranked Steiner nodes validate *in vitro* [17] and can be applied here to guide subsequent analysis of the individual pathway predictions.

#### 4. Discussion

The prize-collecting Steiner forest algorithm is a powerful approach for integrating genomic, proteomic, transcriptional, and epigenomic data to reconstruct signaling pathways. Our multi-sample extension enables PCSF to analyze heterogeneous data, where prizes vary greatly across a collection of samples, and to exploit information from related samples despite the prize-level dissimilarities. Multi-PCSF is an especially pertinent tool for large-scale cancer profiling studies because the most frequently recurring alterations have already been identified (leaving the non-recurrent abnormalities for further interpretation) and we seek to understand the unique causes of oncogenesis in each tumor. The artificial prizes introduced in Multi-PCSF facilitate constructing accurate patient-specific driver pathways despite the presence of numerous passenger mutations by promoting genes that are driver pathway members in other tumors. Algorithms like HotNet can reveal which processes are affected in a patient cohort but do not guide individualized treatment (although recent diffusion-based algorithms [33] aim to lift this limitation). Multi-PCSF is also widely applicable beyond cancer and can model data from noisy biological replicates without initially aggregating all replicates, study responses to a collection of stimuli [34], or compare the immune responses to related viruses [15].

#### Acknowledgements

We thank Nurcan Tuncbag and Fabrizio Altarelli for discussions about Steiner forests as well as Anthony Soltis and Sara Gosline for preparing network data. This work was supported in part by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the US Army Research Office (the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred), by NIH grant U54-CA112967, and by European Grants FET Open No. 265496 and ERC No. 267915, as well as computing resources funded by the National Science Foundation under Award No. DB1-0821391.

#### References

- [1] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz and K. W. Kinzler, *Science* **339**, 1546 (2013).

- [2] M. B. Yaffe, *Sci Signal* **6**, pe13 (2013).
- [3] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway and D. Pe'er, *Cell* **143**, 1005 (2010).
- [4] E. Cerami, E. Demir, N. Schultz, B. S. Taylor and C. Sander, *PLoS ONE* **5**, e8918 (2010).
- [5] F. Vandin, E. Upfal and B. J. Raphael, *J Comput Biol* **18**, 507 (2011).
- [6] G. Ciriello, E. Cerami, C. Sander and N. Schultz, *Genome Res* **22**, 398 (2012).
- [7] J. Zhao, S. Zhang, L.-Y. Wu and X.-S. Zhang, *Bioinformatics* **28**, 2940 (2012).
- [8] M. D. M. Leiserson, D. Blokh, R. Sharan and B. J. Raphael, *PLoS Comput Biol* **9**, e1003054 (2013).
- [9] A. J. Sedgewick, S. C. Benz, S. Rabizadeh, P. Soon-Shiong and C. J. Vaske, *Bioinformatics* **29**, i62 (2013).
- [10] A. Gitter, M. Carmi, N. Barkai and Z. Bar-Joseph, *Genome Res* **23**, 365 (2013).
- [11] R. Brosh and V. Rotter, *Nat Rev Cancer* **9**, 701 (2009).
- [12] C.-H. Yeang, T. Ideker and T. Jaakkola, *J Comput Biol* **11**, 243 (2004).
- [13] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François and R. Zecchina, *Proc Natl Acad Sci* **108**, 882 (2011).
- [14] Y.-A. Kim, S. Wuchty and T. M. Przytycka, *PLoS Comput Biol* **7**, e1001095 (2011).
- [15] A. Gitter and Z. Bar-Joseph, *Bioinformatics* **29**, i227 (2013).
- [16] N. Tuncbag, A. Braunstein, A. Pagnani, S.-S. C. Huang, J. Chayes, C. Borgs, R. Zecchina and E. Fraenkel, *J Comput Biol* **20**, 124 (2013).
- [17] S.-s. C. Huang, D. C. Clarke, S. J. C. Gosline, A. Labadorf, C. R. Chouinard, W. Gordon, D. A. Lauffenburger and E. Fraenkel, *PLoS Comput Biol* **9**, e1002887 (2013).
- [18] N. Atias and R. Sharan, *Mol BioSyst* **9**, 1662 (2013).
- [19] S. J. Pan and Q. Yang, *IEEE Trans Knowl Data Eng* **22**, 1345 (2010).
- [20] I. Biazzo, A. Braunstein and R. Zecchina, *Phys Rev E* **86**, 026706 (2012).
- [21] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
- [22] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. v. Mering, *Nucleic Acids Res* **39**, D561 (2011).
- [23] N. R. Gough, *Ann NY Acad Sci* **971**, 585 (2002).
- [24] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay and K. H. Buetow, *Nucleic Acids Res* **37**, D674 (2009).
- [25] The Cancer Genome Atlas Network, *Nature* **490**, 61 (2012).
- [26] F. Vandin, P. Clay, E. Upfal and B. J. Raphael, *Pac Symp Biocomput*, 55 (2012).
- [27] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, *Nucleic Acids Res* **37**, D767 (2009).
- [28] H. Jeong, S. P. Mason, A. L. Barabási and Z. N. Oltvai, *Nature* **411**, 41 (2001).
- [29] B. Zhang, Y. Zhang, M.-C. Dagher and E. Shacter, *Cancer Res* **65**, 6054 (2005).
- [30] M. Petersen, E. Pardali, G. van der Horst, H. Cheung, C. van den Hoogen, G. van der Pluijm and P. ten Dijke, *Oncogene* **29**, 1351 (2010).
- [31] J. Lennartsson and L. Rönnstrand, *Physiol Rev* **92**, 1619 (2012).
- [32] H. Joensuu, *Nat Rev Clin Oncol* **9**, 351 (2012).
- [33] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler and J. M. Stuart, *Bioinformatics* (2013).
- [34] S. J. C. Gosline, S. J. Spencer, O. Ursu and E. Fraenkel, *Integr Biol* **4**, 1415 (2012).

## DETECTING STATISTICAL INTERACTION BETWEEN SOMATIC MUTATIONAL EVENTS AND GERMLINE VARIATION FROM NEXT-GENERATION SEQUENCE DATA

HAO HU

*Department of Epidemiology  
The University of Texas MD Anderson Cancer Center  
1155 Pressler Street  
Houston, TX, 77030, USA  
Email: [hhu1@mdanderson.org](mailto:hhu1@mdanderson.org)*

CHAD D. HUFF

*Department of Epidemiology  
The University of Texas MD Anderson Cancer Center  
1155 Pressler Street  
Houston, TX, 77030, USA  
Email: [chuff1@mdanderson.org](mailto:chuff1@mdanderson.org)*

The two-hit model of carcinogenesis provides a valuable framework for understanding the role of DNA repair and tumor suppressor genes in cancer development and progression. Under this model, tumor development can initiate from a single somatic mutation in individuals that inherit an inactivating germline variant. Although the two-hit model can be an overgeneralization, the tendency for the pattern of somatic mutations to differ in cancer patients that inherit predisposition alleles is a signal that can be used to identify and validate germline susceptibility variants. Here, we present the Somatic-Germline Interaction (SGI) tool, which is designed to identify statistical interaction between germline variants and somatic mutational events from next-generation sequence data. SGI interfaces with rare-variant association tests and variant classifiers to identify candidate germline susceptibility variants from case-control sequencing data. SGI then analyzes tumor-normal pair next-generation sequence data to evaluate evidence for somatic-germline interaction in each gene or pathway using two tests: the Allelic Imbalance Rank Sum (AIRS) test and the Somatic Mutation Interaction Test (SMIT). AIRS tests for preferential allelic imbalance to evaluate whether somatic mutational events tend to amplify candidate germline variants. SMIT evaluates whether somatic point mutations and small indels occur more or less frequently than expected in the presence of candidate germline variants. Both AIRS and SMIT control for heterogeneity in the mutational process resulting from regional variation in mutation rates and inter-sample variation in background mutation rates. The SGI test combines AIRS and SMIT to provide a single, unified measure of statistical interaction between somatic mutational events and germline variation. We show that the tests implemented in SGI have high power with relatively

modest sample sizes in a wide variety of scenarios. We demonstrate the utility of SGI to increase the power of rare variant association studies in cancer and to validate the potential role in cancer causation of germline susceptibility variants.

## 1. Introduction

In 1971, Alfred Knudson proposed the two-hit hypothesis for retinoblastoma, demonstrating that the distribution of age-of-onset for familial retinoblastoma cases was consistent with inheritance of a germline variant followed by a somatic mutation, while age-of-onset for sporadic cases was consistent with two independent somatic mutations<sup>1</sup>. The gene responsible for this process was identified 15 years later as *RBI*, the first tumor suppressor gene<sup>2,3</sup>. The two-hit hypothesis is now the classic model for DNA repair and tumor suppressor genes, which follow a dominant mode of inheritance but are typically recessive at the cellular level. This model provides a useful framework for understanding cancer predisposition, although DNA repair and tumor suppressor genes can be either dominant or recessive at the cellular level, depending on the context. Germline mutations in the tumor suppressor gene *TP53* follow both one- and two-hit models in Li-Fraumeni syndrome, with some inherited genetic causes resulting from cellular recessive loss-of-function nonsense variants and others resulting from dominant gain-of-function missense variants<sup>4</sup>. The DNA repair genes *BRCA1* and *BRCA2* variants are also either recessive or dominant at the cellular level depending on the type of cancer, with complete loss of the wild type allele in ovarian cancer but occasional haplo-insufficiency in breast cancer<sup>5</sup>. In general, inherited variants in the tumor suppressor gene *APC* are recessive at the cellular level in colorectal cancer<sup>6</sup>, but can exert dominant effects that can lead to chromosomal instability<sup>7</sup>. In contrast to DNA repair and tumor suppressor genes, oncogenes are generally dominant at both the germline and cellular levels, and thus tend to follow a one-hit model. Nonetheless, there are a number of examples of oncogenes that follow a two-hit model<sup>8</sup>. Thus, although one- and two-hit models are sometimes overgeneralizations, many genes display a pattern of somatic mutational events in tumors that occur more or less frequently than expected among individuals that carry particular germline susceptibility variants.

Next-generation sequencing now provides efficient, high-coverage interrogation of nearly the entire genome and is revolutionizing our understanding of somatic mutational events that drive tumorigenesis<sup>9-11</sup>. The use of next-generation sequencing to identify rare germline variants that influence cancer risk also holds great promise but is fundamentally a more difficult problem given that purifying selection ensures that intermediate-penetrance germline variants are usually very rare. A number of rare variant association tests have been developed recently to identify disease-susceptibility genes from case-control next-generation sequence data. The primary advantage of these methods over traditional approaches is that they aggregate rare variants to perform a single statistical test for each gene, which greatly increases power while reducing the multiple testing burden. However, as we have previously shown, although rare variant association tests greatly improve statistical power, studies involving thousands of cases and controls will likely be needed to identify novel gene associations for common cancers<sup>12-14</sup>. The tendency for somatic mutational events to occur more or less frequently than expected given the presence of a germline

susceptibility variant is an additional piece of evidence that can aid in the search for novel gene-cancer susceptibility associations or in the validation and characterization of candidate germline susceptibility variants. The primary motivation of this work is to provide a framework for identifying these statistical interactions between somatic and germline variation in a high-throughput manner that takes advantage of available bioinformatic tools and existing next-generation sequencing capacity. The methods we present are implemented in the Somatic Germline Interaction (SGI) tool.

SGI analyzes next-generation sequencing data from tumor-normal tissue pairs and normal tissue in matched controls to determine whether germline variation in a gene or pathway statistically interacts with the occurrence of somatic events. The two-hit model describes one process that can result in statistical interaction, in which two damaged copies of a gene are required to initiate tumorigenesis. If the two-hits model holds, then the tumors of cancer patients with a deleterious germline variant in a driver gene are likely to have a second somatic mutation event in the same gene. Another process that can result in statistical interaction involves *cis*-acting germline variants that can greatly increase the somatic mutation rate in the local genomic region<sup>15-18</sup>. SGI identifies candidate germline susceptibility variants by interfacing with the Variant Annotation, Analysis and Search Tool (VAAST)<sup>19</sup>. The rare variant association test in VAAST incorporates amino acid substitution severities, phylogenetic conservation, and the distribution of allele frequencies in cases and controls to variants and genes that are likely to influence disease susceptibility<sup>12</sup>. After identifying individuals in the study with candidate germline variants, SGI then analyzes tumor-normal pair sequence data to evaluate whether somatic mutational events occur more or less frequently than expected by testing the null hypothesis that the occurrence of somatic events is independent of the presence or absence of germline variation.

We divide somatic mutational events into two categories: somatic mutations and preferential allelic imbalance. SGI implements the Allelic Imbalance Rank Sum (AIRS) test to evaluate evidence for preferential allelic imbalance. Specifically, within each gene or pathway, AIRS tests whether the chromosomes harboring putatively deleterious germline mutations are preferentially amplified in tumor tissues. Allelic imbalance is an important signal of somatic mutations resulting from copy number variants (CNVs) or loss-of-heterozygosity (LOH) that has been used to identify and validate modest penetrance germline-cancer associations in both humans<sup>15,20,21</sup> and mice<sup>22-24</sup>. In addition to allelic imbalance, SGI also evaluates whether somatic mutations occur more or less frequently than expected in the tumors of individuals that harbor putatively deleterious germline mutations using the Somatic Mutation Interaction Test (SMIT). SMIT only considers single nucleotide and small indel somatic mutations that do not result in LOH or CNVs in a large genomic region, as these larger somatic events are evaluated by allelic imbalance evidence. SGI also combines AIRS and SMIT to provide a single unified framework to detect statistical interaction between germline and somatic variation.

SGI has a number of potential applications. For known germline-susceptibility genes, SGI can validate germline variants of unknown significance. For genes that are known to be significantly mutated in tumors but not known to play a role in cancer predisposition, SGI can search for novel

germline variant associations. SGI can also identify novel cancer-associated genes that would be much more difficult to detect than germline case-control studies or somatic mutational analysis alone due to rarity and/or effect size. Here, we present the methods implemented in SGI and evaluate the performance of the tool in a wide variety of scenarios.

## 2. Methods

### 2.1. Identifying candidate germline variants

SGI processes VAAST output files to identify individuals with candidate germline susceptibility variants. For each gene, any variant that has a VAAST score of greater than 0 is identified as a candidate. SGI then performs the AIRS and SMIT tests based on the binary classification of individuals with and without candidate germline susceptibility variants. The VAAST score threshold is a tunable parameter. Other association tests can be supported, but require combining the association test results with a variant classifier – such as SIFT<sup>25,26</sup>, PolyPhen-2<sup>26</sup>, Align-GVGD<sup>27,28</sup>, or VAAST 2.0<sup>12</sup> – to identify candidate susceptibility variants. For the AIRS and SMIT tests below, set  $A$  contains the affected individuals with candidate germline susceptibility variants, and set  $B$  contains all other affected individuals.

### 2.2. AIRS

AIRS evaluates candidate germline susceptibility variants to test for preferential allelic imbalance. For each individual  $i$  at site  $j$ , we use the raw somatic read counts for the reference and non-reference allele for each germline heterozygous to calculate the binomial one-tail probability,  $p_{ij}$ , that the allele frequency of the non-reference allele is greater than 0.5. To control for inter-sample variation in the distribution of allelic imbalance throughout the genome, we transform  $p_{ij}$  to the percentile rank,  $f_{ij}$ , using the empirical distribution function of binomial p-values among all variant sites throughout the genome for each individual. This transformation does not necessarily require whole-genome data and should effectively control for inter-sample variation in genome-wide levels of allelic imbalance in targeted gene panels that include as few as 50 genes. To control for differences in the level and distribution of allelic imbalance throughout the genome, we restrict the test to variants in or around the gene of interest (by default, all variants between the beginning of the first and the end of the last exon). Let  $G$  equal the set of variants around the gene, and let  $C$  equal the subset of candidate germline susceptibility variants. Our test statistic is a Wilcoxon-Mann-Whitney  $U$  that compares values of  $f_{ij}$  for candidate variants to all other variants in the gene among individuals that do not carry a candidate germline variant:

$$U = \sum_{i \in A} \sum_{j \in C \cap \{f_i\}} \sum_{k \in B} \sum_{l \in G \cap \{f_k\}} I(f_{ij} > f_{kl}) - \frac{v_A(v_A + 1)}{2}, \quad (1)$$

where  $v_A$  is the total number of candidate alleles. When the sample size of either group is under 20, the exact one-tail null probability is calculated. Otherwise, a normal

approximation is assumed.

Although we include only candidate germline alleles from individuals in set A, we include all heterozygous germline alleles from individuals in set B. Including multiple variants from an individual is a violation of the independence assumption in the U test, given that the observation of allelic imbalance in one variant would alter the expected distribution of read counts for other variants in the region. However, in our tests, we observed a modest increase in power and no inflation in Type I error by including all variants from B for sample sizes as small as 40. The inclusion of all heterozygous germline alleles is designed to detect subtle signals of allelic imbalance resulting from low levels of tumor purity or multiclonality. If the allelic imbalance signals are infrequent yet unambiguous, a more powerful alternative is to only include alleles in the rank sum test that are on the tails of the binomial distribution (e.g.,  $p_{ij}$  less than 0.05 or greater than 0.95). These thresholds can be set as optional parameters.

We evaluated two allelic imbalance metrics other than the binomial, the proportion of non-reference alleles and a one-sided Fisher's exact test comparing read counts between normal and somatic tissue. The proportion of non-reference alleles suffered from an inability to account for differences in coverage depth. The Fisher's exact test had the advantage of controlling for allele-specific read count biases that are present in both the normal and somatic data, but this was offset by a modest reduction in power. More sophisticated methods that incorporate haplotype information to test for allelic imbalance, such as Haplotype Amplification in Tumor Sequences (HATS)<sup>29</sup> or Haplotype LOH (hapLOH)<sup>30</sup>, may provide a replacement to the binomial in the future. In all cases, the raw allelic imbalance metric should be transformed using the empirical distribution function for each individual to control for inter-sample variation in the level of allelic imbalance throughout the genome.

### 2.3. SMIT

SMIT is designed to evaluate whether somatic mutations occur more or less frequently than expected for individuals with a candidate germline susceptibility variant in a gene or pathway of interest. More generally, SMIT tests for statistical interaction between somatic mutation frequencies and any binary classifier in a defined genomic feature. SMIT addresses the same general question as the Clinical Correlation Test (CCT) in the Mutational Significance of Cancer package (MuSiC)<sup>31</sup>, but provides the additional advantage of controlling for inter-sample variation in the somatic background mutation rate. Because the same genomic regions are evaluated in the two sample groups, the method is robust to heterogeneity in the mutational process between genomic regions, which is a major potential source of false-positives when searching for cancer-associated genes<sup>10</sup>.

Let  $M$  equal the set of individuals with at least one somatic mutation observed in the genomic feature (typically gene). Let  $t_i$  equal the total number of somatic mutations throughout the genome for sample  $i$ , and let  $l$  equal the proportional length of the gene in base pairs relative to the total sequenced region of the genome. For each sample  $i$ , we estimate the background mutation rate at the gene by the approximation  $r_i = t_i \times l$ . Let  $s_A$  and  $s_B$  equal the probability for sets  $A$  (affected individuals with candidate germline susceptibility variants) and  $B$  (all other individuals),

respectively, that a somatic mutation occurs in the gene through a process that is unrelated to the background mutation rate, which approximates the somatic driver mutation rate. SMIT tests the null hypothesis that  $s_A = s_B$  against the alternative hypothesis that  $s_A \neq s_B$  using a likelihood ratio test:

$$\Lambda = \frac{\prod_{i \in M} r_i + (1-r_i)\hat{s} \prod_{i \notin M} (1-r_i)(1-\hat{s})}{\prod_{i \in A \cap M} r_i + (1-r_i)\hat{s}_A \prod_{i \in \{x \in A | x \notin M\}} (1-r_i)(1-\hat{s}_A) \prod_{i \in B \cap M} r_i + (1-r_i)\hat{s}_B \prod_{i \in \{x \in B | x \notin M\}} (1-r_i)(1-\hat{s}_B)}. \quad (2)$$

We estimate the maximum likelihood of  $s$ ,  $s_A$ , and  $s_B$  using a grid search. Note that when  $r_i$  does not vary between samples and the maximum likelihood of  $s$ ,  $s_A$ , and  $s_B$  are all greater than 0, Eq. 2 collapses to a multinomial likelihood ratio test. We estimate the significance level of the two-tailed test using a chi-square approximation ( $-2\ln\Lambda \sim \chi^2_1$ ). We also implement one-sided tests by applying the appropriate transformations to the significance levels of the two-sided test. The one-tailed test,  $s_A > s_B$  evaluates a cellular recessive (or partially recessive) two-hit hypothesis and the one-tailed test,  $s_B > s_A$  evaluates a cellular dominant (or partially dominant) one-hit hypothesis.

#### 2.4. Somatic-Germline Interaction (SGI) Tool

SGI implements both AIRS and SMIT, and also combines the two tests to evaluate two- and three-hit hypotheses using a Fishers Combined Probability Test (FCPT). We refer to the combined AIRS-SMIT test as the SGI test. We also use the FCPT to perform the VAAST-AIRS, VAAST-SMIT, and VAAST-SGI tests in Figure 4.

#### 2.5. Datasets

The breast cancer samples used in Figures 1 and 3 are from Complete Genomics (CG) whole-genome sequence data of a tumor-normal pair<sup>32</sup>. This sample exhibited high levels of allelic imbalance throughout the genome, with 77% of heterozygous germline SNPs having a somatic allele frequency significantly different from 0.5 at the 0.05 level. In Figures 1 and 4, we used the breast cancer sequence data to establish a distribution of read counts to represent next-generation sequence data in tumors. For individuals without candidate germline susceptibility alleles (group B), we sampled 50 Kb segments with replacement from the breast cancer whole-genome data. To represent the marker density of whole-exome data (approximately 2% of the genome), we performed rejection sampling on each heterozygous germline variant, rejecting each variant with probability 0.98. The top half of Figure 1 was based on the tumor tissue data and represents loci with relatively high levels of allelic imbalance in group B. The bottom half of Figure 1 was based on the normal tissue data and represents loci with very low levels of allelic imbalance in group B. For candidate germline susceptibility alleles in Figure 1 (group A), we simulated the distribution of read counts using the following procedure: For each candidate germline variant in each individual, we first designated it as preferentially amplified with probability  $q$  (between 0.1 and 1). Note that the proportion of samples with higher frequency of the preferred allele is approximately



$q+(1-q)b$ , where  $b$  is the proportion of variants with a higher frequency for the preferred allele in group B (Figure 1). The read counts of alleles not designated as preferentially amplified were randomly sampled from breast cancer whole-genome data. For the remaining variants, we set the total number of reads to a Poisson random variate,  $t$ , with mean equal to 52 to match the mean read count in the normal tissue whole-genome data. We then set the expected proportion of the preferred allele,  $w$ , to between 0.6 and 1 and the number of non-reference to a binomial random variate with parameters  $t$  and  $w$ . In Figure 4, the breast cancer *ATM* case-control sequence data in Figure 4 is from a meta-analysis described in<sup>13</sup>. The genomic variants in group B were simulated by sampling 50 Kb segments from the breast cancer whole-genome data, and the variants in group A were simulated using the same protocol as Figure 1, with  $w$  equal to 1.

3. Results

We evaluated the performance of AIRS, SMIT, and SGI across a range of parameter values using a combination of simulated data and bootstrapped next-generation sequencing datasets (see Methods). In each comparison, we divide the cases into two groups, the normal group and the candidate germline group, representing individuals with and without candidate germline susceptibility variants, respectively.

To benchmark AIRS, we simulated the distribution of read counts according to the parameters in Figure 1 for the candidate germline group. For the normal group, we sampled whole-genome sequence data from the breast cancer tumor-normal pair. We evaluated two scenarios for the normal group, one with very low rates of allelic imbalance and one with relatively high levels of imbalance (see Figure 1). When the level of allelic imbalance in the normal group is low, preferential allelic imbalance in the candidate germline group is easier to detect, but AIRS performs well in both scenarios when 40 or more individuals are included in the candidate germline group or when the proportion imbalanced reads for the preferred allele is high. For example, with complete amplification of the preferred allele, AIRS has approximately 99% power

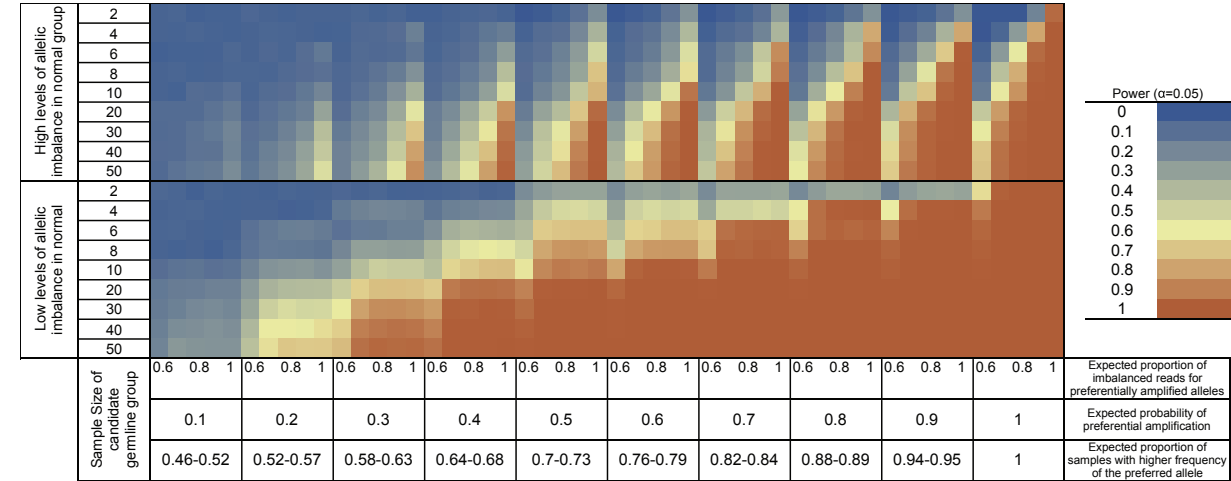


Figure 1. Power of AIRS to detect preferential allelic imbalance with  $\alpha$  of 0.05. Sample size for group A ranged from 2 to 50. Sample size for group B was 200. The expected proportion of preferential amplification summarizes information about both the proportion of true positive germline susceptibility variants and the proportion of true positives that are preferentially amplified.

at the significance level ( $\alpha$ ) of 0.05 with a sample size of only two individuals in the candidate germline group for both scenarios. Because AIRS is designed to detect preferential allelic imbalance, it cannot be used to search for genes that tend to follow a one-hit model.

The performance of SMIT depends heavily on the frequency of somatic mutations in the gene. When the mutation frequency is high in the normal group (e.g. 0.5 for *APC* and colorectal cancer)<sup>11</sup>, SMIT can detect both relative increases and decreases in the candidate germline group (Figure 2). In contrast, when the mutation frequency is very low in the normal group, SMIT can only detect mutation frequency increases in the candidate germline group. Thus, genes that follow a one-hit model can only be detected if somatic mutations are common or if the sample sizes are large. In contrast, genes that strictly follow a 2-hit model can be detected with nearly 100% power at  $\alpha = 0.05$  with sample sizes of just 10 individuals in the candidate germline group, although the detection of subtle increases in mutation frequency require substantially larger sample sizes.

SMIT is designed to control for inter-sample variation in background mutation rates between samples, which can vary by three orders of magnitude<sup>10</sup>. Systematic differences in background mutation rates between the candidate germline group and control group can result from random sampling or differences in sample collection strategies. To investigate this problem, we performed simulations with identical somatic driver mutation rates but highly differentiated background somatic mutation rates between the candidate germline group and the control group. We found SMIT properly controlled for Type I error (Figure 3A). In comparison, a Fisher exact test (e.g. CCT in MuSiC<sup>31</sup>) exhibited a highly inflated Type I error rate (Figure 3B).

SGI is designed to interface with VAAST to increase the power of a rare variant association study by combining case-control and tumor-normal pair sequence data. To demonstrate the utility of this approach, we analyzed a breast cancer case-control sequencing dataset of the gene *ATM* in VAAST, and then applied SGI to evaluate the potential change in performance. We set the number of individuals in the candidate germline group equal to the number of individual variants that had a positive VAAST score from the *ATM* results. We set the frequency of somatic mutations in the normal group to 5%, which is the reported frequency of *ATM* mutations in basal-like breast cancer<sup>9</sup>. For the candidate germline group, we varied the frequency of somatic mutations in the candidate germline group from 0 to 0.5 and set the frequency of preferential allelic imbalance

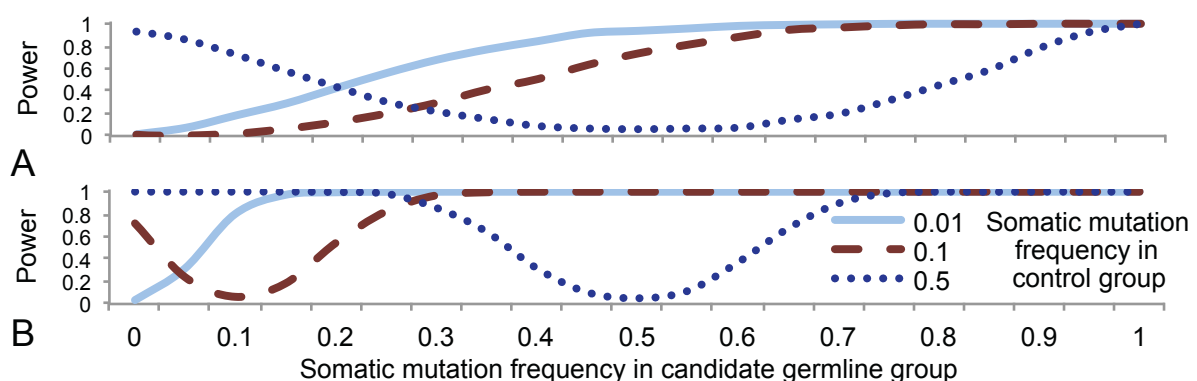


Figure 2. Power of SMIT to detect statistical interaction between germline variation and somatic small indel and point mutations at  $\alpha = 0.05$ . **A)** Sample size of group A is 10 individuals, and **B)** sample size of group A is 100 individuals. Sample size for group B was 200.

equal to the somatic mutation frequency. For each individual, preferential allelic imbalance and somatic mutation were mutually exclusive events. Figure 4 reports the sample size needed to achieve 80% power using SGI alone (4A-4B) and in combined VAAST-SGI analyses (4C). When somatic mutational events are common, combining VAAST with SGI can result in dramatic reductions in required sample sizes.

#### 4. Discussion

SGI incorporates several measures to avoid artifactual findings that can result from studies of somatic mutational events due to heterogeneity in the mutational process<sup>10</sup>. Because all comparisons are restricted to the same genomic regions, we avoid issues resulting from regional variation in mutation rates across the genome, which is the most critical source of mutational heterogeneity<sup>10</sup>. The transformation of binomial probabilities to empirical probabilities for each individual in the AIRS test allows subtle signals from low purity tumor samples to be combined with stronger signals from pure tumor samples while preserving power and controlling for inter-sample variation in genome-wide levels of allelic imbalance. AIRS is comparable to the Amplification Distortion Test (ADT) in that both tests are designed to detect preferential allelic imbalance, with AIRS designed for next-generation sequence data and ADT designed for high-density SNP microarray data<sup>33</sup>. SMIT tests for differences in the frequency of somatic mutational events between two groups at the same locus. SMIT performs the same role as the CCT test in MuSiC<sup>31</sup>, but additionally controls for inter-sample variation by incorporating sample-specific background mutation rates.

The tests we present here are well powered for a broad range of realistic scenarios. Studies of preferential allelic imbalance have reported the proportion of samples with higher frequency of the preferred allele of over 60% in colorectal cancer for a common susceptibility SNP at 8q24.21<sup>21</sup>, 70% in colorectal cancer tumors for a familial susceptibility variant in *AURKA*<sup>20</sup>, over 80% in glioblastoma for common susceptibility SNPs in the *LHFPL3* gene<sup>15</sup>, and 80%, 90%, and 100%, respectively, for skin tumor susceptibility haplotypes in *Skts6*, *Skts1*, and *Skts2* in mice<sup>22-24</sup>. Figure

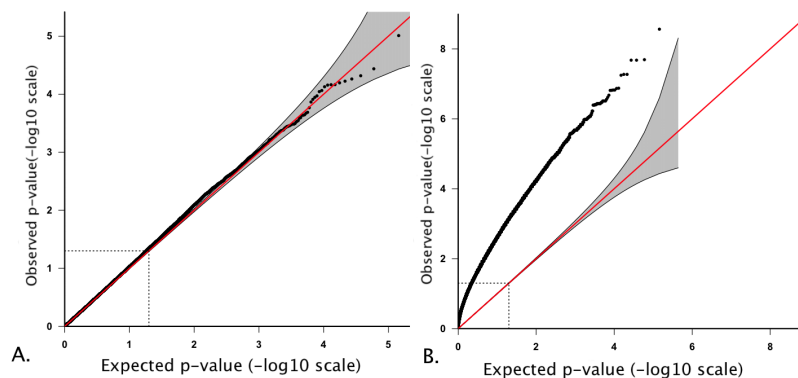


Figure 3. Observed versus expected p-values of two tests of germline-somatic interaction: **A)** SMIT and **B)** Fisher exact test (e.g. CCT<sup>31</sup>). Expected background mutation rate was 0.01 in the control group and 0.1 in the candidate germline group. Somatic driver mutation rate was 0.1 in both groups. Results generated from 100,000 simulations.

1 demonstrates that the sample sizes needed for AIRS to detect such signals are generally modest. For example, when 70% of samples have a higher frequency of the preferred allele, AIRS can detect preferential allelic imbalance with over 85% power from a sample of 20 individuals with germline susceptibility variants and a comparison group of 200 individuals. Unlike allelic imbalance, which can be detected from SNP microarray data, most somatic mutations can only be detected with sequence data, and thus, fewer studies of somatic mutation-germline interaction have been conducted. However, promising examples include a 10-fold increase in somatic mutations (from approximately 5% to approximately 50%) in a specific region of *APC* among carriers of a particular germline susceptibility variant in human colorectal cancer<sup>34</sup>, and an 88% somatic mutation rate in carriers of the *Skts2* susceptibility haplotype in mice<sup>23</sup>. Both scenarios could be detected by SMIT with greater than 80% power with a sample of only 10 individuals with candidate germline variants and a comparison group of 200 individuals (Figure 2).

The example of *ATM* and breast cancer in Figure 4 provides an illustration of how SGI can be combined with VAAST to identify novel cancer-gene associations and to yield new insights for known associations. *ATM* is not a classic two-hit tumor suppressor gene. Some rare missense germline variants have a dominant gain-of-function effect, and nonsense germline variants are reported to primarily increase the risk of breast cancer via haplo-insufficiency<sup>13,35</sup>. However, reports of rare *ATM* germline mutations and loss of the wild-type allele in tumors<sup>35</sup> is suggestive of potential germline-somatic interaction with *ATM* and breast cancer<sup>36</sup>. Figure 4 illustrates the sample size needed to conclusively detect a somatic-germline interaction effect using AIRS, SMIT, or SGI given a range of possible effect sizes. Figure 4 also demonstrates how SGI can be combined with VAAST to reduce the sample size needed to identify novel cancer-gene associations from next-generation sequence data for genes that have patterns of variation similar to *ATM* in breast cancer.

In a number of reported scenarios, preferential amplification tends to occur in conjunction with somatic mutations in a three-hit model, involving a germline susceptibility variant, a somatic point mutation on the same haplotype, and a subsequent CNV or LOH event that amplifies both the germline and somatic variants<sup>16-18,23</sup>. *JAK2* and myeloproliferative neoplasms provides one such example. Somatic mutations preferentially occur on haplotypes with germline risk variants in *JAK2* 80% of the time, and frequent third-hit somatic events result in homozygosity for both the

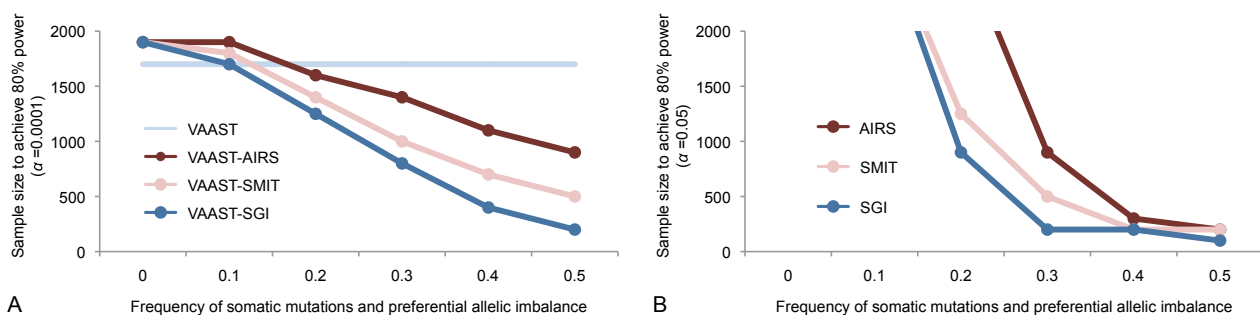


Figure 4. Sample size needed to achieve 80% power to detect **A)** cancer-germline association at  $\alpha = 1 \times 10^{-4}$  and **B)** somatic-germline interaction at  $\alpha = 0.05$ . Sample size indicates the number of controls, normal tissue samples from cases, and paired tumor tissue samples from cases.

germline risk allele and the somatic mutation<sup>16-18</sup>. One potential explanation for this three-hit model is that *cis*-acting germline variants create a hypermutable region of the gene and that the subsequent somatic mutations are then amplified by the second somatic event, after which the somatic mutation (driver) and germline variant (passenger) increase in frequency together by selection<sup>15-18</sup>. This mechanism has been demonstrated experimentally in mice<sup>23</sup>, and the T->A germline variant at *APC* nucleotide position 3920 has been reported as an example of a *cis*-acting hypermutable phenotype that leads to colorectal cancer in humans<sup>34</sup>. A second explanation for this three-hit model is that somatic mutations functionally interact in *cis* with specific germline variants and require the presence of a germline variant to promote tumorigenesis<sup>15-18</sup>. By combining evidence for preferential allelic imbalance and the occurrence of somatic mutations, SGI is well suited for detecting genes that follow a three-hit model.

SGI is designed to detect statistical interaction between somatic mutational events and germline variation from next-generation sequence data. SGI is compatible with existing variant call formats (vcf and CG tsv) and interfaces with VAAST and other variant classifiers to identify candidate germline susceptibility variants in a high-throughput manner. The AIRS test evaluates evidence for preferential allelic imbalance from next-generation sequence data and allows for combined testing of multiple variants in a gene while controlling for inter-sample variation in tumor purity and genome-wide levels of allelic imbalance. SMIT evaluates evidence for statistical interaction between candidate germline susceptibility variants and somatic SNVs and small indels while controlling for inter-sample variation in background mutation rates. SGI combines AIRS and SMIT to provide a unified measure of statistical interaction between candidate germline susceptibility variants and the occurrence of somatic mutational events. SGI can be used to help demonstrate a causal role for candidate germline susceptibility variants or can be combined with rare-variant association tests to increase the power to identify cancer-gene associations.

## 5. Software

SGI can be found at [www.hufflab.org/software/#sgi](http://www.hufflab.org/software/#sgi) and is freely available for academic use.

## Acknowledgments

We thank Sean Tavtigian for providing the *ATM* case-control data. HH was supported by the MD Anderson Cancer Center Odyssey Program. An allocation of computer time on the UT MD Anderson Research Computing High Performance Computing (HPC) facility is gratefully acknowledged.

## References

1. A. G. Knudson, Jr., *Proc Natl Acad Sci U S A.* **68**, 820-3 (1971).
2. S. H. Friend, R. Bernards, S. Rogelj, R. A. Weinberg, et al., *Nature.* **323**, 643-6 (1986).
3. W. H. Lee, R. Bookstein, F. Hong, L. J. Young, et al., *Science.* **235**, 1394-9 (1987).
4. D. Malkin, *Genes Cancer.* **2**, 475-84 (2011).
5. T. A. King, W. Li, E. Brogi, C. J. Yee, et al., *Ann Surg Oncol.* **14**, 2510-8 (2007).

6. C. J. Ceol, D. Pellman and L. I. Zon, *Nat Med.* **13**, 1286-7 (2007).
7. A. Tighe, V. L. Johnson and S. S. Taylor, *J Cell Sci.* **117**, 6339-53 (2004).
8. J. Soh, N. Okumura, W. W. Lockwood, H. Yamamoto, et al., *PLoS One.* **4**, e7464 (2009).
9. *Nature.* **490**, 61-70 (2012).
10. M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, et al., *Nature.* (2013).
11. *Nature.* **487**, 330-7 (2012).
12. H. Hu, Huff, C.D., Moore, B., Flygare, S., Reese, M.G., Yandell, M, *Genetic Epidemiology.* (in press).
13. S. V. Tavtigian, P. J. Oefner, D. Babikyan, A. Hartmann, et al., *Am J Hum Genet.* **85**, 427-46 (2009).
14. F. Le Calvez-Kelm, F. Lesueur, F. Damiola, M. Vallee, et al., *Breast Cancer Res.* **13**, R6 (2011).
15. T. LaFramboise, N. Dewal, K. Wilkins, I. Pe'er, et al., *PLoS Genet.* **6**, e1001086 (2010).
16. D. Olcaydu, A. Harutyunyan, R. Jager, T. Berg, et al., *Nat Genet.* **41**, 450-4 (2009).
17. O. Kilpivaara, S. Mukherjee, A. M. Schram, M. Wadleigh, et al., *Nat Genet.* **41**, 455-9 (2009).
18. A. V. Jones, A. Chase, R. T. Silver, D. Oscier, et al., *Nat Genet.* **41**, 446-9 (2009).
19. M. Yandell, C. Huff, H. Hu, M. Singleton, et al., *Genome Res.* **21**, 1529-42 (2011).
20. T. Hienonen, R. Salovaara, J. P. Mecklin, H. Jarvinen, et al., *Int J Cancer.* **118**, 505-8 (2006).
21. S. Tuupanen, I. Niittymaki, K. Nousiainen, S. Vanharanta, et al., *Cancer Res.* **68**, 14-7 (2008).
22. J. P. de Koning, Y. Wakabayashi, H. Nagase, J. H. Mao, et al., *Oncogene.* **26**, 4171-8 (2007).
23. H. Nagase, J. H. Mao and A. Balmain, *Cancer Res.* **63**, 4849-53 (2003).
24. A. Ewart-Toland, P. Briassouli, J. P. de Koning, J. H. Mao, et al., *Nat Genet.* **34**, 403-12 (2003).
25. P. C. Ng and S. Henikoff, *Annu Rev Genomics Hum Genet.* **7**, 61-80 (2006).
26. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, et al., *Nat Methods.* **7**, 248-9 (2010).
27. S. V. Tavtigian, A. M. Deffenbaugh, L. Yin, T. Judkins, et al., *J Med Genet.* **43**, 295-305 (2006).
28. E. Mathe, M. Olivier, S. Kato, C. Ishioka, et al., *Nucleic Acids Res.* **34**, 1317-25 (2006).
29. N. Dewal, Y. Hu, M. L. Freedman, T. Laframboise, et al., *Genome Res.* **22**, 362-74 (2012).
30. S. Vattathil and P. Scheet, *Genome Res.* **23**, 152-8 (2013).
31. N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, et al., *Genome Res.* **22**, 1589-98 (2012).
32. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, et al., *Science.* **327**, 78-81 (2010).
33. N. Dewal, M. L. Freedman, T. LaFramboise and I. Pe'er, *Bioinformatics.* **26**, 518-28 (2010).
34. S. J. Laken, G. M. Petersen, S. B. Gruber, C. Oddoux, et al., *Nat Genet.* **17**, 79-83 (1997).
35. J. O. Bay, N. Uhrhammer, D. Pernin, N. Presneau, et al., *Hum Mutat.* **14**, 485-92 (1999).
36. C. Schaffner, I. Idler, S. Stilgenbauer, H. Dohner, et al., *Proc Natl Acad Sci U S A.* **97**, 2773-8 (2000).

## SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA<sup>\*</sup>

IN SOCK JANG<sup>1</sup>, ELIAS CHAIBUB NETO, JUSTIN GUINNEY, STEPHEN H. FRIEND, ADAM A. MARGOLIN<sup>1</sup>

Sage Bionetworks

1100 Fairview Ave. N Seattle, WA 98109, USA

Email: [in.sock.jang@sagebase.org](mailto:in.sock.jang@sagebase.org)

Email: [elias.chaibub.neto@sagebase.org](mailto:elias.chaibub.neto@sagebase.org)

Email: [justin.guinney@sagebase.org](mailto:justin.guinney@sagebase.org)

Email: [friend@sagebase.org](mailto:friend@sagebase.org)

Email: [margolin@sagebase.org](mailto:margolin@sagebase.org)

Large-scale pharmacogenomic screens of cancer cell lines have emerged as an attractive pre-clinical system for identifying tumor genetic subtypes with selective sensitivity to targeted therapeutic strategies. Application of modern machine learning approaches to pharmacogenomic datasets have demonstrated the ability to infer genomic predictors of compound sensitivity. Such modeling approaches entail many analytical design choices; however, a systematic study evaluating the relative performance attributable to each design choice is not yet available. In this work, we evaluated over 110,000 different models, based on a multifactorial experimental design testing systematic combinations of modeling factors within several categories of modeling choices, including: type of algorithm, type of molecular feature data, compound being predicted, method of summarizing compound sensitivity values, and whether predictions are based on discretized or continuous response values. Our results suggest that model input data (type of molecular features and choice of compound) are the primary factors explaining model performance, followed by choice of algorithm. Our results also provide a statistically principled set of recommended modeling guidelines, including: using elastic net or ridge regression with input features from all genomic profiling platforms, most importantly, gene expression features, to predict continuous-valued sensitivity scores summarized using the area under the dose response curve, with pathway targeted compounds most likely to yield the most accurate predictors. In addition, our study provides a publicly available resource of all modeling results, an open source code base, and experimental design for researchers throughout the community to build on our results and assess novel methodologies or applications in related predictive modeling problems.

Keywords: Cancer cell lines, pharmacogenomics, machine learning, predictive modeling.

### 1. Introduction

Molecular analysis of cancer has revealed that tumor subtypes differ in pathway activity, progression, and chemotherapeutic response, leading to the development of therapeutic approaches with demonstrated efficacy in molecularly defined cancer subtypes [1-4]. Human cancer cell lines represent an attractive pre-clinical system for identifying molecular characteristics of tumors predictive of therapeutic response.

Recently, two ambitious initiatives, named the cancer cell line encyclopedia [5, 6] and the genomics of drug sensitivity projects [7] have performed large-scale small molecule screens on

---

<sup>\*</sup> Work supported by grant U54CA149237 from the Integrative Cancer Biology Program of the National Cancer Institute.

<sup>1</sup> corresponding authors



panels of hundreds of molecularly characterized cancer cell lines. Both studies also demonstrated that employing modern machine learning algorithms to develop predictors of drug response based on molecular profiling measurements of each tumor could effectively identify known pharmacogenomic predictive biomarkers. These proof-of-concept studies have established cell line-based screens as a viable pre-clinical system for identifying functional biomarkers underlying drug sensitivity or resistance and for suggesting patient selection strategies for clinical trial design.

As computational approaches for modeling therapeutic response become increasingly common in research and translational applications, a study is warranted to systematically assess different modeling approaches, and recommend best practices for future applications. To address this question, we defined important categories of modeling choices, such as the predictive algorithm and genomic features for model inclusion (among others), and performed a large multifactorial experiment with crossed factors, where the modeling choices represent the experimental factors, and the predictive performance measures (derived from model fits, and spanning all possible combinations of modeling choices) represent the response data. This experimental design allows for formal statistical testing and quantification of the relative importance of the modeling choices.

Our results provide statistically principled, data-driven guidelines for best-in-class modeling practices. Our findings suggest the use of elastic net or ridge regression applied to continuous valued response data, summarized using the area under the fitted dose response curve, and using all molecular features (in particular, gene expression data). Moreover, our results suggest that pathway targeted compounds lead to more accurate predictors than classical broadly cytotoxic chemotherapies. In addition, we performed detailed analysis comparing models based on continuous versus discretized response measurements, suggesting that discretizing data (e.g. into sensitive and resistant calls) causes decreased model accuracy. Finally, we report a discordance in reported values across the 2 datasets for the same compounds and suggest that raw dose-response data should be made publicly available to facilitate comparison of the 2 datasets based on the same procedures for processing and summarizing dose-response values.

Our study provides a publicly available interactive resource of modeling results and an open source analysis package. The results for all >110,000 models are available at (<https://www.synapse.org/#!Synapse:syn2009053>), providing a resource for other researchers to interactively browse the results of all models and perform additional downstream analyses. Moreover, we are releasing the open source “predictiveModeling” R package ([https://github.com/Sage-Bionetworks/PredictiveModel\\_pipeline](https://github.com/Sage-Bionetworks/PredictiveModel_pipeline) and <https://github.com/Sage-Bionetworks/predictiveModeling>), containing all code used to infer models in this study, and providing a modular API that may be extended by the community and used to conduct similar research studies.

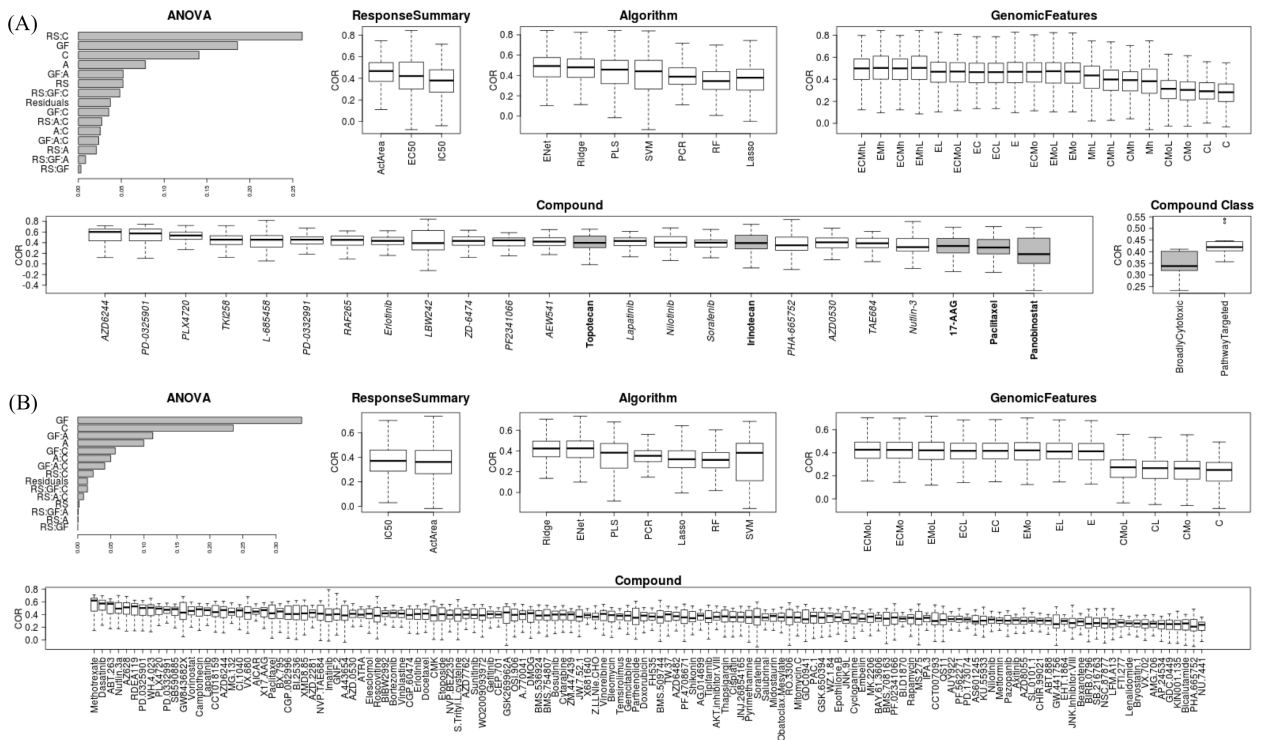
## **2. Material and Methods**

### **2.1 Data Sets**

The CCLE and Sanger datasets contain compound screening data performed on large panels of molecularly characterized cancer cell lines. Both datasets contain genome-wide gene expression and copy number profiling, as well as sequencing data on a subset of genes (described in the next section). Gene expression, copy number, and mutation data were summarized to gene-level features. The Sanger panel is composed of 30,672 genomic features and 138 compounds profiled



on 714 cell lines (535 cell lines contain all measurement types). The CCLE panel is composed of 41,814 genomic features and 24 compounds profiled on 504 cell lines (411 cell lines contain all measurement types). All data was normalized as described in the original papers [5-9]. Mutation data was summarized to binary gene-level variables represented as 0 (wild type) and 1 (mutation). We also annotated each cell line with a representative “tumor type” label, derived by manually curating the provided meta-data annotations. Each tumor type was then included as a binary feature variable.

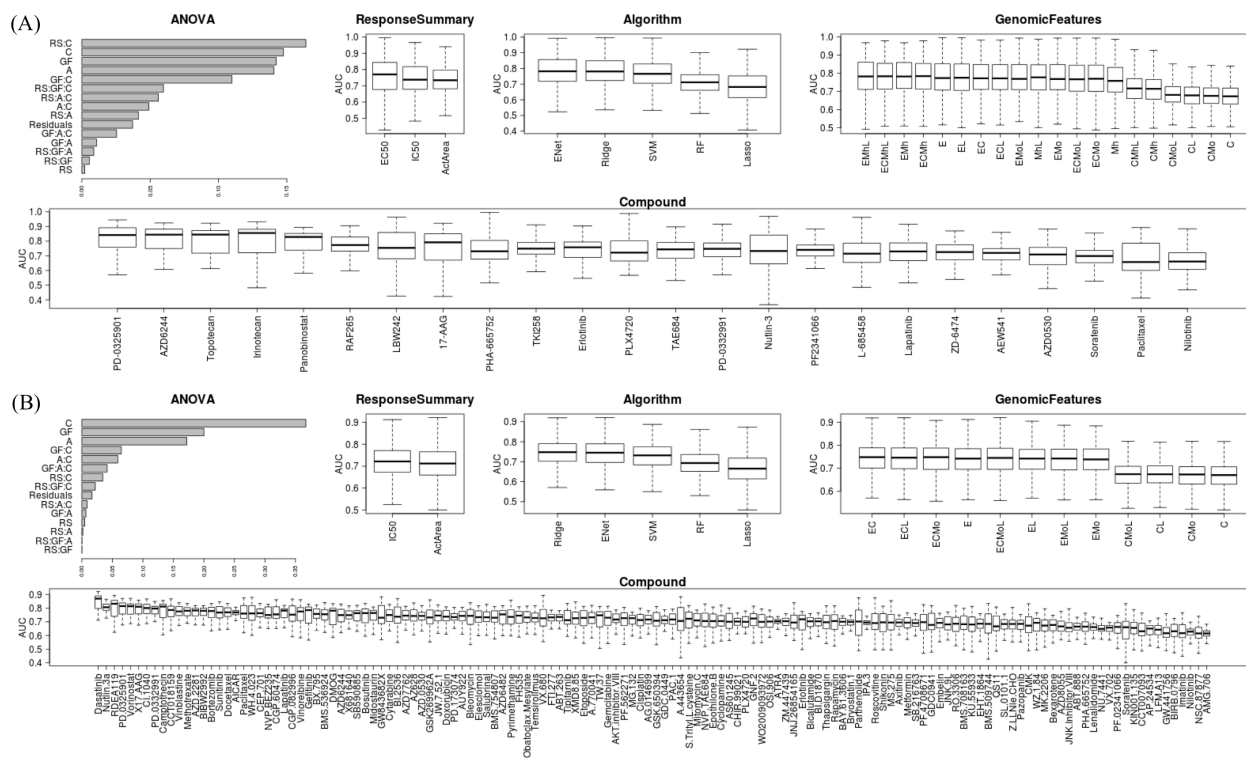


**Figure 1 – Summary of evaluation of regression models.** (A) Results for CCLE. (B) Results for Sanger. The left panel displays the percent variance of predictive accuracy (COR) explained by each category of modeling choice after fitting our 4-way ANOVA model. The panels labeled *Compound*, *ResponseSummary*, *Algorithm*, and *GenomicFeatures* correspond to each of our tested categories of modeling choices, and display the distribution of predictive performance (COR) scores for each modeling choice (factor levels) within the category. For the CCLE *Compound* panel, compounds classified as “BroadlyCytotoxic” are displayed as shaded boxes and bold text, and compounds classified as “PathwayTargeted” are displayed as white boxes and non-bold text. The panel titled *Compound Class* displays the distribution of predictive performance scores for the BroadlyCytotoxic vs. PathwayTargeted compound classes.

Both studies provide multiple statistics used to summarize dose-response curves to compound sensitivity values for each cell line (described in the next section). We used the summarized sensitivity values reported in each dataset, as raw dose-response values were not available to process both datasets using the same procedures.

## 2.2 Definition of modeling choices

Our goal was to systematically assess the effect of modeling choices on predictive performance given a *drug response vector* and a *molecular feature matrix*. We enumerated the following 5 categories of modeling choices, as well as the possible choices of modeling factors within each category



**Figure 2 – Summary of evaluation of classification methods.** (A) Results for CCLE. (B) Results for Sanger. Results are presented as described in **Figure 1**, based on evaluation of classification models using the AUC predictive performance statistic.

**GenomicFeatures:** Represent the distinct data types used as features in the predictive algorithms. In Sanger we have 4 distinct types: gene expression measurements (E) on 12,024 genes; copy number variation measurements (C) on 18,601 genes; cell line tumor type classifications (L) according to 93 distinct tumor lineages; and mutation profiling (Mo) on 47 genes. We tested 12 distinct data type combinations as shown in the *GenomicFeatures* panels in Figure 1B and Figure 2B (specifically, we tested 12 combinations other than those corresponding to small feature sets, such as L+Mo). For the CCLE panel we have 5 distinct data types: gene expression measurements (E) on 18,897 genes; copy number measurements (C) on 21,217 genes; cell line tumor type classifications (L) of 97 tumor lineages; mutation profiling (Mo) on 33 genes using the oncomap 3.0 platform [10]; and mutation profiling of 1,667 genes using hybrid capture sequencing (Mh). We tested 20 distinct data type combinations shown in the *GenomicFeatures* panels in Figure 1A and Figure 2A.

**Compound:** Represents the anti-cancer compounds screened by the cell line projects. There are 138 compounds in Sanger and 24 in CCLE.

**ResponseSummary:** Represents the statistic used to summarize the dose response curves to a single number, corresponding to the degree of sensitivity of a given cell line to a given compound. For Sanger, the choices are: AUC – the area under the fitted dose response curve; IC50 – the concentration at which the compound reaches 50% reduction in cell viability. For CCLE, the choices are: ActArea – the area above the fitted dose response curve (inverse measure of AUC in Sanger); IC50 – the same as in Sanger; EC50 – the concentration at which the compound reaches 50% of its maximum reduction in cell viability. We note that although they use the same terminology, both studies used different procedures for fitting dose response curves and generating summary statistics.

**Continuous vs. categorical models:** Whether predictions are made based on continuous or discretized *ResponseSummary* measurements. We tested multiple discretization schemes, including: mean and median based deviation statistics; Gaussian mixture models; and upper/lower third quartile thresholds. We report results based on upper/lower third quartile thresholds, which was the discretization scheme that achieved the highest average classification accuracy (AUC).

**Algorithm:** Represents the predictive algorithms compared in this study. In the analysis of continuous response variables, we compared: principal component regression (PCR); partial least square regression (PLS); least squares support vector machine regression with linear kernels (SVM); random forests (RF); least absolute shrinkage and selection operator (LASSO); ridge regression (RIDGE); and elastic net regression (ENet) [11-19, 27]. For the analysis of binary response variables, we considered: least squares support vector machine classification with linear kernels (SVM); random forests (RF); binomial least absolute shrinkage and selection operator (LASSO); ridge binomial regression (RIDGE); and elastic-net binomial regression (ENet) [8, 11, 12, 14, 15, 20].

### 2.3 Model fitting procedures

We employed a multifactorial experimental design and tested all combinations of modeling choices (e.g. the cross product of all choices of *ResponseSummary*  $\times$  *Compound*  $\times$  *GenomicFeatures*  $\times$  *Algorithm*  $\times$  *Discretization*, excluding application PCR and PLS to discrete data). This resulted in testing a total of 114,048 models.

For Sanger and CCLE the input dataset was divided into five non-overlapping sample groups, used as cross-validation folds for training and testing data. For each cross-validation fold, each model was trained on 4/5<sup>th</sup> of the samples, and used to make predictions of sensitivity for the held out 1/5<sup>th</sup> of samples. Within each training step, a separate 5-fold cross-validation procedure was employed for parameter tuning of each model.

Predicted vs. observed response vectors were compared to assess the performance of each algorithm. The predicted response vector was computed by concatenating the prediction vectors for each cross-validation fold. For continuous models we computed the Pearson correlation coefficient (COR). For discrete models we computed area under the receiver operating characteristics curves (AUC).

### 2.4 Statistical Analysis

We evaluated the effect of modeling choices on predictive performance using multiway-ANOVA with crossed factors. For instance, in the analysis of continuous models in the CCLE panel, we adopted COR as the response variable, and performed ANOVA using 4 factors:

*GenomicFeatures*, composed of 20 levels representing distinct data type combinations; *Compound*, composed of 24 levels, each representing one of the anti-cancer compounds tested in the CCLE panel; *ResponseSummary*, represented by levels ActArea, EC50, and IC50; and *Algorithm* represented by levels ENet, RIDGE, PLS, SVM, PCR, LASSO, and RF. For each one of the possible  $20 \times 24 \times 3 \times 7 = 10,080$  modeling choice combinations, we fit a predictive model and recorded the correlation between the observed and predicted outcome as the response variable. Since we only have a single observation per modeling choice combination, our design corresponds to a multiway-ANOVA with 4 crossed factors and a single observation per cell. Hence, we cannot fit a complete model (i.e., with all interaction terms up to order 4) and we restrict our analysis to interactions of order up to 3. In addition to the analysis described above, we also performed analogous ANOVA analyses for the evaluation of continuous models in Sanger, discrete models in CCLE, and discrete models in Sanger.

### 3. Results

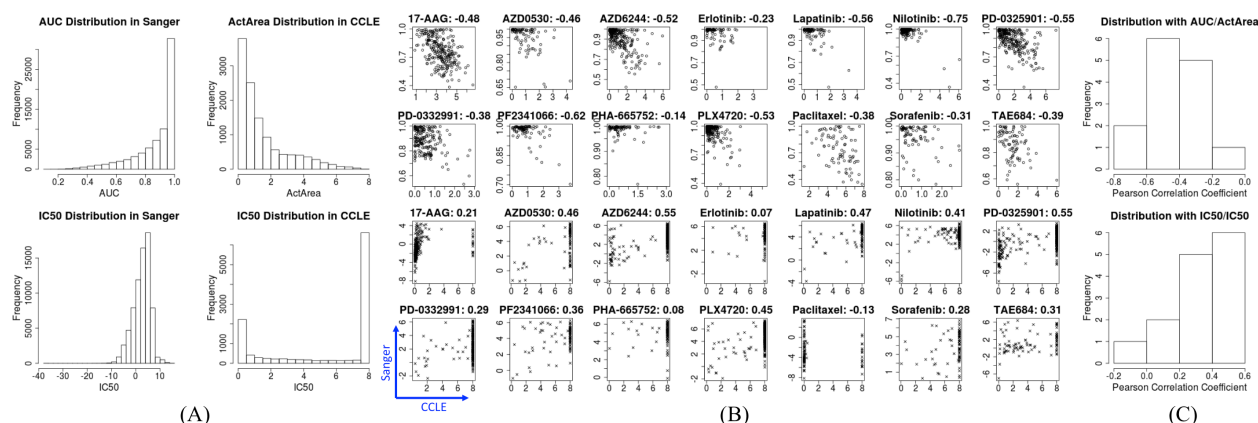
#### *Modeling factors influencing predictive performance*

In order to assess the individual contributions of each category of modeling choices (and their interactions) to explaining the total variability of the predictive performance statistic (COR or AUC), we examined the decomposition of the total sum of squares of the predictive performance variable into residual sum of squares plus sum of squares terms for each one of the factors and factor interactions in our 4way-ANOVAs, including all possible interactions of order up to 3. We first describe results for continuous models. The left panels in Figure 1A and B present barplots in which each bar represents the sum of squares of the respective term divided by the total sum of squares.

For both the CCLE and Sanger datasets, most of the variance of predictive accuracy is explained by the modeling factors considered in our study, as indicated by the small percent of variance attributable to residuals. For both CCLE and Sanger the modeling factors explaining the highest percent variance are: 1) the type of molecular features used to build the model; and 2) the compound being predicted by the model. The third most important modeling factor is the type of algorithm, although this factor is considerably less important than the first 2. This result is consistent with previous studies [21], suggesting that input data is the dominant factor related to model performance, whereas the specific modeling strategies are of secondary importance.

The CCLE dataset contains a strong interaction term between *Compound* and *ResponseSummary*, suggesting that model performance depends both on the compound being modeled, and the ability to summarize the compound's dose response measurements. By contrast, *ResponseSummary* has negligible effect in the Sanger dataset. We point out that, although Sanger and CCLE both report response data in terms of IC50 and AUC (referred to as ActArea in CCLE) summarizations, the 2 studies use quite different procedures for fitting dose response curve and summarizing them to IC50 or AUC statistics. The discordant importance of the *ResponseSummary* factor between the 2 studies, compared with the highly concordant importance of all other factors, suggests that the procedures for summarizing dose response curves to summary statistics may be inconsistent between the 2 studies. Indeed, comparison of IC50 and AUC values for compounds profiled in both datasets suggests a relatively high degree of inconsistency (Figure 3). Unfortunately, raw dose response data used for curve fitting is not available in either study,

limiting our ability to investigate this issue further. This result highlights the importance of making raw forms of data publicly available, in addition to computed summary statistics, such that the community may more transparently analyze and improve the value of the data resource.



**Figure 3 – Comparison of IC50 and AUC summary statistics for 14 compounds and 283 cell lines in common between the Sanger and CCLE datasets.** (A) Distribution of IC50 and AUC/ActArea values in Sanger and CCLE. Note that the AUC value reported in Sanger corresponds to the area under the dose response curve in which values of 0 correspond to complete reduction in cell viability and values of 1 correspond to no reduction in cell viability. The ActArea value reported in CCLE corresponds to the area over the dose response curve in which values of -100 correspond to complete reduction in cell viability and values of 0 correspond to no reduction in cell viability. Therefore a negative correlation is expected between AUC and ActArea values. (B) Scatter plots comparing AUC/ActArea values (top) and IC50 values (bottom) across the 2 studies. (C) Histograms of the distribution of correlations across the 2 studies for the 14 common compounds based on ActArea/AUC (top) and IC50 (bottom).

### Assessment of best performing modeling strategies

The ANOVA analysis detected highly significant interaction and main effects in explaining predictive performance, indicating the importance of some modeling choices over others. Figure 1 and Figure 2 depict boxplot panels for each one of the modeling choice factors in our analyses, showing the distribution of predictive performance as a function of the modeling factor levels. For both datasets, expression data was the most informative molecular feature type, as all of the best performing models included use of expression data. Models using other molecular features types in addition to expression data performed slightly better than using expression data alone, although performance improvements were modest. For both datasets, elastic net and ridge regression were the top performing algorithms. For the CCLE dataset, summarizing dose response values based on ActArea achieved the highest performance. For Sanger, response summarization had little effect on model performance, warranting closer investigation starting from raw dose response data.

For both datasets, some compounds were easier to predict than others, as clearly shown by the *Compound* panels in Figure 1. Inspection of predictability scores for CCLE compounds suggested a general trend. Compounds with low predictability scores tended to be more classical chemotherapeutics that disrupt broad cellular processes (e.g. topoisomerase inhibitors). Compounds with high predictability scores tended to target proteins in specific pathways, primarily related to mitogen signaling (e.g. MEK inhibitors). To test this hypothesis, we manually annotated each compound in one of these 2 classes, which we called “BroadlyCytotoxic (BC)” and

“PathwayTargeted (PT)”. Indeed, PT compounds displayed significantly higher predictability scores compared to BC ( $P=0.003529$  by Wilcoxon rank sum test, as shown in the top-right panel of Figure 1).

### *Assessment of categorical models*

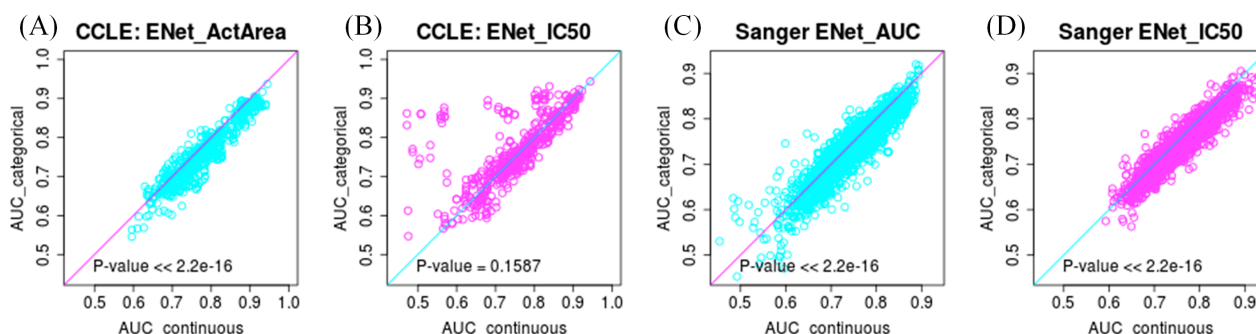
An alternative strategy to modeling the drug response as a continuous-valued variable is to discretize the response vector into a binarized “sensitive” and “resistant” vector. To evaluate this strategy, we implemented the categorical analogues of lasso, ridge, elastic net, random forests, and support vector machines, and discretized each response summarization (IC50, EC50, AUC or ActArea) base on the upper and lower third quartiles.

Results from this analysis were highly consistent with results from our continuous models (Figure 2). For both CCLE and Sanger, the relative importance of model factors was consistent with results for continuous models (e.g. *GenomicFeatures* and *Compound* being most important, followed by *Algorithm*). The relative performance of modeling choices was also consistent between categorical and continuous models (e.g. the order of predictive performance of algorithms is fully consistent).

One advantage of categorical models is the ability to interpret AUC values as the probability of correctly classifying a new sample as sensitive or resistant. For example, analysis of the distribution of AUC scores suggests that sensitive vs. resistant samples can be classified with >70% accuracy for 22 of 24 (91.7%) compounds in CCLE and 83 of 138 (60.1%) compounds in Sanger. More specific analysis of the AUC curves can be used to determine the expected trade-offs between false positives and false negatives. We suggest that such analysis may be useful in assessing the potential clinical utility of a predictive model, for example, by applying criteria such as requiring less than a 5% false positive rate (e.g. correctly prescribing a drug to 95% of patients who might benefit) at the expense of a less than 20% false negative rate (e.g. failing to prescribe the drug to 20% of the patients who will benefit from it). Of course, such statistics derived from cell line studies are unlikely to directly translate in a clinical context, but may be useful to identify predictive models that should be prioritized for further clinical studies.

### *Comparison of continuous vs. categorical models*

In order to directly compare the performance of continuous vs. categorical models, we computed the AUC scores of the rank-ordered predictions in comparison to the discretized response data. That is, we calculated the sensitivity and specificity at each threshold of the rank-ordered predictions in order to compute an ROC curve for each model. We based our comparison on the best performing regression and classification methods, which was elastic net in both cases (results were similar for other methods). In general, regression models, trained using continuous *ResponseSummary* values, outperformed classification models, trained using discretized *ResponseSummary* values ( $P < 2.2 \times 10^{-16}$  for Sanger, based on AUC;  $P < 2.2 \times 10^{-16}$  for Sanger, based on IC50;  $P < 2.2 \times 10^{-16}$  for CCLE, based on ActArea;  $P=0.1587$  for CCLE, based on IC50. See Figure 4. Classification methods outperformed regression methods only when using the CCLE IC50 values, as explained by the fact that these values are inherently discretized. Sanger IC50 values utilized extrapolations of the curve fits beyond the tested concentration range. By contrast, out of 11,670 IC50 values reported in CCLE (426 excluding NA values), 6,499 (55.69%) were set to a value of 8, corresponding to the maximum tested compound dose of 8  $\mu$ M (Figure 3A).



**Figure 4 – Comparison of predictive performance of continuous (regression) vs. categorical (classification) models.** Results were compared for the continuous and categorical versions of elastic net, which were the best performing continuous and categorical models. (A) CCLE data with ActArea, (B) CCLE data with IC50, (C) Sanger data with AUC, and (D) Sanger data with IC50.

#### 4. Discussion

As large-scale complex genomic resources become increasingly available, there is a pressing need to develop community standards and robust assessment methods to determine the best performing approaches for analyzing such data. Pharmacogenomic screens performed on genomically characterized cancer cell lines provide rich data resources, and application of machine learning methodologies to such data have demonstrated evidence of uncovering genomic mechanisms underlying drug response.

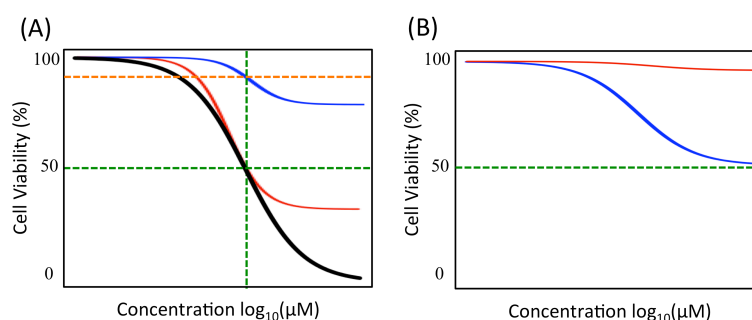
From an analytical perspective, such pharmacogenomic data resources are particularly well suited to application of statistical learning methods by representing genomic and compound sensitivity data, respectively, as predictive features and response variables in a supervised learning scheme. In this study, we performed a controlled analysis of many modeling choices that may be used in this application. We believe this work contributes to the community in 3 ways: 1) by providing a set of recommended best practices for inferring pharmacogenomic predictive models, and a study on the relative importance of each; 2) by establishing a resource of over 110,000 modeling results, providing a baseline set of scores that researchers may use in future studies to demonstrate improved performance of novel methodologies; 3) by providing an experimental design template, and open source modeling package, that can be extended for use in other predictive modeling applications.

Our study suggests a statistically principled set of recommended best modeling practices: using **elastic net or ridge regression** with input features from all genomic profiling platforms, most importantly, **gene expression features**, to predict **continuous-valued** sensitivity scores summarized using the **area under/over the dose response curve**, with **pathway targeted** compounds will most likely yield the most accurate predictors.

The use of elastic net regression is consistent with modeling choices reported in previous studies [14][16], and is a particularly attractive option due to the ability to perform feature selection based on inferred feature weights. We investigated several methods that have previously been shown to achieve superior predictive accuracy, but lead to less interpretable models, such as support vector machines, random forests, and principal components regression [22, 23].

Nonetheless elastic net regression achieves the highest predictive accuracy without requiring a trade-off of model interpretability. Moreover, elastic net is designed to seek the optimal trade-off of model complexity penalties imposed by lasso and ridge regression. While the sparse feature selection encouraged by lasso indeed leads to inferior predictive performance, elastic net performs as well as ridge regression based on predictive accuracy, suggesting that elastic net effectively balances the strengths of the two methods by encouraging sparser models without compromising predictive accuracy. We note that although we employed standard and well-accepted cross-validation schemes for parameter tuning of all models, it is possible that alternative methods could improve the performance of some models.

The observation that gene expression features provide the most informative predictors might be explained by the increased “information content” of gene expression data. In particular, copy number values are highly correlated with each other and the mutation data profiles only a small subset of genes. Although gene expression data provides advantages in predictive accuracy, genomic (e.g. somatic mutation and copy number) data possess advantages in potential translation to clinical biomarkers. From a technical standpoint, the increased molecular stability of DNA compared to RNA facilitates easier development of clinical assays, even from archival samples. Perhaps more importantly, features derived from genomic data are more likely to correspond to functional driver events related to drug sensitivity, whereas features derived from gene expression may be correlative, rather than causal, biomarkers. Thus genomic features are more likely amenable to functional validation experiments, such as testing if knockdown or overexpression of predicted functional biomarkers confers the predicted suppression or enhancement of sensitivity. By extension, genomic predictors of drug resistance may suggest targets for combination therapies [24].



**Figure 5 – Illustration of differences in dose response curves not captured by IC50 or EC50 statistics.** (A) The red curve and black curve achieve 50% reduction in cell viability at the same compound concentration, but the black curve achieves increased reduction in cell viability at higher compound concentrations. Both curves correspond to the same IC50 value (vertical dotted green line), while the area under the dose response curve (AUC) captures the increased sensitivity shown in the black curve. The blue curve illustrates a sample with limited maximal reduction in cell viability at high compound concentrations. The EC50 statistic would be the same for the blue and black curves (vertical dotted green line), while the AUC statistic captures the increased response of the black curve. (B) The red and blue curves fail to reach 50% reduction in viability within the tested concentration range. The IC50 statistic would be set to the maximum tested concentration in both cases (or extrapolated outside the tested range), while the AUC statistic naturally captures the increased sensitivity displayed in the blue curve.



We also investigated alternative methods of assigning a summary statistic representing the sensitivity of a given cell line to a given compound. Predictive accuracy was improved by computing the area under/over the dose response curve (AUC/ActArea), as opposed to the more traditional metric of IC50. Following the theme described above, we suggest that AUC/ActArea captures more information from the experiment than IC50. Specifically, IC50 assumes a canonical sigmoidal shape of dose response curves, with zero growth inhibition in the absence of compound and 100% growth inhibition at high compound doses. This assumption fails to differentiate samples that achieve 50% growth inhibition at the same dose, even if one of the samples achieves far higher growth inhibition at higher doses (Figure 5A). An alternative statistic, EC50, is designed to account for this situation by computing the concentration at which a sample achieves 50% of its maximal growth inhibition; but this statistic suffers from additional degeneracies. Moreover, many samples do not achieve 50% growth inhibition within the tested dose range (Figure 5B). Therefore, IC50 calculations must set all such cases to a single threshold value (e.g. the highest tested dose, as reported for CCLE), or attempt to extrapolate based on fitted curves (as reported for Sanger). By contrast, the AUC/ActArea statistic is able to discriminate the examples listed above, and captures additional information contained in the dose response curves related to differential sensitivity (see Figure 5).

Our observation that continuous regression models, in general, outperform discrete classification models also follows the general theme of using data with the maximal amount of information as model inputs. Discretization of sensitivity data reduces the amount of information contained in the continuous valued data. Such a trade-off may be desirable if discretization reduces noise in the data (e.g. by only modeling the tails of the data, which are more likely to correspond to true differences in sensitivity and resistance, while ignoring the noisy intermediate values). Although this argument may apply in selective cases, it is highly dependent on choosing an accurate discretization scheme. We investigated several alternatives, including mixture models and mean and median-based deviation statistics (not shown). We observed that each scheme worked in some cases but not others; e.g. deviation-based statistics may classify no samples as sensitive or resistant for some compounds, while quartile-based statistics do not capture variable numbers of samples that may be sensitive to different compounds.

In addition to assessing the performance of modeling choices within our evaluated categories, we also assessed the relative importance of the categories themselves. Consistent with previous studies [21], our general conclusion is that the choice of input data (which molecular features are used and which compound is being predicted) dominates in explaining the high or low accuracy of a model. The choice of modeling algorithm also matters, but far less than the input data. While this conclusion may be sobering for data analysts (such as ourselves) in pursuit of the next great algorithm, we point out that our study was limited to machine learning methods designed to operate on specified feature and response data. Thus we suggest that optimization of methodologies in this context are unlikely to achieve dramatic improvements over current state-of-the-art methods; however, methodologies that incorporate additional information sources, such as other large-scale genomic datasets or information from pathway databases, were not tested in our study and may yield such improvements. This intuition is consistent with our observation that the quality and information content of input data dominates predictive performance, as such strategies augment the amount of information used to build a predictor. Indeed, in a recent community-based assessment of genomic predictors of breast cancer survival, the best performing method integrated information from all of TCGA in addition to the dataset directly used to build predictors [25, 26].

We note that our study does not assess all possible modeling choices. For example, we utilized the normalized genomic data provided by the CCLE and Sanger resources and did not assess the impact of alternative normalization or data processing procedures. We invite researchers throughout the community to build on and improve our work to investigate the myriad of additional approaches. Indeed, we hope the resource released by our study serves as initial input to a community effort promoting critical assessment of modeling methodologies. Innovative approaches developed by any researcher may be assessed in comparison to our results, thus providing a pre-defined set of performance criteria and baseline model scores against which novel approaches may objectively demonstrate their value.

## REFERENCES

1. Ferte, C., et al., *Nature Reviews Clinical Oncology* **7**, 367-380 (2010).
2. Roche-Lestienne, C., et al., *New England Journal of Medicine* **348**, 2265-2266 (2003).
3. Peggs, K. and S. Mackinnon, *New England Journal of Medicine* **348**, 1048-1050 (2003).
4. Savage, D.G. and K.H. Antman, *New England Journal of Medicine* **346**, 683-693 (2002).
5. Barretina, J., et al., *Nature* **483**, 603-607 (2012).
6. Marum, L., *Pharmacogenomics* **13**, 740-741 (2012).
7. Garnett, M.J., et al., *Nature* **483**, 570-U87 (2012).
8. Forbes, S.A., et al., *Nucleic Acids Research* **39**, D945-D950 (2011).
9. Bindal, N., et al., *Genome Biology* **12**, 5-5 (2011.).
10. MacConaill, L.E., et al., *Plos One*, **4**, 11 (2009).
11. Jolliffe, I.T., *Journal of the Royal Statistical Society Series C* **31**, 300-303 (1982).
12. Wold, S., et al., *Chemometrics and Intelligent Laboratory Systems* **58**, 109-130 (2001).
13. Balabin, R.M. and E.I. Lomakina, *Analyst* **136**, 1703-1712(2011).
14. Wu, Y.F. and S. Krishnan, *Journal of Experimental & Theoretical Artificial Intelligence* **23**, 63-77 (2011).
15. Breiman, L., *Machine Learning* **45**, 5-32 (2001).
16. Tibshirani, R., *Journal of the Royal Statistical Society Series B* **58**, 267-288 (1996).
17. Tibshirani, R., *Journal of the Royal Statistical Society Series B* **73**, 273-282 (2011).
18. Friedman, J., et al., *Journal of Statistical Software* **33**, 1-22(2010).
19. Hoerl, A.E. and R.W. Kennard, *Technometrics* **42**, 80-86 (2000).
20. CCLE data portal. Available from: <http://www.broadinstitute.org/ccle/home>.
21. Shi, L.M., et al., *Nature Biotechnology* **28**, 827-U109 (2010).
22. Xu, C.J., et al., *Plos One* **7**, 8 (2012).
23. Niculescu-Mizil, R.C.A., *ICML2006*, 161-168 (2006).
24. Wei, G., et al., *Cancer Cell* **21**, 547-562 (2012).
25. Margolin, A.A., et al., *Science Translational Medicine*, **5**, 181 (2013).
26. Cheng, W.Y., et al, *Science Translational Medicine* **5**, 181 (2013).
27. Statnikov, A., et al, *BMC Bioinformatics* **9**, 319 (2008)

## INTEGRATIVE ANALYSIS OF TWO CELL LINES DERIVED FROM A NON-SMALL-LUNG CANCER PATIENT – A PANOMICS APPROACH

OLEG MAYBA<sup>1\*</sup>, FLORIAN GNAD<sup>1\*</sup>, MICHAEL PEYTON<sup>4\*</sup>, FAN ZHANG<sup>3</sup>, KIMBERLY WALTER<sup>2</sup>,  
PAN DU<sup>1</sup>, MELANIE A. HUNTLEY<sup>1</sup>, ZHAOSHI JIANG<sup>1</sup>, JINFENG LIU<sup>1</sup>, PETER M. HAVERTY<sup>1</sup>, ROBERT  
C. GENTLEMAN<sup>1</sup>, RUIQIANG LI<sup>3</sup>, JOHN D. MINNA<sup>4</sup>, YINGRUI LI<sup>3</sup>, DAVID S. SHAMES<sup>2</sup>,  
ZEMIN ZHANG<sup>1#</sup>

*Departments of <sup>1</sup>Bioinformatics and Computational Biology and <sup>2</sup>Development Oncology Diagnostics,  
Genentech, Inc., South San Francisco, CA 94080, USA*

*<sup>3</sup>BGI-Shenzhen, Shenzhen 518083, China*

*<sup>4</sup>Hamon Center for Therapeutic Oncology Research, UT-Southwestern Medical Center, Dallas, TX 75390, USA*

*\* These authors contributed equally to this work.*

*# To whom correspondence should be addressed: Zemin Zhang (zemin@gene.com)*

Cancer cells derived from different stages of tumor progression may exhibit distinct biological properties, as exemplified by the paired lung cancer cell lines H1993 and H2073. While H1993 was derived from chemo-naïve metastasized tumor, H2073 originated from the chemo-resistant primary tumor from the same patient and exhibits strikingly different drug response profile. To understand the underlying genetic and epigenetic bases for their biological properties, we investigated these cells using a wide range of large-scale methods including whole genome sequencing, RNA sequencing, SNP array, DNA methylation array, and de novo genome assembly. We conducted an integrative analysis of both cell lines to distinguish between potential driver and passenger alterations. Although many genes are mutated in these cell lines, the combination of DNA- and RNA-based variant information strongly implicates a small number of genes including *TP53* and *STK11* as likely drivers. Likewise, we found a diverse set of genes differentially expressed between these cell lines, but only a fraction can be attributed to changes in DNA copy number or methylation. This set included the ABC transporter *ABCC4*, implicated in drug resistance, and the metastasis associated *MET* oncogene. While the rich data content allowed us to reduce the space of hypotheses that could explain most of the observed biological properties, we also caution there is a lack of statistical power and inherent limitations in such single patient case studies.

### 1. Introduction

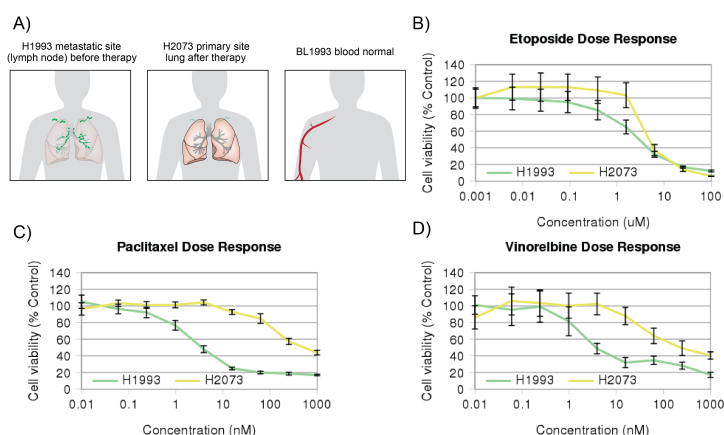
Cancer arises as a result of genomic or epigenomic alterations that change a wide range of cellular processes, leading to uncontrolled tumor cell proliferation and other tumor-specific characteristics (1). Cytotoxic agents and targeted therapies have been developed to treat cancer patients. However, one major challenge during treatment is the potential development of drug resistance (2). Lung cancer, the leading cause of cancer-related death (3), is one of the most heterogeneous of cancer types in terms of underlying molecular characteristics and therapy response. It is biologically and clinically important to understand the underlying genetic lesions influencing cancer cell behaviors such as differential drug response. Recent advances in high-throughput sequencing allow the elucidation of genomewide patient-specific molecular profiles that reveal individual tumor drivers and form the basis for personalized treatments (4). However, most

identified genetic variation is usually difficult to interpret, as the vast majority of alterations are passenger mutations. In addition, not all genomic features can be obtained by a single technology. Integrative approaches have the potential to capture the combination of patient-specific characteristics on various levels for a better understanding and targeting of the molecular basis of specific cancers – a rising field termed “panomics”. It is however not clear if comprehensive and deep analyses of a small number of patients, or single patients, might reveal new insights of the genetic basis of patient phenotype.

In this study, we performed a wide spectrum of genomic analyses to study a lung cancer patient, who underwent chemotherapy but relapsed with tumor regrowth at the primary site. Two cell lines were derived from this patient: one from a lymph node metastasis isolated prior to chemotherapy, and the other from the lung tumor regrowth months after chemotherapy. Although derived from the same individual, these two cell lines have distinct drug response profiles. To understand the underlying genetic basis for their phenotypic differences, we performed whole genome sequencing, transcriptome sequencing, SNP array, DNA methylation array, and de novo whole genome assembly to thoroughly interrogate genetic and epigenetic events. We conducted an integrative analysis of both cell lines and constructed a model that might explain the development of the patient’s cancer and drug resistance after chemotherapy.

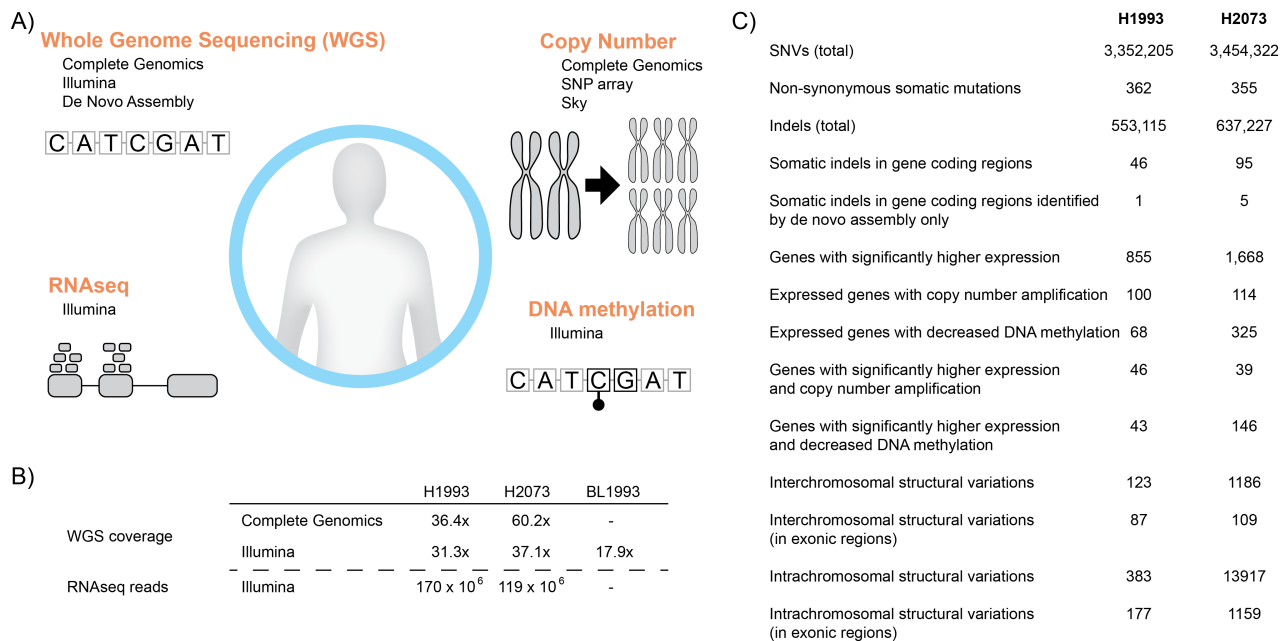
## 2. Sample Description, Drug Response and Screening Overview

Cell line H1993 was derived from the lymph nodes of a 47 year old female Caucasian with history of smoking and diagnosed with non-small cell lung cancer in 1988 (Figure 1A). After treatment with cisplatin and etoposide, H2073 was derived from the resected lung tumor of the same patient. We performed drug response studies as previously described (5). As expected, H2073 shows resistance to etoposide (Figure 1B). Interestingly, the spectrum of drug resistance of H2073 cells encompasses a broader range of therapeutics including paclitaxel and vinorelbine (Figures 1C-D), which target mitotic division.



**Figure 1. Sample description and cytotoxic drug resistance of H2073**

A) Cell line H1993 was derived from a metastatic site in patient’s lymph nodes, while H2073 originated from the primary lung tumor after treatment with cisplatin and etoposide. BL1993 was derived from lymphoblastoid cells of the same patient, thus representing the matched normal blood sample. (B-D) In comparison to H1993, H2073 cells show higher viability upon treatment with etoposide, paclitaxel and vinorelbine. Error bars indicate standard deviation.



**Figure 2. Integrative analysis of H1993 and H2073: a panomics approach**

**A)** We applied an integrative analysis of H1993 and H2073 based on whole genome sequencing, RNA sequencing, DNA methylation quantification and copy number investigation. **B)** The genome of each cell line was sequenced at minimum 30x coverage. **C)** The panomics approach allowed us to analyze the landscape of single nucleotide variants, indels, differential gene expression, copy number changes, and structural variations. The numbers of detected aberrations are shown for these two cell lines.

To elucidate the development of the patient's cancer and to understand the drug resistance after chemotherapy, we applied an integrated analysis of somatic exonic mutations, messenger RNA sequencing, DNA copy number, and promoter DNA methylation (Figure 2A).

Whole genome sequencing (WGS) of both cell lines was conducted on two independent platforms: Complete Genomics (CG) and Illumina, to a minimum depth of 30x (Figure 2B). In addition, we constructed DNA libraries with variable insert sizes for both cell lines, performed Illumina-based paired-end sequencing, and used the resulting reads for de novo genome assembly, in order to identify genomic features missed by reference-based approaches. We also carried out Illumina WGS on DNA isolated from BL1993, a lymphoblastoid cell line from the same patient, representing the matched normal blood sample.

To identify genes differentially expressed between H1993 and H2073, we collected RNA-Seq data in 3 replicates. DNA methylation was measured by Illumina Infinium array, and copy number analysis with the Illumina OMNI 2.5M SNP array, processed by a modified version of the PICNIC algorithm, as previously described (6, 7). Results are summarized in Figure 2C.

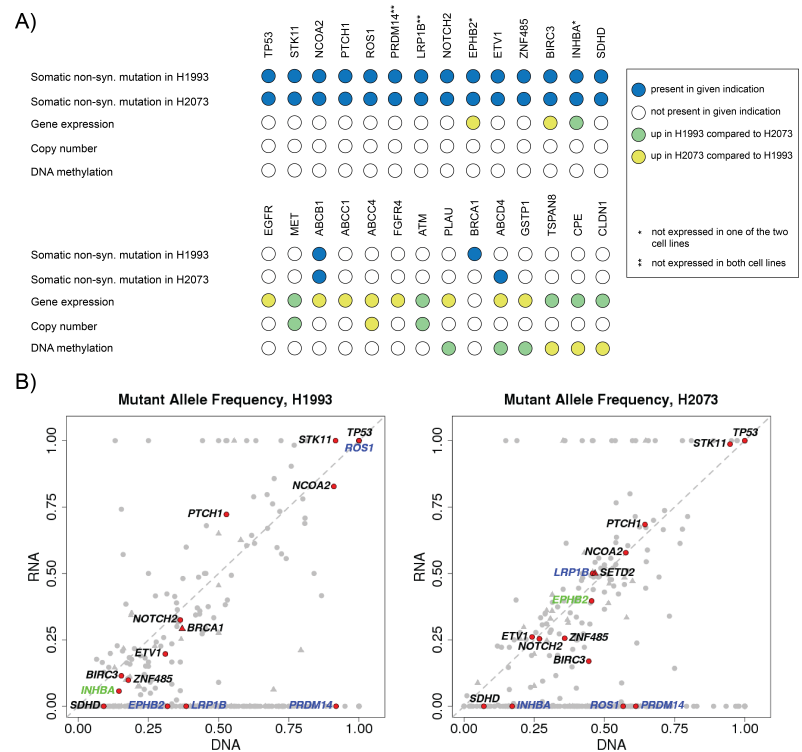
### 3. Mutation Landscapes and the Identification of Expressed Variants

Somatic mutations were identified by comparing the variant calls in H1993 and H2073 with BL1993. We selected non-synonymous mutations with a minimum support of five reads and excluded known germline variants from a variety of sources (see Methods). Any variants listed in COSMIC database of somatic mutations in cancer (8) were retained. This resulted in 313 somatic non-synonymous single-base substitutions in common between H1993 and H2073, of which 290 were missense mutations, 21 resulted in stop gain, and 2 resulted in stop loss. Consistent with the patient's smoking history, we observed an enriched fraction of C:G > A:T transversions, the smoking-related mutation signature, in the tumor-specific variants (data not shown).

Both cell lines harbor non-synonymous mutations in genes known to be altered in lung cancer, including *TP53*, *STK11*, *EPHB2*, *LRP1B*, *INHBA*, *ZNF458*, and *PRDM14* (Figure 3A). Other somatically mutated cancer genes, which are listed in the Cancer

Gene Census (CGC) (9), include *NOTCH2*, *BIRC3*, *PTCH1*, *ETV1*, *ROS1*, *SDHD* and *NCOA2*. To prioritize these putative drivers, we used RNA-Seq to eliminate genes with little or no expression (RPKM<0.5). This expression based filtering reduced the number of common mutations from 313 to 106 (96 missense, 10 stop gain), eliminating a large fraction of candidate genes at the risk of possibly discarding low expressed drivers (Figure 3B).

We further hypothesized that the mutant alleles for driver mutations should be selected for, leading to higher mutant allele frequencies for driver genes. We then assessed mutant allele frequencies in DNA and RNA data and grouped the mutations into three frequency classes (Figure



**Figure 3. Genomic landscapes and pathway alterations of H1993 and H2073**

**A)** Multiple cancer related genes were somatically mutated in both cell lines (upper panel) or differentially expressed between the cell lines (lower panel). **B)** Integrating gene expression and focusing on instances of high mutant allele frequency enabled us to substantially reduce the set of candidate drivers. Known cancer related genes are highlighted. Genes with low expression (<0.5 RPKM) in both cell lines are shown in blue, while genes with low expression in one cell line are shown in green. Triangles indicate cell line-specific mutations, while circles correspond to common mutations.

3B): high ( $>0.9$ , class 1), medium (0.3 to 0.9, class 2), and low/inconsistent (class 3). Mutations at loci with a total DNA or RNA read coverage  $< 10$  were also assigned to class 3. Class 1 comprised only 10 genes, 8 of which had stop gain or missense mutations that were predicted to be deleterious (10) based on Polyphen2 (11) and SIFT (12) calculations. In this reduced set of candidate drivers were tumor protein 53 (*TP53*) and serine/threonine kinase 11 (*STK11*, also known as *LKB1*), the two most significantly mutated tumor suppressors in lung cancer (13). Both mutations were observed in regions with loss of heterozygosity. The homozygous *TP53* missense mutation C242W was also observed in other cancer types including breast (14) and stomach (15) cancer, while the homozygous stop gain mutation on position 199 within the kinase domain of *STK11* has been previously reported in other lung cancer samples (16). Thus, integrating WGS and RNA-Seq data on the two cell lines allowed us to reduce a set of non-synonymous mutations to two likely drivers of oncogenesis in this patient.

While 106 non-synonymous mutations in expressed genes were common to both cell lines, 20 and 22 were specific to H1993 and H2073, respectively. These included Cancer Gene Census genes *SETD2* (class 2) in H2073, and *BRCAl* (class 2) in H1993. Inactivation of *BRCAl* is associated with tumor aggressiveness and invasion (17), consistent with the metastatic state of H1993. None of the cell line specific mutations was assigned to class 1. Overall, the limited difference between H1993 and H2073 mutation profiles indicates that unique point mutations are unlikely to explain the phenotypical variations between them.

Among 138 somatic coding indels detected in either cell line, 7 affected Cancer Gene Census genes. All of these were cell line-specific, with frame-shifting indels observed in genes *SF3B1*, *BMPRIA*, and *GPHN* in H1993, and in genes *JUN*, *MLL3*, *NR4A3* in H2073. We also observed an in-frame insertion in gene *MLL2* in H2073. It is unclear what role, if any, these mutations may play in the observed phenotypic differences between the two cell lines. While histone methyltransferases *MLL2* and *MLL3* have been linked to *TP53*-mediated DNA damage response pathway (18, 19), our cell lines exhibited lack of a functional copy of *TP53*, rendering any additional mutations to this pathway inconsequential.

#### 4. Differentially Expressed Genes and the Relationship with DNA Changes

Our RNA-Seq analysis identified 2,523 differentially expressed genes between H1993 and H2073 (Figure 4A), of which 1,668 (67%) were over-expressed in H2073. Classical markers for epithelial/mesenchymal status, including *CDH1*, *CDH2*, *VIM* and *FNI*, were not consistently differentially expressed between the two cell lines, suggesting that the observed differences between the primary and the metastatic cell line were not due to epithelial-to-mesenchymal transition.

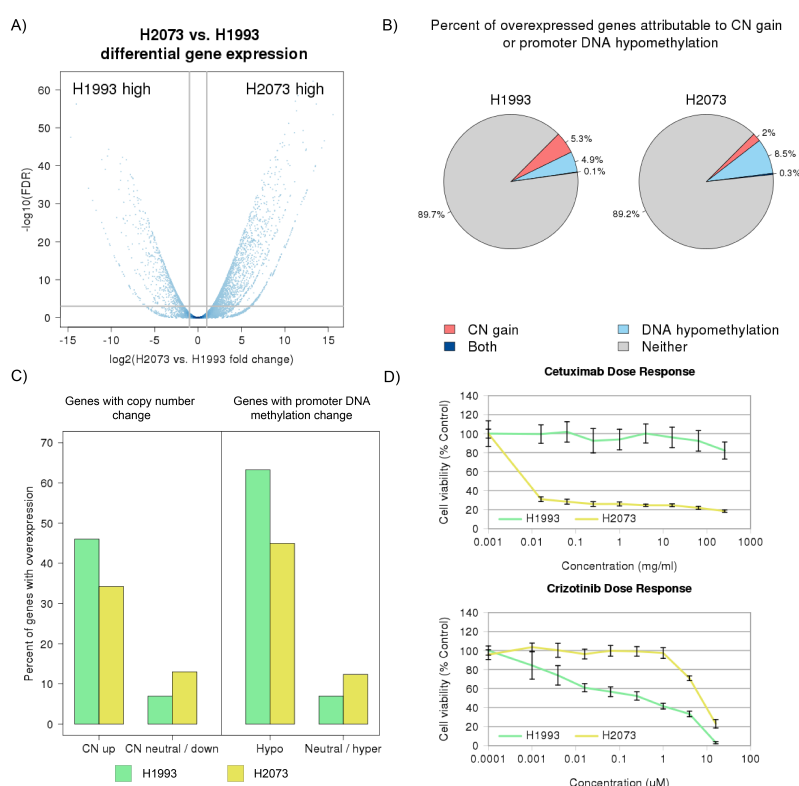
The large number of differentially expressed genes also suggests that most of these expression changes are downstream effects of the causal events. We hypothesized that the primary expression differences should have certain degree of genetic or epigenetic basis. We therefore focused on differentially expressed genes that can be directly attributed to changes in copy number or DNA methylation state. We found that 39 out of 1,668 (2.3%) genes overexpressed in H2073 are in regions amplified in H2073 relative to H1993 (ploidy adjusted CN fold change  $\geq 2$ ). Ploidy adjustment was carried out because H1993 is mostly tetraploid, while H2073 has average ploidy



between 2 and 3, consistent with cytogenetic results (data not shown). Similarly, we observed that 46 out of 885 (5.4%) genes overexpressed in H1993 are in genomic regions amplified in H1993 relative to H2073 (Figure 4B).

Overall, regions amplified in H1993 and H2073 contained 100 and 114 expressed genes, respectively, out of which 46 (46%) and 39 (34%) were overexpressed according to our cutoffs, exhibiting higher rate of overexpression events than non-amplified regions (Figure 4C, Fisher exact test p-value  $<7 \times 10^{-9}$  for both cell lines). In total, we identified seven amplified regions in either cell line longer than 1 Mb, six of which (three in each cell line) accounted for 82 out of 85 differentially expressed genes with underlying CN changes. One of the H2073 amplicons included transporter gene *ABCC4*, previously implicated in drug resistance and showing 3-fold overexpression in H2073. The region on chromosome 7, highly amplified ( $>10$  copies) in H1993, contained oncogene *MET* (Figure 5A), which is known to be involved in tumor cell invasion and metastasis (20). We found *MET* to be 7-fold overexpressed in H1993, consistent with the metastatic character of H1993. The dependence of H1993 on *MET* is confirmed by its low viability in the presence of *MET* inhibitors (Figure 4D). Another highly amplified genomic region was located on chromosome 11 and contained the oncogene *ATM* (4-fold overexpression in H1933), which was also reported to promote metastasis (21).

Comparing the two cell lines further, we found that 427 genes expressed in at least one cell line showed differentially methylated regions (DMRs) within 2kb of their transcription start site (TSS). Out of 1,668 genes overexpressed in H2073, 166 (9.9%) contained DMRs (Figure 4B). In 146 cases (82%), the extent of methylation was higher in H1993, consistent with down-regulation of



**Figure 4. Differential gene expression analysis between H1993 and H2073**

**A)** Volcano plot illustrating fold changes and false discovery rates for all human genes as calculated by differential gene expression analysis. **B)** Percentage of overexpressed genes with significant copy number gain or DNA hypomethylation. **C)** The sets of expressed genes with copy number amplification or promoter DNA hypomethylation were enriched for overexpressed genes. **D)** Treating both cell lines with an EGFR inhibitor Cetuximab reveals lower viability of H2073 in comparison to H1993. Treating the two cell lines with a MET inhibitor Crizotinib reveals lower viability in H1993. Error bars indicate standard deviation.

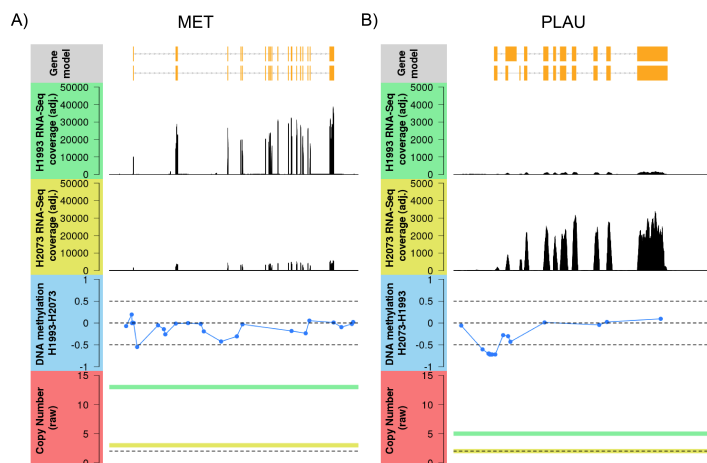


expression by hypermethylation. In comparison, 61 out of 885 (6.9%) genes overexpressed in H1993 contained DMRs within 2kb of TSS, with 43 (70%) exhibiting higher methylation in H2073. In total, hypomethylated DMRs were associated with 68 and 325 genes in H1993 and H2073, respectively, out of which 43 (63.2%) and 146 (44.9%) showed overexpression, exhibiting higher rate of overexpression events than hypermethylated or non-differentially methylated regions (Figure 4C, Fisher exact test p-values  $< 2 \times 10^{-32}$  for both cell lines). Several of the genes with overexpression and promoter DNA hypomethylation in H2073 have been implicated in apoptosis evasion and drug resistance, including *PLAU* (Figure 5B), *SNCG*, *BNIP3*, *GSTP1*, *ETS1*, and *MSLN*. Interestingly, we found the binding partners *PLAU* and *PLAUR* to be overexpressed in H2073, suggesting co-regulation of their expression. Binding of *PLAU* to *PLAUR* can activate the *ERK* pathway and contribute to cancer development (22).

Genes overexpressed and hypomethylated in H1993 included the metastasis effectors *RAB25*, *TSPAN8*, and *CPE*, as well as *CLDN1*, whose up-regulation has been associated with cisplatin sensitivity (23), consistent with cisplatin resistance in H2073. Overall, 10.8% of genes overexpressed in H2073 and 10.3% of genes overexpressed in H1993 are associated with either differential DNA methylation or copy number rearrangements. Thus, the integration of these two additional data types allowed us to substantially reduce the number of candidate drivers, while possibly omitting driver genes activated via alternative mechanisms.

Guided by our discovery of the amplification of transporter gene *ABCC4* in the drug-resistant cell line H2073, we tested for differential expression of other transporter genes. While one transporter gene, *ABCB10*, showed overexpression in H1993, several others were overexpressed in H2073 and are known to play a role in drug resistance. We found that the multi-drug resistance transporter *MDR1/ABCB1* was expressed in H2073 but almost absent from H1993.

Both *ABCC4* and *ABCC1*, also implicated in drug resistance, were also at least 3-fold overexpressed in H2073 (24, 25). Furthermore we found 9-fold higher expression of *FGFR4* in H2073. A recent study reported that inhibition of *FGFR* reverses *ABCB1*-mediated drug resistance in cancer (26). Overall, these results suggest an efflux-based drug resistance mechanism developed

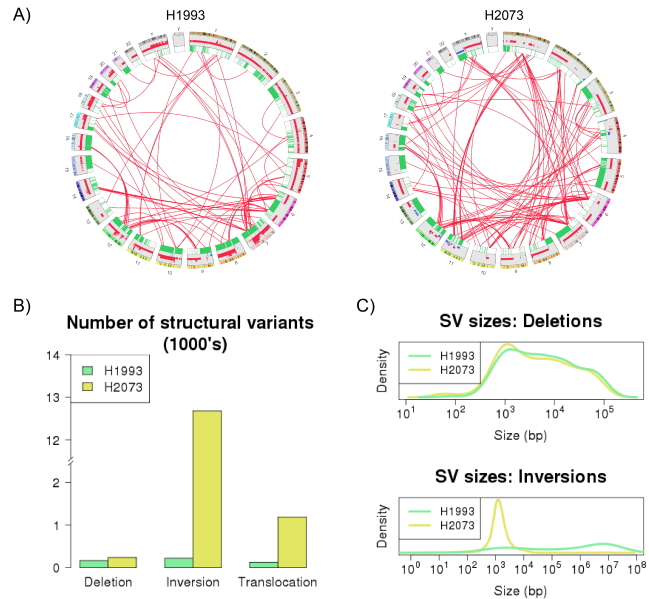


**Figure 5. Overexpression of MET in H1993 is associated with copy number gain (A), while overexpression of PLAU in H2073 is associated with decreased promoter methylation (B).**

The panels show gene structure (top, individual transcript isoforms), expression normalized by sequencing depth (H1993: second from the top, H2073: third from the top), difference in DNA methylation (second from bottom, dashed lines correspond to differences of 0.5, 0, and -0.5), and raw copy number (bottom, green line: H1993, yellow line: H2073, dashed black line: CN=2 (baseline)).

by H2073, which involves *ABCC4*, *ABCB1*, and possibly other transporter proteins that were not over-expressed in H2073 based on our cutoffs.

Integrating information on changes in DNA copy number and methylation allowed us to reduce a large set of differentially expressed genes 10-fold to candidate drivers with clear underlying mechanism of differential expression. Close examination of these candidate drivers revealed a number of genes overexpressed in H1993 and known to be involved in metastasis. This allowed us to construct a drug resistance model for H2073. However, this reductionist approach has its limitations, as not all meaningful differential expression can be attributed to a change in either DNA copy number or methylation. As an example, the expression of the well-known cancer gene *EGFR* is 8-fold higher in H2073 than in H1993, and the dependence of H2073 on *EGFR* for survival and proliferation is strongly suggested by its higher sensitivity to *EGFR* inhibitors (Figure 4D). However, the observed overexpression of *EGFR* was not associated with either a copy number change or differential promoter DNA methylation in this study. It is likely that other types of genetic or epigenetic alterations, such as histone mark changes, are responsible for the observed *EGFR* expression change but are not captured by our existing assays.



**Figure 6. Structural variations in H1993 and H2073**

**A)** Illustration of genomic alterations in H1993 and H2073 using Circos plots. Candidate interchromosomal structural variations identified by the Complete Genomics Platform are shown as red lines. Copy number changes detected by Illumina SNP arrays are illustrated as bar plots. Loss of heterozygosity regions are shown in green. **B)** Structural variations, in particular smaller inversions **(C)**, were more frequent in the cell line derived after chemotherapy (H2073)

## 5. Structural Variation Analysis

Based on WGS by the Complete Genomics platform, we observed 164 large deletions (50bp-100kb), 219 inversions, and 123 translocations in H1993, supported by at least 5 reads (Figure 6A-B). H2073 showed substantially more structural variants, with 237 large deletions, 13680 inversions, and 1186 translocations. This significant increase in the number of structural variants, in particular short inversions (Figure 6C), might be due to the stress imposed on the cell by the chemotherapy (27). This is consistent with the fact that H1993 was derived from tumor cells prior to chemo-treatment, while H2073 was derived afterward and therefore is chemo-resistant.

## 6. De Novo Genome Assembly Reveals Additional Variant Information

To discover genomic alterations that might be missed by standard WGS analysis, we performed de novo assembly of H1993 and H2073 genomes, based on paired-end Illumina sequences. The insert

size ranged from 200bp to 40kb, in order to aid longer range DNA assembly. The resulting assembled sequences span 2.96 (H1993) and 2.89 (H2073) Gb, including 2.29 and 2.48 Gb of fully resolved (non-gapped) sequence. The N50 values were 1.9 Mb and 1.26 Mb, respectively, reflecting a large portion of the sequence in scaffolds of substantial (>1Mb) size.

We aligned the assembled sequences to the reference genome to identify insertions or deletions, which may have been missed by resequencing-based approaches. We identified 2 insertions and 3 deletions that were exclusively detected by the assembly approach and that affected exons (Table 1). These indels ranged in size from 51 to 123 bp, indicating the utility of the assembly approach in detecting medium size indels, that are not short enough to be detected by most resequencing-based indel callers, but are not long enough to be detected by the copy number or structural variation analyses. We note that the observed frame-shifting deletion in TSPAN8 in H2073 may have contributed to its lower expression in that cell line, alongside the hypermethylation component, described above.

Table 1. Assembly-specific exonic indels.

Indel Type	Coordinate	Length (bp)	Affected gene	Cell line
Deletion	Chr1:7,889,973-7,890,026	54	PER3	Both
Deletion	Chr2:27,324,254-27,324,304	51	CGREF1	H2073
Insertion	Chr12:71,523,133-71,523,134	109	TSPAN8	H2073
Deletion	Chr14:104,645,583-104,645,705	123	KIF26A	H2073
Insertion	Chr20:62,196,017-62,196,018	57	PRIC285	H2073

## 7. Conclusions

The expansion of high-throughput assays for analyzing cellular states has provided new opportunities for integrative analyses. Here we used several genome-scale analyses of 2 cancer cell lines to ask whether we could better explain their observed biological similarities and differences. Perhaps the most significant challenge in interpreting genomic data is to pinpoint the most relevant genomic changes from a large collection of data points, and the panomics approach by definition epitomizes this problem. While it might be practically impossible to achieve statistical significance for such panomics approaches, we believe that prior knowledge and logical combination of different data could dramatically reduce the search space and propose biologically meaningful models.

In this study, while variant analysis revealed more than 300 non-synonymous mutations, combining this analysis with expression data reduced the number of candidate drivers 3-fold. Integrating allele frequencies on both DNA and RNA levels further reduced the focal set to 8 homozygously mutated genes, including likely drivers *TP53* and *STK11*. Similarly, while expression analysis alone revealed thousands of differentially expressed genes between the two cell lines, only a small fraction of such genes were associated with the underlying genetic and epigenetic changes. Among these small number of genes, *MET* was present in a highly amplified region and showed 7-fold overexpression in H1993, and *ABCC4* was amplified and overexpressed in the drug resistant cell line H2073. Although we could not exclude other genomic changes that

might also explain the phenotypic differences between these two cell lines, our integrated analyses readily produced a working model that is consistent with our knowledge of the samples.

It is worth noting that although H1993 and H2073 have been independently cultured *ex vivo* for decades, they show remarkable similarity and display largely overlapping point mutations. This shows that any new mutations acquired during the cell culturing steps are at the minimum if they exist. This finding boosts the validity of these cell lines as stable model systems for cancer studies. From the technology point of view, our *de novo* assembly of both cell lines revealed a number of additional insertions and deletions, missed by the reference-based assembly. Only 5 of these altered protein coding regions, indicating that reference-based assembly captures most of the actionable variants.

It should also be noted that this study is exploratory by nature. With such small sample size, the statistical power is nonexistent, so it is currently impossible to draw any causal relationships with any confidence. This approach should be viewed as a hypothesis generating method. Alternatively, this approach can be viewed as a “hypothesis-supporting method”. Our current knowledge of lung cancer and drug resistance has led us to propose genes like *EGFR*, *MET*, and *ABCC4* as functionally relevant culprits in these cell lines, but an improved knowledge in the field might implicate a different set of genes. It is therefore necessary to view the panomics data with a grain of salt, as the interpretation of these data can be influenced by the current biological knowledge. Nevertheless, the maturation of this field will enhance our ability to better analyze panomics data, as no single assay can provide a full picture of the cell state or to point in the direction of possible therapeutic actions.

## 8. Materials and Methods

### 8.1. Whole Genome Sequencing and Variant Calling

Whole genome sequencing (WGS) of H1993 and H2073 was performed by Complete Genomics, as described (7). Independently, WGS of H1993, H2073, and BL1993 was performed by Illumina sequencing (100bp paired-end reads), using libraries with insert sizes of 200, 500, 2000, 5000, 10000, 20000, and 40000 bp. Reads were aligned to reference human genome (hg19) using BWA (28). Single nucleotide variants (SNV) and indels were called by the Complete Genomics WGS processing pipeline. Several variant callers were applied to Illumina WGS data. We used SOAPsnp (29) to identify germline SNVs in all 3 cell lines, and VarScan (30) to identify cell line-specific SNVs in every possible 2-cell line comparison. Only variants supported by 5 or more reads and separated by 10 or more base pairs from the nearest variant were retained. Somatic mutations were identified by requiring that no variant-supporting reads be detected in BL1993 WGS. Unless the variant was listed in COSMIC database of cancer mutations, we further required that it was covered by 10 or more reads in BL1993, and not present in dbSNP (v.132) (31) or among variants from 1000 Genomes Project (32), 6515 normal exomes published by NHLBI (33), or 69 normal genomes sequenced by Complete Genomics and made available to the public (34). We used Dindel (35) to identify germline indels and GATK (36) to identify cell line-specific indels. Indels were declared somatic if no compatible indel was detected in BL1993 by Dindel, and if the indel was not part of a set of known normal indels obtained from 1000 Genomes Project

and 69 publicly available Complete Genomics sequenced normal genomes. Structural variants were obtained from the Complete Genomics pipeline.

## **8.2. Copy Number Analysis**

H1993 and H2073 cell lines were assayed with Illumina OMNI 2.5M SNP array and processed with a modified version of PICNIC (7). When calculating copy number fold change between the two cell lines, adjustment was made for average cell line ploidy. This was calculated as the average copy number per base pair, and was 3.9 for H1993 and 2.4 for H2073.

## **8.3. Messenger RNA Sequencing**

Three temporally separate, standard RNA-seq library preparations and subsequent sequencing data were collected for each of the two cell lines. One of the libraries for each cell line was sequenced on an Illumina GAI, while the remaining libraries were sequenced on an Illumina HiSeq machine. The resulting RNA-seq data was filtered for read quality, ribosomal RNA contamination, and then aligned to the human reference genome (NCBI Build 37) using the GSNAP alignment tool (37). Alignments were permitted a maximum of 3 mismatches per 75 base pair sequence and used the following GSNAP parameters: “-M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1 --pairmax-rna=200000”. These steps, and the downstream processing of the resulting alignments to obtain read counts and RPKMs per gene (over coding exons of RefSeq gene models) per replicate are implemented in the Bioconductor package, HTSeqGenie (v 3.10.0) (38).

We compared the gene expression profiles of the two cell lines using the gene count data described above, and the Bioconductor package edgeR (39). Each of the three temporally separate RNA-seq libraries per cell line was used as biological replicates for dispersion estimates within edgeR. Genewise exact tests for differential gene expression were performed, and resulting summary statistics reported. We used the cutoffs of FDR<0.001, fold change > 2, and RPKM  $\geq 0.5$  (in cell line with overexpression) to declare differential expression.

## **8.4. DNA Methylation Analysis**

DNA methylation was measured by Illumina Infinium Human Methylation 450K BeadChips and preprocessed using the Bioconductor lumi package (40). Methylation status was measured in beta-values ranging from 0 (unmethylated) to 1 (methylated). A probe was considered to show significant methylation change, if the difference between H1993 and H2073 beta-values was larger than 0.5. Nearby (less than 2kb) differentially methylated probes were merged into differentially methylated regions (DMR). Final differential methylation calls were based on DMRs near gene transcription starting site (TSS) (2kb upstream of TSS or overlapping with the first exon of the gene) with a minimum of two supporting probes.

## **8.5. Data Access**

The results of Complete Genomics WGS and copy number SNP array assays have been previously published (7). The remaining data will be made available to the public and the repository location and accession information can be obtained from the authors.

## Acknowledgments

We thank Jens Reeder and Gregoire Pau for development of the transcriptome sequencing analysis pipeline, Allison Bruce for help with the design of figure graphics, and Gerard Manning for critical review of this manuscript.

## References

1. Vogelstein B, Kinzler KW. 2004. *Nat Med* 10: 789-99
2. Longley DB, Johnston PG. 2005. *J Pathol* 205: 275-92
3. Siegel R, Naishadham D, Jemal A. 2012. *CA Cancer J Clin* 62: 10-29
4. Roychowdhury S, Iyer MK, Robinson DR, et al. 2011. *Sci Transl Med* 3: 111ra21
5. Gandhi J, Zhang J, Xie Y, Soh J, Shigematsu H, et al. 2009. *PLoS One* 4: e4576
6. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, et al. 2012. *Genome Res* 22: 593-601
7. Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, et al. 2012. *Genome Res* 22: 2315-27
8. Forbes SA, Bhamra G, Bamford S, et al. 2008. *Curr Protoc Hum Genet* Chapter 10: Unit 10 1
9. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. 2004. *Nat Rev Cancer* 4: 177-83
10. Gnadt F, Baucom A, Mukhyala K, Manning G, Zhang Z. 2013. *BMC Genomics* 14 Suppl 3: S7
11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, et al. 2010. *Nat Methods* 7: 248-9
12. Kumar P, Henikoff S, Ng PC. 2009. *Nat Protoc* 4: 1073-81
13. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, et al. 2008. *Nature* 455: 1069-75
14. Chevillard S, Lebeau J, Pouillart P, de Toma C, et al. 1997. *Clin Cancer Res* 3: 2471-8
15. Kubicka S, Claas C, Staab S, Kuhnel F, Zender L, et al. 2002. *Dig Dis Sci* 47: 114-21
16. Imielinski M, Berger AH, Hammerman PS, Hernandez B, et al. 2012. *Cell* 150: 1107-20
17. Albiges L, Andre F, Balleyguier C, Gomez-Abuin G, et al. 2005. *Ann Oncol* 16: 1846-7
18. Guo C, Chang CC, Wortham M, Chen LH, et al. 2012. *Proc Natl Acad Sci U S A* 109: 17603-8
19. Lee J, Kim DH, Lee S, Yang QH, Lee DK, et al. 2009. *Proc Natl Acad Sci U S A* 106: 8513-8
20. Jeffers M, Rong S, Vande Woude GF. 1996. *J Mol Med (Berl)* 74: 505-13
21. Sun M, Guo X, Qian X, Wang H, Yang C, et al. 2012. *J Mol Cell Biol* 4: 304-15
22. Nguyen DH, Hussaini IM, Gonias SL. 1998. *J Biol Chem* 273: 8502-7
23. Fortier AM, Asselin E, Cadrin M. 2013. *J Biol Chem* 288: 11555-71
24. Nath S, Daneshvar K, Roy LD, Grover P, Kidiyoor A, et al. 2013. *Oncogenesis* 2: e51
25. Oprea-Lager DE, Bijnsdorp IV, RJ VANM, AJ VDE, et al. 2013. *Anticancer Res* 33: 387-91
26. Patel A, Tiwari AK, Chufan EE, et al. 2013. *Cancer Chemother Pharmacol* 72: 189-99
27. Portugal J, Mansilla S, Bataller M. 2010. *Curr Pharm Des* 16: 69-78
28. Li H, Durbin R. 2009. *Bioinformatics* 25: 1754-60
29. Li R, Li Y, Fang X, Yang H, Wang J, et al. 2009. *Genome Res* 19: 1124-32
30. Koboldt DC, Chen K, Wylie T, Larson DE, et al. 2009. *Bioinformatics* 25: 2283-5
31. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. *Nucleic Acids Res* 29: 308-11
32. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. 2012. *Nature* 491: 56-65
33. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. 2013. *Nature* 493: 216-20
34. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. 2010. *Science* 327: 78-81
35. Albers CA, Lunter G, MacArthur DG, McVean G, et al. 2011. *Genome Res* 21: 961-73
36. McKenna A, Hanna M, Banks E, Sivachenko A, et al. 2010. *Genome Res* 20: 1297-303
37. Wu TD, Nacu S. 2010. *Bioinformatics* 26: 873-81
38. Pau G, Reeder J. 2012. *R package* R package version 3.10.0
39. Robinson MD, McCarthy DJ, Smyth GK. 2010. *Bioinformatics* 26: 139-40
40. Du P, Kibbe WA, Lin SM. 2008. *Bioinformatics* 24: 1547-8

# AN INTEGRATED APPROACH TO BLOOD-BASED CANCER DIAGNOSIS AND BIOMARKER DISCOVERY

Martin Renqiang Min<sup>ad\*</sup>, Salim Chowdhury<sup>bd</sup>, Yanjun Qi<sup>a</sup>, Alex Stewart<sup>c</sup>, and Rachel Ostroff<sup>c</sup>

<sup>a</sup>NEC Labs America, Princeton, NJ 08540, USA, <sup>b</sup>Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA, <sup>c</sup>SomaLogic, Inc., Boulder, CO 80301, USA

E-mail: renqiang@nec-labs.com, salim@cmu.edu, yanjun@nec-labs.com, {astewart, rostroff}@somalogic.com

<sup>d</sup>These authors contributed equally to this work

Disrupted or abnormal biological processes responsible for cancers often quantitatively manifest as disrupted additive and multiplicative interactions of gene/protein expressions correlating with cancer progression. However, the examination of all possible combinatorial interactions between gene features in most case-control studies with limited training data is computationally infeasible. In this paper, we propose a practically feasible data integration approach, QUIRE (QUadratic Interactions among infoRmative fEatures), to identify discriminative complex interactions among informative gene features for cancer diagnosis and biomarker discovery directly based on patient blood samples. QUIRE works in two stages, where it first identifies functionally relevant gene groups for the disease with the help of gene functional annotations and available physical protein interactions, then it explores the combinatorial relationships among the genes from the selected informative groups. Based on our private experimentally generated data from patient blood samples using a novel SOMAmer (Slow Off-rate Modified Aptamer) technology, we apply QUIRE to cancer diagnosis and biomarker discovery for Renal Cell Carcinoma (RCC) and Ovarian Cancer (OVC). To further demonstrate the general applicability of our approach, we also apply QUIRE to a publicly available Colorectal Cancer (CRC) dataset that can be used to prioritize our SOMAmer design. Our experimental results show that QUIRE identifies gene-gene interactions that can better identify the different cancer stages of samples, as compared to other state-of-the-art feature selection methods. A literature survey shows that many of the interactions identified by QUIRE play important roles in the development of cancer.

**Keywords:** Blood-based Cancer Diagnosis; Biomarker Discovery; Feature Interactions; Sparse Learning; Aptamer; SOMAmer Prioritization.

## 1. Introduction

In this paper, we focus on the task of biomarker discovery and cancer diagnosis directly based on patient blood samples in the setting of limited training data. Although cancer diagnosis based on microarray datasets has been extensively studied, blood-based cancer status prediction is still a challenging problem, because complex diseases like cancers are the results of multiple genetic and epigenetic factors and their manifestation in blood samples is even more complicated than in tumor samples. It is very difficult to identify these complicated factors solely based on limited information provided by training data. Previous studies on single gene markers can provide valuable information about disease status prediction, but they offer limited insight into the complex interplay among the molecular factors responsible for progression of complicated diseases such as cancers. So, recently, research in complex diseases shifts towards the identification of multiple genes that interact directly or indirectly in contributing their association to the target disease. Several complex interactive partners from previous

---

\*To whom correspondence should be addressed.

studies have been shown to give important insight into the mechanism of breast cancer<sup>1</sup> and colorectal cancer,<sup>2</sup> but none of these approaches addressed the problem of disease diagnosis based on blood samples or considered the multiplicative effect of gene/protein expressions.

The identification of groups of genes that show differential behavior in the manifestation of complex diseases is computationally infeasible due to the combinatorial nature of the problem. Several recent methods propose to reduce the search space using orthogonal prior knowledge about connections amongst the genes, such as interactions collected from protein-protein interaction (PPI) network<sup>3</sup> or grouping information from functional annotations of proteins. One notable computational method named Group Lasso<sup>4</sup> incorporates such prior interaction or grouping among the genes to detect gene groups that contribute to human disease, by enforcing sparsity at the group level in a supervised regression framework. Group Lasso is extended by Jacob *et al.*<sup>5</sup> to a more general setting that incorporates groups whose overlaps are nonempty. Such overlaps in groups is biologically significant, because many genes participate in multiple pathways and manifest themselves in several biological processes. Although (Overlapping) Group Lasso is very useful in identifying biologically relevant groups of genes and proteins, it cannot capture complex combinatorial relationships among the features within and across the groups, and it often outputs too many features as biomarkers. Also, current PPI network data is inherently noisy due to experimental constraints.<sup>6</sup> Algorithmic approaches based solely on these noisy prior information can result in many false positive interactions which are absent in the real genome space.

Our goal in this paper is to identify the complex combinations of single genes and multiplicative pairwise interactions among genes that help us (1) better perform cancer diagnosis based on blood samples, and (2) gain novel insights into the mechanistic basis of the diseases. Since the total number of possible pairwise human gene interactions is huge, it is computationally infeasible to examine all possible combinatorial combinations of them when trying to understand their relevance to the diseases under consideration. We propose a two-stage approach in a sparse learning framework, named as QUIRE, i.e. to detect QUadratic Interactions among infoRmative fEatures which show differential behavior for diagnosing a target disease using protein or mRNA expressions. Based on our own experimentally generated data from patient blood samples using a novel SOMAmer technology,<sup>7</sup> we apply QUIRE to blood-based cancer diagnosis for RCC and OVC, and we also apply QUIRE to a publicly available CRC dataset that can be used to prioritize our SOMAmer design. QUIRE can discover complementary sets of markers and pairwise interactions that can better classify samples from different stages of cancer and predict post-cancerous events, like cancer recurrence and death from cancer with higher accuracy than other state-of-the-art feature selection methods. To the best of our knowledge, QUIRE is the first proposed method to identify combinatorial patterns among the pairs of informative genes for studying complex diseases like cancers. Subsequent functional analysis of the interactions identified by QUIRE reveals that it can indeed identify genes relevant to the progression of diseases under study.



## 2. Problem and Method

The identification of single gene markers in a genome-wide case-control study is an ill-posed problem, because the number of genes in human cells is much larger than that of available samples. For such problems, Lasso, proposed by Tibshirani *et al.*<sup>8</sup> is very popular for selecting a small number of features relevant to the problem under study. When a set of features are highly correlated to each other, Lasso selects one from that set, ignoring others. So there is a possibility that Lasso leaves out biologically relevant genes from its set of selected informative features.

Suppose we have a data set  $S$  containing  $n$  samples and  $p$  gene features  $(\mathbf{x}^i, y^i)$  with response variable  $y \in R$  and feature vector  $\mathbf{x} \in R^p$ , where  $i \in \{1, \dots, n\}$ , and we assume that the feature values and the  $y$ s are centered in  $S$ . The Lasso approach optimizes the following objective function,

$$\begin{aligned}\ell(\mathbf{w}) &= \sum_{i=1}^n (y_i - \sum_{j=1}^p w_j x_j^i)^2, \\ \ell_{lasso}(\mathbf{w}) &= \ell(\mathbf{w}) + \lambda \sum_{j=1}^p |w_j|,\end{aligned}\tag{1}$$

where  $\ell(\mathbf{w})$  is the loss function of linear regression, and  $\mathbf{w}$  is the weight parameter. The  $\ell_1$  norm penalty in lasso induces sparsity in the weight space for selecting features. It is obvious that the sum of the least squared errors and the  $\ell_1$  norm are convex functions with respect to the weights  $\mathbf{w}$ . Lasso has a global optimum, which can be identified by any convex optimization technique.

In spite of the computational efficiency and the popularity of Lasso for feature selection, its formulation prevents it from capturing any prior information on possible group structures among the features. Group Lasso<sup>4</sup> proposed using  $\ell_{2,1}$  penalty to select groups of input features which are partitioned into non-overlapping groups. The group penalty is the sum of the  $\ell_2$  norm on the features belonging to the same group. Overlapping Group Lasso<sup>5</sup> extends Group Lasso to handle groups of features with overlapping group members by duplicating input features belonging to multiple groups in the design matrix. Because a lot of real applications involve overlapping feature groupings, Overlapping Group Lasso is a more natural choice than Group Lasso for biomarker discovery. Suppose that we partition  $p$  features in data set  $S$  into  $q$  overlapping groups  $G = \{g_1, g_2, \dots, g_q\}$ , the following objective function is minimized,<sup>5</sup>

$$\ell_{oglasso} = \ell(\mathbf{w}) + \lambda \sum_{g \in G} \|\mathbf{w}_g\|,\tag{2}$$

where  $\lambda$  is the regularization parameter,  $\mathbf{w}_g$  denotes the vector of weights associated with features in group  $g$ , and  $\|\cdot\|$  is the Euclidean norm. The above optimization problem is separable, so we can use block coordinate descent to optimize the weights associated with each group  $g$  separately. Although considering grouping structure among input features is very important for feature selection, Overlapping Group Lasso only encourages sparsity at the feature group level and there is no sparsity penalty within feature groups. Therefore, Overlapping Group Lasso often outputs a much larger number of selected features than Lasso. Furthermore, Lasso

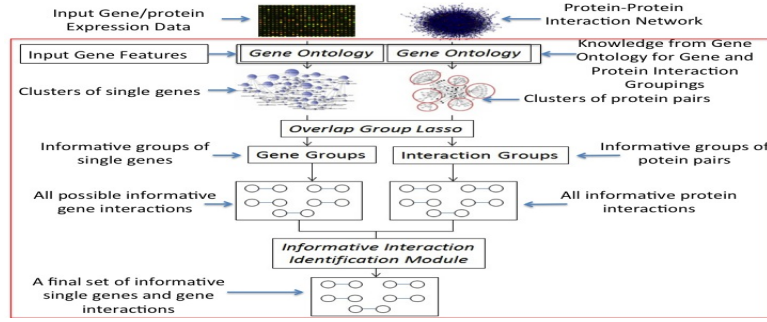


Fig. 1. Working model of QUIRE. QUIRE takes as input, gene or protein expression levels of a set of samples, disease status of those samples and physical interactions amongst the gene products. Then it uses gene ontology based functional annotation to group the genes and cluster the interaction network. Overlapping group lasso is run next on the expression and interaction space to identify informative set of genes and interactions. QUIRE then enumerates all pairwise binary interactions amongst the selected gene features. Finally the proposed novel objective function is applied on the selected single gene features, the informative protein protein interactions and the quadratic interactions amongst these genes to identify the final set of interactions and gene markers. and Overlapping Group Lasso only consider single gene features for prediction, which is limited for disease status prediction and biomarker discovery as shown by our experimental results.

For cancer diagnosis and biomarker discovery from blood samples or tissue samples, we consider all possible combinations of single gene features and quadratic gene interaction features. Ideally, we want to optimize the following optimization problem to identify discriminative features given the dataset  $S$ ,

$$\begin{aligned} \ell(\mathbf{w}, \mathbf{U}) = & \sum_{i=1}^n (y^i - \sum_{j=1}^p w_j x_j^i - \sum_{j=1}^{p-1} \sum_{k=j+1}^p U_{jk} x_j^i x_k^i)^2 \\ & + \lambda_1 \sum_{j=1}^p |w_j| + \lambda_2 \sum_{j=1}^{p-1} \sum_{k=j+1}^p |U_{jk}|, \end{aligned} \quad (3)$$

where  $\mathbf{U}$  is the weights associated with pairwise feature interactions. However, the above model has  $O(p^2)$  features and is not applicable to genome-wide biomarker discovery studies because  $p^2$  is too large. Provided that the training data is often very limited, it is almost impossible to identify the discriminative single or quadratic interaction features by solving the above optimization problem. We propose QUIRE (QUadratic Interactions among infoRmative fEatures) to address these challenges, which is based on Overlapping Group Lasso and Lasso. And it takes advantage of both of these feature selection methods.

The underlying idea of QUIRE is to incorporate all possible complementary biological knowledge (see Figure 1) into the above intractable optimization problem to reduce search space. By restricting discriminative gene interactions to happen only between genes in some informative gene groups, we can use existing functional annotations of input genes to identify these groups thereby to throw away a lot of interaction terms during the optimization. In addition, available physical interactions between the protein products of input genes can also be used to cut the search space, although discriminative gene feature interactions for prediction do not always necessarily correspond to physical interactions. The general working model of QUIRE is shown in Figure 1. In details, QUIRE takes the expression profile of  $n$  samples over  $p$  genes (proteins), the physical interaction network among the genes products (i.e. protein-

protein interaction network) and the disease status of these samples as input, and it outputs a (small) set of discriminative genes and gene interactions with corresponding learned weights for predicting the disease status of any incoming test sample. The step-by-step working model of QUIRE is given below:

(1) *Functional group generation:*

- (a) QUIRE groups the  $p$  input gene features into  $q$  overlapping functional categories according to the existing Gene Ontology (GO) based functional annotations such as Cellular Colocalization (CC).
- (b) QUIRE clusters the given interaction network (i.e. PPI) into subsets of overlapping gene products based on CC.<sup>b</sup>

(2) *Informative genes and functional interactions selection:*

- (a) Given the GO functional grouping of input gene features, Overlapping Group Lasso is run to select  $m$  top discriminative genes for disease status prediction according to the absolute values of the learned weights of gene features.<sup>c</sup>
- (b) Overlapping group lasso is run on the clustered interaction network to select informative groups of protein-protein interactions. In this case, each cluster is considered as a group and the products of pairwise gene/protein feature values among the interacting proteins in a group are used as interaction feature values.

(3) *Selection of most informative interactions and genes:* QUIRE first enumerates all possible quadratic feature interactions among the informative genes selected at step 2(a). Then it takes these quadratic interactions, single informative gene features and the informative functional interactions identified at step 2(b) as input and it outputs the final selected gene interactions and single genes as biomarkers.

In order to identify the discriminative combinations of single gene features and quadratic interactions among pairwise informative genes, we define our proposed objective function for Lasso as follows,

$$\begin{aligned} \ell(\mathbf{w}, \mathbf{U}, \mathbf{R}) = & \sum_{i=1}^n (y^i - \sum_{j=1}^m w_j x_j^i - \sum_{j=1}^{m-1} \sum_{k=j+1}^m U_{jk} x_j^i x_k^i - \sum_{l=1}^r R_l I_l)^2 \\ & + \lambda_1 \sum_{j=1}^m |w_j| + \lambda_2 \sum_{j=1}^{m-1} \sum_{k=j+1}^m |U_{jk}| + \lambda_3 \sum_{l=1}^r |R_l|, \end{aligned} \quad (4)$$

where  $j$  and  $k$  index the  $m$  seed informative genes and  $l$  indexes the  $r$  informative protein-protein interactions selected by the Overlapping Group Lasso in the previous step,  $\mathbf{U}$  and  $\mathbf{R}$  are weights associated with feature interactions, and  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$  are regularization parameters<sup>d</sup>. The objective function contains  $\ell_1$  penalties at single gene level, pairwise gene interaction level, and protein interaction level. The intuition behind this formulation is that it captures

<sup>b</sup>We chose CC as final functional grouping of gene/protein features because it produces groups with reasonable size (see experiment section for details) and it is the most relevant annotation to blood-based diagnosis.

<sup>c</sup> $m$  is selected by 5-fold cross validation.

<sup>d</sup>In our experiments, we make  $\lambda_1 = \lambda_2 = \lambda_3$  and set it by 5-fold cross validation.

the interactions that are complementary to the individual informative genes. Because it is computationally infeasible to consider every pair of interaction in a genome-wide case-control study, QUIRE reduces the search space by using the features that are selected by Overlapping Group Lasso as the informative ones, and then it relies on Lasso with  $\ell_1$  penalties to identify the discriminative combination of informative individual gene features and gene interaction features, which provides an approximation to the problem of searching an exponential number ( $O(2^{p+p^2})$ ) of all possible combinations of single features and pairwise interaction features. Our current implementation of QUIRE is based on the standard Lasso package from glmnet<sup>9</sup> and the Overlapping Group Lasso programs from Jacob *et al.*, 2009.<sup>5</sup>

### 3. Experimental Results and Discussion

In this section, we present experimental results of QUIRE on three different cancer datasets: blood-based cancer diagnosis and biomarker discovery for (1) Renal Cell Carcinoma (RCC) and (2) Ovarian Cancer (OVC) based on our private datasets, and cancer recurrence and death prediction for (3) Colorectal Cancer (CRC) based on a public microarray dataset, in which the identified markers can be used to prioritize our SOMAmer design. We compare the performance of QUIRE to the state-of-the-art feature selection techniques, Lasso, Overlapping Group Lasso and SVM. We then perform a literature survey and enrichment analysis of the informative interactions selected by QUIRE and show that they are relevant to the progression of the disease.

#### 3.1. Our Blood-based Datasets Generated by the SOMAmer Technology

To predict cancer progression status directly from blood samples, we generated our own datasets<sup>e</sup>. All samples and clinical information were collected under Health Insurance Portability and Accountability Act compliance from study participants after obtaining written informed consent under clinical research protocols approved by the institutional review boards for each site. Demographic data was collected by self-report and clinical data by chart review. Blood was processed within 2 hours of collection according to established standard operating procedures. To predict RCC status, serum samples were collected at a single study site from patients diagnosed with RCC or benign renal mass prior to treatment. Definitive pathology diagnosis of RCC and cancer stage was made after resection. Outcome data was obtained through follow-up from 3 months to 5 years after initial treatment. To predict OVC status, plasma samples were collected from women with a suspicious pelvic mass prior to treatment. Definitive pathology diagnosis of ovarian cancer stage or benign mass was made after resection. CA-125 (Carbohydrate Antigen 125 also known as MUC16) was measured by a routine clinical laboratory assay. To generate RCC and OVC datasets, the SOMAmer based proteomic technology<sup>7</sup> is used to measure the concentration of a selected set of about 1000 proteins in Relative Fluorescence Unit. The CRC samples belong to a publicly available microarray dataset collected from gene expression omnibus (GEO), and referenced by accession number

<sup>e</sup>Due to conflict of interest, the datasets are not publicly available. Data requests should be sent to the last author of this paper.

*GSE17536* (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536>).<sup>10</sup>

Our RCC dataset contains 212 RCC samples from benign and 4 different stages of tumor. Expression levels of 1092 proteins are collected. The number of Benign, Stage 1, Stage 2, Stage 3 and Stage 4 tumor samples are 40, 101, 17, 24 and 31 respectively. Our OVC dataset contains 845 proteins' expressions for 248 samples across Benign and 3 different stages of ovarian cancer. The number of Benign, Stage 1, Stage 2 and Stage 3 tumor samples are 134, 45, 8 and 61 respectively. The public CRC microarray dataset (*GSE17536*) contains 177 samples from 4 different stages (Stage 1 to Stage 4) of CRC. Expression levels of 20125 genes are collected. Besides stage information, this dataset also has records for each patient, the binary valued information of "Cancer Recurrence" and "Death from Cancer". Out of 177 patients, 55 had recurrence of cancer and 68 died from cancer.

In order to group the genes using gene ontology terms, we use the web based tool "Database for Annotation, Visualization, and Integrated Discovery" (DAVID, <http://david.abcc.ncifcrf.gov/>).<sup>11</sup> There are a set of parameters that can be adjusted in DAVID based on which the functional classification is done. This whole set of parameters is controlled by a higher level parameter "Classification Stringency", which determines how tight the resulting groups are in terms of association of the genes in each group. In general, a "High" stringency setting generates less number of functional groups with the member genes tightly associated and more genes will be treated as irrelevant ones into an unclustered group. We set the stringency level to "Medium" which results in balanced functional groups where the association of the genes are moderately tight. The total number of groups based on CC annotations for RCC and OVC datasets are 56 and 23 respectively, and the number of groups for the CRC dataset is 520.

Besides using it for selecting informative single gene features, we use Overlapping Group Lasso to select the informative protein protein interactions. We download the binary protein protein interactions (PPI) data from HPRD (<http://www.hprd.org/>). For each feature group  $G_i$ , we identify the pairs of member genes of  $G_i$  whose products interact directly with each other in the PPI network. The set of all such pairs where both interacting partners are members of  $G_i$  forms a group. For a pair of interacting genes  $x_j$  and  $x_k$  in a group, we use their quadratic interaction term  $x_j x_k$  as their expression level. Usage of the quadratic interaction formulation in Overlapping Group Lasso helps us to integrate the resulting informative protein protein interactions into the formulation of QUIRE directly without any transformation. Thus the total number of groups are same in the case of interactions and single gene features. But the cardinality of each group and the expression levels of the members are different.

### 3.2. Experimental Design

We perform three stage-wise binary classification experiments using RCC samples: Classification of Benign samples from Stage 1 – 4 samples, Classification of Benign and Stage 1 samples from Stage 2 – 4 samples, and Classification of Benign, Stage 1, 2 samples from Stage 3, 4 samples. In the OVC dataset, *CA125* is a well-known marker in ovarian cancer.<sup>12</sup> Concentration of *CA125* is used to measure the progression of the disease. The suspicious cutoff level of *CA125* is 40 U/mL, meaning that concentration level above 40 of this marker might be

indicative of OVC. But *CA125* is not a good indicator of early detection of the disease onset, especially when the concentration of this biomarker is between 40 and 100.<sup>13</sup> So we use samples with *CA125* concentration level between 40 and 100 as our test set in this experiment. The remaining samples, with concentration of *CA125* below 40 and above 100 are used as training set. We perform the following experiments: Classification of Benign samples from Stage 1 – 3 samples, Classification of Benign, Stage 1 samples from Stage 2, 3 samples, and Classification of Benign, Stage 1, 2 samples from Stage 3 samples. On the CRC dataset, we perform binary classifications to predict whether there is disease-free survival in the follow-up time or not for cancer recurrence prediction and whether there is death from CRC across all pathological stages of the disease for death from colorectal cancer prediction.

### 3.3. Classification performance of QUIRE

In this section, we report systematic experimental results on classifying samples from different stages of RCC and OVC and in predicting CRC recurrence and death from CRC. In the first stage of QUIRE, we use Overlapping Group Lasso to identify the biologically relevant groups of features and pairwise protein interactions, which in turn, is used in the subsequent stage to explore the set of informative markers and quadratic interactions. However, for the RCC and OVC datasets, we do not use protein protein interactions for prediction purpose. This is because, these datasets include only selected marker proteins distributed sparsely across the protein interaction network and thus most of them do not interact with each other directly.

After we run Overlapping Group Lasso on the gene groups, we sort the genes based on the weight value assigned to it by the algorithm. We used cross validation to select the optional parameter  $m$  in QUIRE from  $\{100, 200, 300, 400, 500\}$ , and  $m = 200$  was selected for all our experiments. For classification of CRC samples, Overlapping Group Lasso on average selects 1000 PPIs as informative ones. We use this whole set of selected protein interactions as input to QUIRE to be considered besides the paired quadratic interactions.

The predictive performance of the markers and pairwise interactions selected by QUIRE is compared against the markers selected by Lasso, linear Support Vector Machine (SVM) and Overlapping Group Lasso. We use `glmnet`<sup>9</sup> and `Liblinear`<sup>14</sup> packages for implementation of Lasso and SVM respectively. We use the Group Lasso implementation (with overlapping groups) from.<sup>5</sup> The overall performance of the algorithms are shown in Figure 2. In this figure, we report average AUC (Area Under the Curve) score for ten runs of five-fold cross validation experiments for cancer stage prediction in RCC (Figure 2(A)) and for predicting cancer recurrence and death from cancer in CRC (Figure 2(C)). In five fold cross validation experiments, we divide the samples equally into five disjoint sets or folds. We keep one fold for testing. On the remaining four folds, we use Overlapping Group Lasso to identify the informative set of markers and protein protein interactions (for CRC). We train QUIRE on these four folds using these markers to identify the pairwise interactions and markers and use the set-aside test set for prediction purpose. For each run, this procedure is repeated for each of the five folds and average AUC score is reported for ten such runs. For OVC, we report average AUC score (Figure 2(B)) for predicting the cancer stage of the samples with intermediate levels of *CA125* (concentration of *CA125* is between 40 and 100) using the

remaining samples for training and informative feature selection. In cancer stage prediction

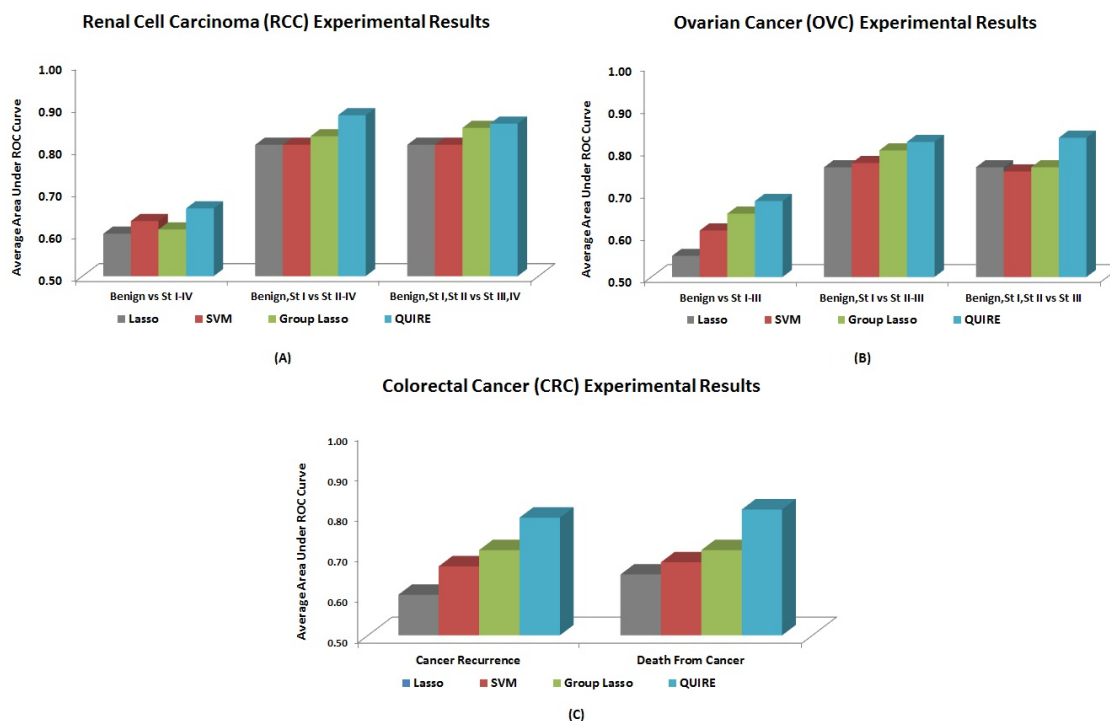


Fig. 2. Comparison of the classification performances of different feature selection approaches with QUIRE in identifying the different stages of (A) RCC, (B) OVC and (C) in predicting CRC recurrence and death from CRC. In (A) and (C), five fold cross validation is repeated ten times and average AUC score is reported. For (B), samples with *CA125* marker's expression level between 40 and 100 are used as test cases, while the remaining samples are used for training. This experiment is also repeated ten times and average AUC score is reported.

experiments for RCC and OVC, we see from Figure 2 that the combination of informative markers and pairwise interactions identified by QUIRE show better classification performance in every case, as compared to the markers selected by Lasso, SVM and Overlapping Group Lasso. For early detection of the diseases (classification of Benign, Stage 1 vs. rest of the samples), QUIRE achieves average AUC scores of 0.88 and 0.82 for RCC and OVC respectively. Overlapping group lasso shows next best performance with average AUC scores of 0.83 and 0.80 respectively. Lasso and SVM, which do not use any grouping or interaction information amongst the features, show the worst performance in all of the classification tasks apart from one. As QUIRE markers show consistently better performance across all the stages of RCC and OVC, they can be used for improved diagnosis and prognosis of these two different types of cancers. Also QUIRE helps better prediction of OVC progression for samples with intermediate levels of *CA125*; so it can be used by the physicians for early detection of this disease.

From Figure 2(C), we can see that gene-gene interactions help us better predict both CRC recurrence and death from CRC, as compared to the other feature selection mechanisms. In the events of cancer recurrence and death from cancer, the average AUC values achieved by features selected with QUIRE are 0.79 and 0.81 respectively, while markers identified by Overlapping Group Lasso show the next best performance with average AUC value of 0.71 in

both of these categories. Markers identified by Lasso show the worst performance in prediction of both of these events. The performance gap between QUIRE and the other three popular feature selection techniques hint to the fact that QUIRE can identify interactions that might help us better understand the mechanistic basis of CRC.

These experimental results show that QUIRE identifies markers and interactions that complement each other in such a way that they not only help better diagnosis and prognosis of cancer, but also can predict the advanced events of recurrence of cancer and survival after cancer with higher accuracy than other state-of-the-art algorithms. For each of these datasets, identification of informative pairwise interactions using brute-force enumerative technique is computationally impractical due to the huge dimensionality of the search space. QUIRE helps reducing this space by a large margin. The total running time of QUIRE is dominated by the Overlapping Group Lasso stage which takes around one hour to identify biologically relevant groups of genes and protein interactions in traditional desktop computers for the types of problems we study. After the dimensionality is reduced, QUIRE exhaustively enumerates all the pairwise interactions and use the protein interactions identified in the previous stage on this low dimensional space in a couple of minutes. QUIRE is computationally very fast considering that it identifies discriminative pairwise gene interactions at a genome-wide scale.

### 3.4. Informative QUIRE markers and interactions associated with cancer

Cancer is a genetic disease, which originates and develops through a process of mutations. Mutations in individual gene not only disrupts its own function, but also affects its interaction patterns with other genes. As complex diseases like cancer is a result of dysregulation in the interactions among the genes, researchers focus on identifying those relevant interactions to gain more insight into the molecular basis of the disease. On the CRC dataset, QUIRE selects about 120 quadratic interactions and single features in total on average as informative ones for both CRC recurrence and death from CRC. On the other hand, the average number of markers selected by Overlapping Group Lasso and Lasso on the same prediction tasks are about 1100 and 150 respectively. Therefore, Overlapping Group Lasso itself is unsuitable for biomarker discovery although it produced the second best performance.

An investigation of the pairwise interactions identified by QUIRE on CRC dataset reveals that many of these interactions are indeed relevant to the progression of cancer in general. Some of such interactions identified for prediction of CRC recurrence include *JAK2* - *LYN*,<sup>15</sup> Transforming growth factor beta (*TGFβ*) - *SMAD*,<sup>16</sup> Epidermal growth factor receptor (*EGFR*) - Caveolin (*CAV*),<sup>17</sup> *TP53* - TATA binding protein (*TBP*),<sup>18</sup> Connective tissue growth factor (*CTGF*) - Vascular endothelial growth factor (*VEGF*),<sup>19</sup> Edoglin (*ENG*) - Transforming growth factor beta receptor (*TGFβR*).<sup>20</sup> Further investigations of the interactions identified by QUIRE might reveal novel gene partners associated with cancer and thus lead to testable hypothesis.

Disturbance in pairwise interactions among the genes affects the pathways in which they are located in. Cancer pathways are a set of pathways dysregulations in which have been shown to be associated with initiation and progression of the disease. A pictorial view of the well-known cancer pathways can be found in the KEGG database(<http://www.genome.jp/kegg/pathway/hsa/hsa05200.html>).<sup>21</sup> We per-



form a pathway enrichment analysis where we test if the set of the markers and interactions identified by QUIRE on the CRC dataset reside in the cancer pathways. As part of this experiment, we first use the partner genes identified by QUIRE as part of the informative interactions while predicting CRC recurrence. We use DAVID to identify the statistically significant pathways that are enriched in these genes. An investigation of the enriched pathways returned by DAVID indicates that many of them are indeed responsible for cancer or related to functions dysregulation in which results in cancer. Some of such KEGG pathways include Apoptosis (p-value  $4.7 \times 10^{-4}$ ), Focal adhesion (p-value  $3 \times 10^{-3}$ ), Cell adhesion molecules (p-value  $9.2 \times 10^{-4}$ ), p53 signaling pathway (p-value  $1.3 \times 10^{-2}$ ), Gap junction (p-value  $1.3 \times 10^{-2}$ ), MAPK signaling pathway (p-value  $4.5 \times 10^{-2}$ ), ErbB signaling pathway (p-value  $5.8 \times 10^{-2}$ ), Cell cycle (p-value  $6.6 \times 10^{-2}$ ), Pathways in Cancer (p-value  $7.2 \times 10^{-4}$ ), Colorectal cancer (p-value  $10^{-3}$ ). Repeating the same analysis on the interacting partners identified by QUIRE while predicting “Death from CRC” result in identification of similar pathways (data not shown).

### 3.5. *Significance of feature interactions in QUIRE*

We also perform classification experiments to observe the performance of PPIs and informative single features on predicting CRC recurrence and death from CRC without quadratic feature interactions. For this experiment, we use the single gene markers and the PPIs selected by Overlapping Group Lasso as input to QUIRE and enumeration of the pairwise interactions among the marker genes is avoided. For ten runs of five fold cross validation experiment on this modified feature set, we observe average AUC score of 0.71 for both classification tasks. If we only use informative single features with the same experimental setting, the average AUC score we got is 0.70. These results show that besides physical interactions and single features, indirect higher level interactions among the informative genes must be taken into account to understand the basic mechanism of complex diseases.

## 4. Conclusion

In this paper, we propose a computational approach, QUIRE, to identify combinatorial interactions among the informative genes in complex diseases, like cancer. Our algorithm uses Overlapping Group Lasso to identify functionally relevant gene markers and protein interactions associated with cancer. It then explores the pairwise interactions among these relevant genes within this reduced space exhaustively and the selected pairwise physical protein interactions to discover the combination of individual markers and gene-gene interactions that are informative for prediction of the disease status of interest. The application of QUIRE on three different types of cancer samples collected using two different techniques shows that our approach performs significantly better than the state-of-the-art feature selection methods such as Lasso and SVM for biomarker discovery while selecting a smaller number of features, and it also shows that our approach can capture discriminative interactions with high relevance to cancer progression. Further investigations show that QUIRE can identify markers and interactions that have been associated previously with pathways associated with cancer. Moreover, the good performance of QUIRE on the CRC dataset suggests that applications of QUIRE on genome-wide microarray experimental data can be used to help prioritize SOMAmer design

for blood-based cancer diagnosis. And QUIRE applied to blood-based experimental data has the great potential to impact the field of practical medical diagnosis.

## Acknowledgement

We thank Hans Peter Graf for valuable comments and discussions.

## References

1. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee and T. Ideker, *Mol. Syst. Biol.* **3**, p. 140 (2007).
2. S. A. Chowdhury, R. K. Nibbe, M. R. Chance and M. Koyuturk, *J. Comput. Biol.* **18**, 263 (Mar 2011).
3. S. Lee and E. P. Xing, *Bioinformatics* **28**, i137 (June 2012).
4. M. Yuan and Y. Lin, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49 (2006).
5. L. Jacob, G. Obozinski and J.-P. Vert, Group lasso with overlap and graph lasso, in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09* (ACM, New York, NY, USA, 2009).
6. H. Yu, P. Braun, M. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis *et al.*, *Science* **322**, 104 (2008).
7. L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, J. Carter, A. B. Dalby, B. E. Eaton and T. Fitzwater *et al.*, *PLoS ONE* **5**, p. e15004 (12 2010).
8. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, pp. 267 (1996).
9. J. H. Friedman, T. Hastie and R. Tibshirani, *Journal of Statistical Software* **33**, 1 (2 2010).
10. J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, S. Eschrich, C. Kis, S. Levy, M. K. Washington, M. J. Heslin, R. J. Coffey, T. J. Yeatman, Y. Shyr and R. D. Beauchamp, *Gastroenterology* **138**, 958 (Mar 2010).
11. G. Dennis Jr, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane and R. Lempicki, *Genome Biol* **4**, p. P3 (2003).
12. K. S. Suh, S. W. Park, A. Castro, H. Patel, P. Blake, M. Liang and A. Goy, *Expert Rev. Mol. Diagn.* **10**, 1069 (Nov 2010).
13. E. L. Moss, J. Hollingworth and T. M. Reynolds, *Journal of clinical pathology* **58**, 308 (March 2005).
14. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang and C. J. Lin, *Journal of Machine Learning Research* **9**, 1871 (2008).
15. A. Samanta, S. Chakraborty, Y. Wang, H. Kantarjian, X. Sun, J. Hood, D. Perrotti and R. Arlinghaus, *Oncogene* **28**, 1669 (2009).
16. W. Grady, *Clinical cancer research* **11**, 3151 (2005).
17. K. Dittmann, C. Mayer, R. Kehlbach, H. Rodemann *et al.*, *Mol Cancer* **7**, 17 (2008).
18. D. Crighton, A. Woiwode, C. Zhang, N. Mandavia, J. Morton, L. Warnock, J. Milner, R. White and D. Johnson, *The EMBO journal* **22**, 2810 (2003).
19. I. Inoki, T. Shiomi, G. Hashimoto, H. Enomoto, H. Nakamura, K. Makino, E. Ikeda, S. Takata, K. Kobayashi and Y. Okada, *The FASEB Journal* **16**, 219 (2002).
20. E. Fonsatti, M. Altomonte, P. Arslan and M. Maio, *Current drug targets* **4**, 291 (2003).
21. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi and M. Tanabe, *Nucleic acids research* **40**, D109 (2012).

## MULTIPLEX META-ANALYSIS OF MEDULLOBLASTOMA EXPRESSION STUDIES WITH EXTERNAL CONTROLS

ALEXANDER A. MORGAN

*Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: alexmo@stanford.edu*

MATTHEW D. LI

*Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: mdli@stanford.edu*

ACHAL S. ACHROL

*Neurosurgery  
Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: achrol@stanford.edu*

PURVESH J. KHATRI

*Institute for Immunity, Transplant and Infection  
Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: pkhatri@stanford.edu*

SAMUEL H. CHESHER

*Neurosurgery, Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: cheshier@stanford.edu*

We propose and discuss a method for doing gene expression meta-analysis (multiple datasets) across multiplex measurement modalities measuring the expression of many genes simultaneously (e.g. microarrays and RNAseq) using external control samples and a method of heterogeneity detection to identify and filter on comparable gene expression measurements. We demonstrate this approach on publicly available gene expression datasets from samples of medulloblastoma and normal cerebellar tissue and identify some potential new targets in the treatment of medulloblastoma.

### 1. Background

Highly multiplex gene expression studies using microarrays or RNAseq are very useful for probing the functional genomics of a wide range of biological processes. The analysis of gene expression data typically involves some sort of comparison between samples. Often this comparison is between samples drawn from different conditions. Possible comparisons include samples from tissue treated with different pharmaceuticals, samples drawn from different tissue types or different developmental stages, or samples taken from diseased tissue compared with samples taken from healthy tissue. Many cancer types have been the focus of extensive gene expression analysis, both to identify new molecular subtypes of cancer by comparing different cancer samples, one to another, but also to compare the gene expression differences between healthy tissue and cancerous tissue to help elucidate the molecular processes in different forms of neoplasia. Comparing gene expression levels across thousands of genes in healthy tissue and cancer is a powerful tool in investigating cancer pathogenicity and the development of new pharmacological agents to treat cancer. In many types of cancer such as breast or prostate cancer, it is standard practice during therapeutic surgical removal of a tumor to remove an accompanying portion of nearby healthy tissue surrounding the tumor (i.e. the margin). This provides material from which paired mRNA can be extracted for comparison between healthy and neoplastic tissue.

However, for many types of cancer this is not possible. For primary brain tumors, surgical resection of the tumor is often a balanced tradeoff between removing as much neoplastic material as possible, while leaving as much essential (eloquent) tissue structures as possible to maintain as much function as possible. In aggressive brain tumors, the border of the malignancy and the healthy tissue may not be distinct or clearly separable. For obvious ethical reasons, it is not possible to obtain brain biopsies of healthy tissue from volunteers, unlike tissue types such as skin or blood. This makes having samples for multiplex comparison of gene expression between tumor and healthy brain tissue very difficult.

Medulloblastoma is a type of highly malignant primary brain tumor that typically originates in the cerebellum below the tentorium cerebelli in the posterior fossa. Gene expression studies of samples taken from medulloblastoma solid tumor tissue have focused on identifying different genomic subtypes of medulloblastoma that might lead to new targeted therapies or stratify prognosis [1,2]. Although it might be possible to do a post-mortem analysis of gene expression changes between samples drawn from tumor tissue and nearby brain in those unfortunate individuals who succumb to the disease, most victims of medulloblastoma are treated with radiation, chemotherapy or both, which can cause dramatic gene expression changes in both tumor and non-neoplastic tissue, making a true comparison of tumor with “normal” tissue difficult. Some of the only gene expression datasets of healthy normal brain tissue come from samples taken from freshly deceased cadavers, often from individuals tragically killed in accidents who pre-arranged to donate biological samples to research or whose families do so on their behalf.

Recent developments in techniques of multiplex meta-analysis have led to techniques that synthesize multiple highly multiplex gene expression studies (e.g. microarray or RNAseq) to help remove batch effects, increase statistical power, and identify differences more likely to be biologically relevant and to be reproduced in subsequent studies [3–6]. In short, these approaches typically involve two steps, one is to identify if the measurements of gene expression across studies are even comparable, or if there is too much variation. The second step is to develop some overall estimate of the relative variation in gene expression across the studies and its statistical significance, against the typical null hypothesis of no difference in underlying expression between conditions.

One possible way to address this problem of gene expression samples without matched controls is to find a way to identify genes expression profiles which look the same within datasets studying a particular condition (e.g. medulloblastoma and healthy cerebellar tissue), and then look for genes that then vary between datasets. To make this intuition more formal, we propose using a statistical measure of heterogeneity across datasets for medulloblastoma and healthy cerebellum respectively to identify genes that are consistently expressed at an equivalent level within the datasets studying each condition (i.e., low heterogeneity implies homogeneity of expression). At the same time, we compute a meta-estimate of effect (expression level) with an appropriate meta-estimate of a confidence interval in that expression level across datasets and compare these two differences between conditions. Figure 2 shows some contrasting patterns of expression across datasets that demonstrate these concepts pictorially.

In order to investigate this concept further, we searched through the Gene Expression Omnibus (GEO) [7] to identify publicly available datasets of gene expression of medulloblastoma and

healthy, normal cerebellar tissue. To make the best comparison possible, we focused on control brain samples from the cerebellum. Gene expression samples were excluded if they were associated with a particular diagnosis (e.g. Huntington's disease) or a drug treatment. We obtained a total of 191 cerebellar control microarrays, and a total of 414 medulloblastoma microarrays for a total of 605 microarrays. We also collected a dataset of 20 microarrays on brain aging to compare differences in gene expression in the tumor samples with normal brain aging. The datasets we collected are summarized in Table 1. With any large meta-analysis, not all datasets are completely consistent in their methodology or content. The Fiaschetti20011 mRNA is from medulloblastoma tissue culture, not primary tumor tissue, and the Remke2011 and Northcott2012 datasets share some overlap in the tumor source for 15 samples (~5% of the Northcott2012 dataset), but these were processed at different times on different microarray platforms, and we consider them as

Table 1. Gene expression datasets used in this paper. For the multiplex meta-analysis of gene expression in medulloblastoma, four studies of medulloblastoma and four studies of healthy cerebellar cortex were synthesized. An additional dataset of gene expression in the brain as a function of age (individuals age 26-73 were used) was also used to compare gene expression changes in aging against gene expression changes found in medulloblastoma. Note that the GPL570 platform (Affymetrix U133 Plus 2.0) has been used for both some control datasets and some medulloblastoma datasets, setting a point of relative comparison between conditions.

Dataset	Number of Arrays	Sample Type	Gene Expression Series	Pubmed ID	Publication	GEO Platform
Gibbs2010	146	Cerebellar Control	GSE15745	20485568	JR Gibbs, PLoS Genetics, 2010	GPL6104
Hodges2006	27	Cerebellar Control	GSE3790	16467349	A Hodges, Hum Mol Genet, 2006	GPL96
Roth2006	9	Cerebellar Control	GSE3526	16572319	RB Roth, Neurogenetics 2006	GPL570
Roth2007	9	Cerebellar Control	GSE7307		Unpublished	GPL570
Fiaschetti2011	3	Medulloblastoma	GSE22139	21317922	G Fiaschetti, Oncogene, 2011	GPL570
Kool2008	62	Medulloblastoma	GSE10327	18769486	M Kool, PLoS One, 2008	GPL570
Northcott2012	285	Medulloblastoma	GSE37382	22832581	PA Northcott, Nature, 2012	GPL11532
Remke2011	64	Medulloblastoma	GSE28245	21911727	M Remke, J Clin Oncol, 2011	GPL6480
AgingCortex	20	Frontal Cortex	GSE1572	15190254	T Lu, Nature 2004	GPL8300

independent datasets.

## 2. Results of Analysis

For each microarray dataset, the expression data was obtained from the Gene Expression Omnibus (GEO) [7] and quantile normalized. The probe identifiers for each sample were mapped to Entrez Gene identifiers using AILUN [8]; probes that mapped to multiple identifiers were excluded. If multiple probes mapped to a single gene in a study, the median expression of all probes was taken for that gene. The expression levels of 7724 different genes were measured in all medulloblastoma and cerebellar datasets, but there was also some missing expression levels in individual microarrays, leaving us with 7015 genes with sufficient data to compare expression across

datasets. The genes were quantile normalized the genes across all microarrays together to get a normalized expression level across datasets.

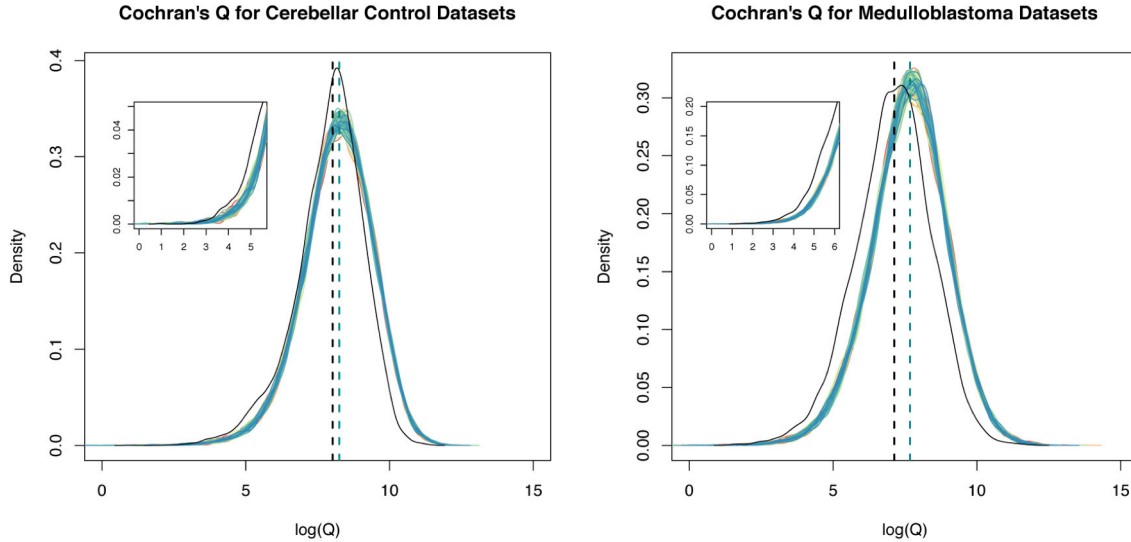


Figure 1: Smoothed histograms of the distribution of Cochran's Q across 100 randomizations (spectrum of colors, visible in blue) compared with the distribution of the measure of heterogeneity in the actual samples in black. The median Q for the actual sample labels is shown in the vertical dashed black line, and then the median for the 100 randomization tests is shown in the blue vertical line to the right. The inset in each panel highlights that at the lower levels of heterogeneity there is substantially more genes showing expression homogeneity in the real data compared with the randomized samples (black line lies above the collection of colored lines, one for each of the 100 randomization).

We then performed a meta-analysis for each gene in the cerebellar and medulloblastoma datasets separately. For each gene in each dataset, we computed the mean expression rank and the standard error of that mean. We used the meta-analysis method proposed by Hedges, et al. [9] which creates a meta-effect estimate based on a random effects linear model, weighting the contribution of the effect (rank expression level) estimate from each included dataset inversely with the standard error of that estimate. This method has been widely used for microarray meta-analysis [5,10,11]. We computed a meta-effect size estimate and we computed a measure of heterogeneity, Cochran's Q [12] for each gene across the cerebellar and medulloblastoma datasets, respectively. This gave us a consensus measure of relative expression of each gene across the cerebellar studies, a confidence interval around that estimate, and a measure of how heterogeneous/homogenous expression of that gene was across studies. We created the corresponding meta-statistics for expression across the medulloblastoma studies.

By identifying the genes with the lowest 20% of heterogeneity in the cerebellar datasets and the genes with the lowest 20% of lowest heterogeneity in the medulloblastoma datasets, and then taking the intersection, we were left with 318 genes. These represent 318 genes that are consistently expressed at about the same level across all the cerebellar datasets and consistently expressed at about the same level across all the medulloblastoma datasets, but may differ in expression between the two conditions. To test the robustness of this result, we performed 100

random reshufflings of the dataset labels and repeated this analysis. Figure 1 shows that there was more heterogeneity in the randomized samples compared to the actual datasets. The median heterogeneity was always less in the actual data compared to the randomized samples. This suggests that it is possible to find a highly specific set of genes that are more homogenous in expression across datasets for each of the two conditions than random chance.

Of the 318 genes homogeneous in the datasets for both conditions (lower 20% of homogeneity in cerebellar and medulloblastoma datasets), 20 varied in meta-expression difference between medulloblastoma datasets and cerebellar controls greater than the computed meta-confidence interval (Equation 4). In the 100 randomized reshufflings of the dataset labels, there were a median of 8 genes (mean 7.97) genes which met the set criteria for heterogeneity and significant different, suggesting a false discovery rate of 40%.

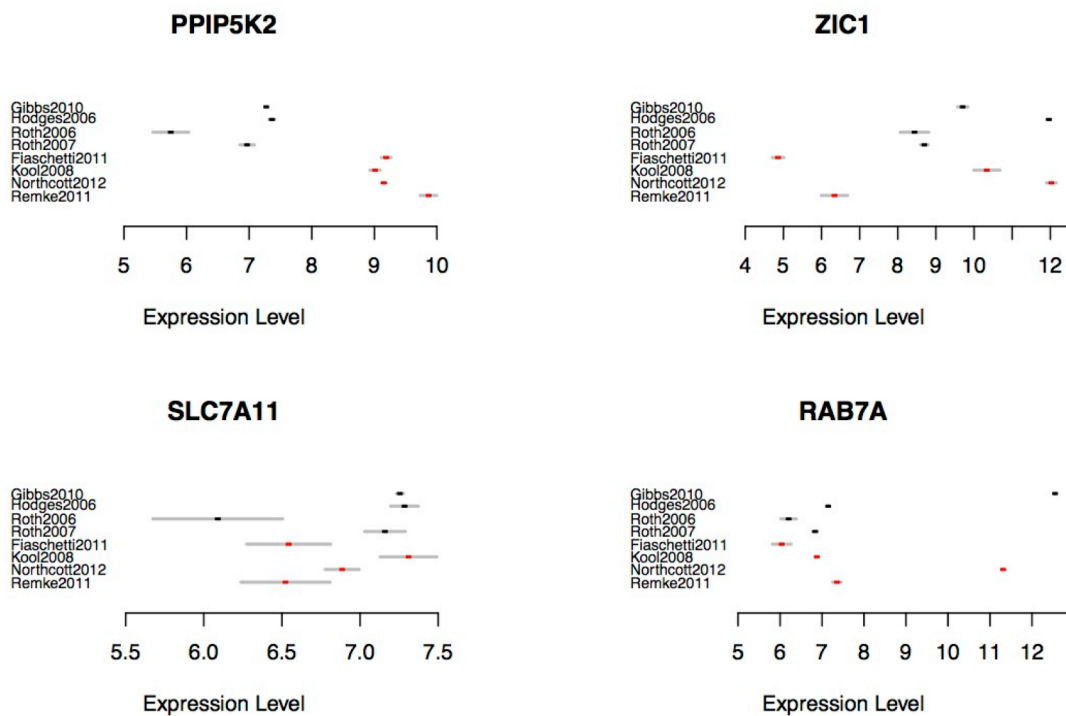


Figure 2: Four forest plots of gene expression across datasets. The four cerebellar controls are shown in black, the medulloblastoma datasets in red. The expression pattern of PPIP5K2 shows narrow confidence intervals and also low heterogeneity, as the expression values across medulloblastoma datasets are similar to one another and the same is true for the controls. PPIP5K2 also shows a pronounced difference in expression between these two conditions; this is the type of expression pattern across datasets that shows strong evidence of increased expression in medulloblastoma. In contrast ZIC1, shows high heterogeneity and thus would be filtered out even though the meta-estimate would suggest substantially lowered expression in medulloblastoma, as the expression levels within datasets studying medulloblastoma vary widely, while the confidence intervals around the expression measurements are also narrow. SLC7A11 shows low heterogeneity, as the expression levels within datasets have broad confidence intervals that nearly overlap, but there is also no significant difference between medulloblastoma and cerebellar controls and thus would be filtered out. RAB7A both has high heterogeneity and low meta-estimate of the difference between medulloblastoma and cerebellar controls and would then be filtered out for both reasons.

Of the 20 genes found to be differentially expressed, two were located on chromosome X or Y and may reflect a gender imbalance in samples and were removed, the remaining 18 are shown in Table 2. Figure 3 shows the expression pattern of ENC1 which encodes Ectodermal-Neural Cortex 1, a gene induced by P53 and which interacts with the Retinoblastoma protein. The PRKAR2B gene shows a similar pattern of expression (Figure 3); it is greatly increased in expression in medulloblastoma compared to the healthy cerebellar tissue. PRKAR2B encodes a regulatory subunit that plays a role in regulation of energy metabolism in the cell.

The filtering by heterogeneity is intended to limit the false positive rate, but we might want to focus on sensitivity and expand the coverage of our meta-analysis. Ignoring the filtering by heterogeneity, we can focus on the top genes whose meta-effect estimate significantly differs

Table 2. Top differentially expressed genes with consistent homogeneous expression levels across studies within each condition (medulloblastoma or cerebellar control).

Symbol	Map Location	Description	Cerebellar Control Q	Medullo Q	Cerebellar Expression Level	Medullo Expression Level	Differential Expression
B3GALNT1	3q25	beta-1,3-N-acetylgalactosaminyltransferase 1 (globoside blood group)	501.1	54.5	6.89 ± 0.22	8.00 ± 0.13	1.10 ± 0.25
DNAJC1	10p12.31	DnaJ (Hsp40) homolog, subfamily C, member 1	533.1	108.7	7.07 ± 0.11	8.25 ± 0.10	1.18 ± 0.15
ENC1	5q13	ectodermal-neural cortex 1 (with BTB-like domain)	559.7	4.3	7.27 ± 0.32	8.92 ± 0.11	1.64 ± 0.34
FAM115A	7q35	family with sequence similarity 115, member A	37.8	84.7	7.50 ± 0.07	9.03 ± 0.15	1.52 ± 0.16
FZD7	2q33	frizzled family receptor 7	190.9	159.1	8.52 ± 0.27	7.42 ± 0.47	-1.10 ± 0.54
LBH	2p23.1	limb bud and heart development homolog (mouse)	384.0	66.2	6.82 ± 0.17	8.06 ± 0.21	1.24 ± 0.27
LMNB1	5q23.2	lamin B1	251.9	155.2	6.86 ± 0.14	8.94 ± 0.24	2.08 ± 0.27
LRIF1	1p13.3	ligand dependent nuclear receptor interacting factor 1	169.6	128.1	6.91 ± 0.09	8.36 ± 0.17	1.45 ± 0.19
MORC3	21q22.13	MORC family CW-type zinc finger 3	511.0	50.8	7.35 ± 0.33	8.52 ± 0.09	1.17 ± 0.34
OSBPL8	12q14	oxysterol binding protein-like 8	561.9	38.6	7.88 ± 0.31	9.42 ± 0.06	1.54 ± 0.32
PAX6	11p13	paired box 6	451.4	144.4	9.52 ± 0.32	7.98 ± 0.51	-1.54 ± 0.60
PODXL	7q32-q33	podocalyxin-like	201.8	63.7	7.16 ± 0.17	8.97 ± 0.15	1.82 ± 0.23
PPIP5K2	5q21.1	diphosphoinositol pentakisphosphate kinase 2	153.7	120.2	6.95 ± 0.10	9.30 ± 0.12	2.34 ± 0.15
PRKAR2B	7q22	protein kinase, cAMP-dependent, regulatory, type II, beta	165.8	32.3	7.09 ± 0.09	9.34 ± 0.17	2.25 ± 0.19
SACS	13q12	spastic ataxia of Charlevoix-Saguenay (sacsin)	145.4	128.6	7.08 ± 0.13	9.04 ± 0.29	1.96 ± 0.32
STMN1	1p36.11	stathmin 1	188.0	177.3	7.53 ± 0.10	8.77 ± 0.12	1.23 ± 0.15
TRMT11	6q11.1-q22.33	tRNA methyltransferase 11 homolog ( <i>S. cerevisiae</i> )	183.4	103.4	7.24 ± 0.17	8.40 ± 0.17	1.17 ± 0.24
ZFP36	19q13.1	zinc finger protein 36, C3H type, homolog (mouse)	306.7	121.1	8.00 ± 0.64	6.94 ± 0.33	-1.06 ± 0.72



between cerebellar and medulloblastoma datasets (354 genes were found to be increased in medulloblastoma compared to the controls when ignoring heterogeneity), and do an analysis for functional enrichment of DAVID [13]. This analysis shows that these genes are highly over-enriched relative to the background of the genes measured across all datasets in such functional annotations such as cell cycle ( $10^{-19}$ , Benjamini corrected p-value for multiple hypothesis testing), M phase of mitosis ( $10^{-15}$ ), cell division ( $10^{-11}$ ) and being involved with cancer ( $10^{-5}$ ), as might be expected. The 483 genes with lowered expression in medulloblastoma compared to healthy cerebellum (again ignoring the heterogeneity criterion) were highly enriched for genes annotated to be involved in the synapse ( $10^{-11}$ ), transmission of nerve impulses ( $10^{-9}$ ), synaptic transmission ( $10^{-9}$ ), the transport of neurotransmitters ( $10^{-7}$ ), psychiatric disorders ( $10^{-6}$ ), and the regulation of nerve impulse transmission ( $10^{-5}$ ). All this is clearly in line with our understanding of medulloblastoma replacing cells essential to the neurological functioning of the brain with cells focused on rapid replication and suggests that this multiplex meta-analysis approach for using external controls is producing differentially expressed genes with biological relevance to our understanding of medulloblastoma.

To address one of the larger potential biases in our datasets, we also investigated the relationship to differential expression of genes due to normal aging. Although we don't have age information for all of our samples, medulloblastoma is a type of neurological cancer that preferentially targets younger individuals. At the same time, most of the healthy cerebellar brain samples are likely from recently deceased older adults, so there may be a bias toward discovering genes which vary in expression in the cerebellum due to development and aging. We do not have access to expression datasets from healthy cerebellar tissue in children of different ages; however, we do have some expression data on tissue from aging brains in adults. If we look at the dataset on aging of brain samples taken from the frontal cortex of samples taken from recently deceased adults, we can look to see if there is any evidence that the gene expression differences between medulloblastoma and healthy cerebellum could be attributable to simple differences in age. This is not a perfect comparison, but simple compromise based on what data we have available.

Using a dataset from Lu, et al. [14] obtained from the Gene Expression Omnibus [7], we obtained gene expression levels from microarrays made from samples from the frontal cortex from twenty individuals aged 26-73. The original dataset contains additional expression measurements from older brain samples, but we wanted to focus our analysis on gene expression changes in younger adults, and our exploratory analysis found that when using all the data the changes we identified were substantially driven by the samples drawn from much older individuals. Data was again quantile normalized, and we simply looked at the significance test for the Pearson correlation between age and gene expression level. The significance estimates were adjusted using the Benjamini-Hochberg method for addressing multiple hypothesis testing. Examples are shown in Figure 4.

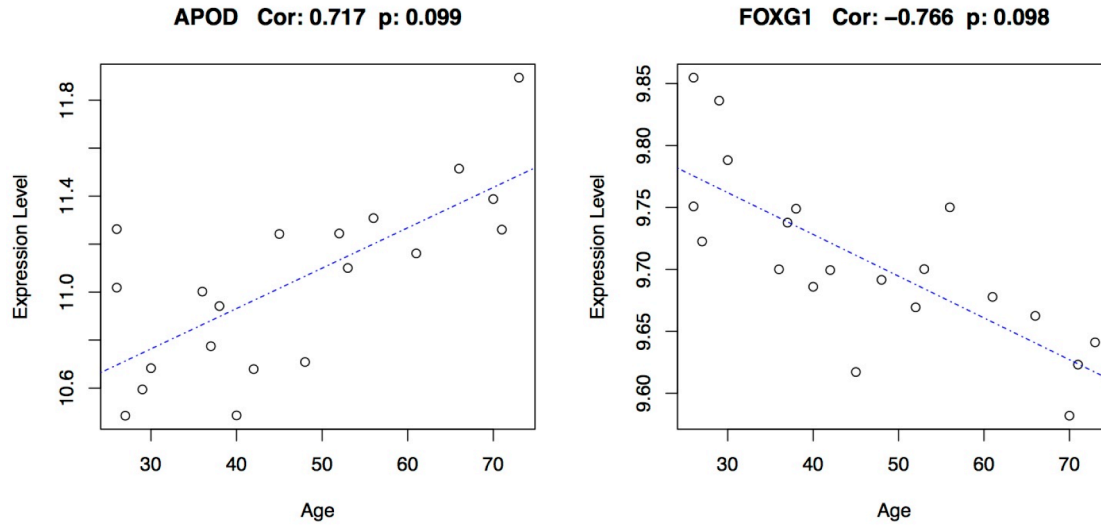


Figure 4: Two examples of genes found to vary in expression with increasing age in the frontal cortex. The expression level is plotted on the vertical axis and the age of the individual on the horizontal. The correlation coefficient and the corrected p-value appear at the top. APOD codes for apolipoprotein D, and FOXG1 codes for a member of the forkhead transcription factors that plays a role in brain development.

None of the same genes were found to be significantly (adjusted  $p < 0.1$ ) differentially expressed with age that were identified in the multiplex meta-analysis of the medulloblastoma and cerebellar controls. Although this does not prove that gene expression changes that we identified in medulloblastoma are not due to differences in the age of individuals sampled, it does suggest that we are not identifying changes in gene expression solely based on the most dramatic age-related changes.

### 3. Statistical Methods

Computations were done using the 'meta' package in R [15]. For each gene with an average expression level  $f_i$  in each dataset,  $i$ , the meta-expression estimate of that gene within all the datasets studying a given condition was estimated by taking a weighted average of the expression levels across the gene level median of the probeset expression levels, where the weighting is the inverse of the sum of the within study variance and the estimate of the variance in expression between datasets. For a given gene with expression of  $f_i$  and a within dataset variance of  $v_i$  in each of  $k$  datasets, the estimate of meta fold-change,  $M$ , for that given gene is shown in Eq. (1) and as described by Hedges & Olkin [9]. This analysis was done using quantile normalized gene expression levels.

$$M = \frac{\sum_{i=1}^k w_i f_i}{\sum_{i=1}^k w_i} \quad (1)$$

The weight for the contribution from each dataset,  $i$ , is given by adding the estimate of the variance between each dataset and within each dataset and inverting, as shown in Eq. (2).

$$w_i = T^2 + v_i \quad (2)$$

We estimate the between study variance,  $T^2$ , using the method of moments, Eq. (3)

$$T^2 = \frac{\sum_{i=1}^k \frac{f_i^2}{v_i} - \frac{\left(\sum_{i=1}^k \frac{f_i}{v_i}\right)^2}{\sum_{i=1}^k \frac{1}{v_i}} - k + 1}{\sum_{i=1}^k \frac{1}{v_i} - \frac{\sum_{i=1}^k \frac{1}{v_i^2}}{\sum_{i=1}^k \frac{1}{v_i}}} \quad (3)$$

The confidence intervals and then p-values for the meta fold-change,  $M$ , are computed from the estimate of the variance,  $v_M$ , which is computed from the inverse weights, Eq. (4).

$$v_M = \frac{1}{\sum_{i=1}^k w_i} \quad (4)$$

The homogeneity test statistic,  $Q$ , is computed by Eq. (5)

$$Q = \sum_{i=1}^k w_i f_i^2 - \frac{\left(\sum_{i=1}^k w_i f_i\right)^2}{\sum_{i=1}^k w_i} \quad (5)$$

#### 4. Discussion

We have presented a possible method for the multiplex meta-analysis of gene expression with external controls amenable for use in gene expression studies of some types of cancer, and have presented a set of genes differentially expressed in medulloblastoma compared to cerebellar control tissue. There are considerable batch effect differences that usually make directly comparing two gene expression datasets for differences in expression challenging or impossible. There are also significant differences between gene expression platforms that make a single cross platform analysis impossible. However, the power of looking at multiple studies enables investigation of shared features across datasets to identify commonalities of expression that enable comparison of differences between collections of datasets. We have only begun to scratch the surface of what is possible using the vast resources of the constantly expanding publicly available

data on gene expression. Additional strategies and tools for merging data across varying datasets will be crucial for leveraging the full power of all this data.

Multiplex gene expression measurement modalities are not the only datasets in need of such approaches. For example, the analysis of data drawn from sequencing often is based on a comparison against shared or pooled controls that has its own biases as a semi-external control set. Other highly multivariate (multiplex) measurement modalities will have similar problems. Often when doing such analyses we are interested in an analysis with very high specificity, such as identifying new biomarkers or drug targets, and it is acceptable to filter aggressively, such as by requiring very high levels of homogeneity within datasets of a particular condition, making such an approach tenable. A further investigation of approaches would include a greater examination of non-parametric, rank based approaches, as have been previously investigated for comparing against external controls [16]. It is also possible to use much larger datasets with existing included controls (such as other forms of cancer) to demonstrate accuracy and consistency of results across a variety of cancers and other pathologies or use information about heterogeneity of expression across large numbers of datasets [17].

Although a false discovery rate estimated at 40% may seem unimpressive, it is also highly context dependent. To go from a list of tens of thousands of potential genes, down to a few dozen, with only half of them potentially being false positives may have use in many applications, including biomarker development. Also, the general approach of filtering for genes of within sample class heterogeneity could be used with RNAseq data, which should have substantially less platform variability, but still has experimental and technical biases which would confound direct sample to sample comparison of expression [18].

Another important avenue of future investigation is to look closer at molecular subtypes of medulloblastoma separately. The original research work that provided these medulloblastoma expression datasets identified several clinically different subtypes with characteristic gene expression profiles [1]. Our analysis grouped all the medulloblastoma samples together, looking only at shared properties in expression patterns, but this opens up exciting new possible avenues of hierarchical meta-analytic methodology development.

We hope this work can lead to greater insight into the genomic and molecular pathogenicity of aggressive primary brain tumors like medulloblastoma, and although it will be only one part of future large scale data integration across experimental modalities, it will facilitate further methods of investigations in the absence of custom made control data. The full data from these analyses is available on request from the authors (alexmo@stanford.edu).

## 5. Acknowledgments

This work was supported by Stanford University School of Medicine's Medical Scholars Program.

## References

- [1] P. A. Northcott, D. J. H. Shih, M. Remke, Y.-J. Cho, M. Kool, C. Hawkins, C. G. Eberhart, A. Dubuc, T. Guettouche, Y. Cardentey, E. Bouffet, S. L. Pomeroy, M. Marra, D. Malkin, J. T. Rutka, A. Korshunov, S. Pfister, and M. D. Taylor, *Acta Neuropathologica* **123**, 615 (2012).
- [2] M. D. Taylor, P. A. Northcott, A. Korshunov, M. Remke, Y.-J. Cho, S. C. Clifford, C. G. Eberhart, D. W. Parsons, S. Rutkowski, A. Gajjar, D. W. Ellison, P. Lichter, R. J. Gilbertson, S. L. Pomeroy, M. Kool, and S. M. Pfister, *Acta Neuropathologica* **123**, 465 (2012).
- [3] A. A. Morgan, P. Khatri, R. H. Jones, M. M. Sarwal, and A. J. Butte, *BMC Bioinformatics* **11 Suppl 9**, S6 (2010).
- [4] A. Campain and Y. H. Yang, *BMC Bioinformatics* **11**, 408 (2010).
- [5] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, *PLoS Med* **5**, e184 (2008).
- [6] F. Hong and R. Breitling, *Bioinformatics* **24**, 374 (2008).
- [7] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, *Nucleic Acids Res* **35**, D760 (2007).
- [8] R. Chen, L. Li, and A. J. Butte, *Nat Methods* **4**, 879 (2007).
- [9] L. V Hedges and I. Olkin, *Statistical Methods for Meta-analysis* (1985).
- [10] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, *Bioinformatics* **21**, 4348 (2005).
- [11] A. A. Morgan, P. Khatri, R. H. Jones, M. M. Sarwal, and A. Butte, in *AMIA Summit on Translational Bioinformatics* (San Francisco, 2010).
- [12] Cochrane Collaboration, *Cochrane Handbook for Systematic Reviews of Interventions* (2008).
- [13] G. Dennis Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, *Genome Biol* **4**, P3 (2003).
- [14] T. Lu, Y. Pan, S. Y. Kao, C. Li, I. Kohane, J. Chan, and B. A. Yankner, *Nature* **429**, 883 (2004).
- [15] G. Schwarzer, *Package "Meta"* (University of Freiburg, 2013).
- [16] H. Haerberle, J. T. Dudley, J. T. C. Liu, A. J. Butte, and C. H. Contag, *Neoplasia* (New York, N.Y.) **14**, 666 (2012).
- [17] A. A. Morgan, J. T. Dudley, T. Deshpande, and A. J. Butte, *Physiol Genomics* **40**, 128 (2010).
- [18] K. D. Hansen, R. A. Irizarry, and Z. Wu, *Biostatistics* (Oxford, England) **13**, 204 (2012).

## COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY

S. JOSHUA SWAMIDASS

*Department of Pathology and Immunology  
Washington University School of Medicine  
St. Louis, Missouri 63130  
Email: swamidass@wustl.edu*

ZHIYONG LU<sup>†</sup>

*National Center for Biotechnology Information (NCBI)  
Bethesda, MD, 20894 USA  
Email: zhiyong.lu@nih.gov*

PANKAJ AGARWAL

*Systematic Drug Repositioning, Computational Biology  
GlaxoSmithKline Pharmaceuticals R&D  
King of Prussia, PA 19406, USA  
Email: pankaj.agarwal@gsk.com*

ATUL J. BUTTE<sup>¶</sup>

*Division of Systems Medicine  
Stanford University School of Medicine  
Stanford, CA 94305, USA  
Email: abutte@stanford.edu*

Despite increasing investments in pharmaceutical R&D, there is a continuing paucity of new drug approvals. Drug discovery continues to be a lengthy and resource-consuming process in spite of all the advances in genomics, life sciences, and technology. Indeed, it is estimated that about 90% of the drugs fail during development in phase 1 clinical trials<sup>1</sup> and that it takes billions of dollars in investment and an average of 15 years to bring a new drug to the market<sup>2</sup>.

Meanwhile, there is an ever-growing effort to apply computational power to improve the effectiveness and efficiency of drug discovery<sup>3</sup>. Traditional computational methods in drug discovery were focused on understanding which proteins could make good drug targets, sequence analysis, modeling drugs binding to proteins, and the analysis of biological data. With the attention on translational research in recent years, a new set of computational methods are being developed which examine drug-target associations and drug off-target effects through system and network approaches. These new approaches take advantage of the unprecedented large-scale high-throughput measurements, such as drug chemical structures and screens<sup>4,5</sup>, side effect profiles<sup>6,7</sup>, transcriptional responses after drug treatment<sup>8,9</sup>, genome wide association studies<sup>10</sup>, and combined knowledge<sup>11,12</sup>. More importantly there are increasing reports of these findings being validated in

---

<sup>†</sup> Work supported by NIH Intramural Research Program, National Library of Medicine

<sup>¶</sup> Work supported by National Institute for General Medical Science (R01 GM079719).

experimental models<sup>6, 8, 13, 14</sup>, thus clarifying the value proposition for computational drug discovery. As a result, now is an exciting time for computational scientists to gain evidence for reusing an existing drug for a different use or generate testable hypotheses for further screening<sup>15</sup>.

Despite the progress, there is clearly room for technical improvement with regard to computational repurposing approaches. Furthermore, to materialize the true potential and impact of these methods, much work is needed to show that they can be successfully adopted into practical applications. Hence, the aim of our session is to provide a forum to bring together the research community for a serious examination of these important issues.

The six papers accepted to this year's session reflect both the value of integrating disparate sources of data and an emerging emphasis in the field on target prediction using improvements on chemical informatics methods

Brubaker *et al.*, using data from the Cancer Genome Project, present a study on the sensitivity of cancer lines to a large group of drugs. Looking at gene expression, copy number data, mutational data, known mechanisms of drugs, and the known targets of drugs, they make mechanistic inferences about the mechanisms of drug resistance and sensitivity. Extracting this type of knowledge from large, complex repositories of screening data will be increasingly important in the coming years. This study also explains how these genomic changes may affect the efficacy of drugs, which connects repurposing with personalized medicine.

Zhu *et al.* present a semantic reasoner that identifies repurposing opportunities for breast cancer. Instead of using machine learning, as do the other papers in the session, their approach looks to connect disparate pieces of information, from several sources, to make a logical case that supports repurposing a effort.

Ng *et al* propose an interesting random-walk based approach to finding repurposing opportunities for malaria. The authors rightly identify specific challenges in applying chemical informatics in infection disease, and there method seems, nonetheless, to make good progress towards overcoming these challenges. Molecules are connected to one another if they are structurally similar and are annotated with their known targets. They show how random walks on this molecule network can identify the targets of molecules known to inhibit Malaria and also suggest potential repurposing opportunities with FDA approved drugs.

Yang *et al*, similarly, propose a promising approach to predicting the protein targets of molecules, a key tool in identifying repurposing opportunities. They use a conditional random field to integrate information from chemical similarity, protein similarity, and known side-effects. This approach predicts the targets of molecules with high accuracy, and is exciting because it integrates critical but disparate data in a unified approach.

Yera *et al.* propose another approach to predicting the targets of molecules. They use a combination of 2D structural similarity, 3D structural similarity, and clinical effect (as

reported in package label) similarity. Their best models get a performance boost from including the clinical effect information from package inserts. They also see strong predictive performance in identifying known off-targets of drugs.

Blucher *et al.* makes the point that there are substantial issues in the metadata, data quality and completeness of public repositories of chemical assay data, like PubChem and ChemBank. Many computational approaches to repositioning seek to identify patterns in publically available chemical assay data, so the issues they identify are critical for the whole field. In particular, we hope their request for improved data submission standards and guidelines will be heeded. Moreover, the next steps forward for the target prediction methods that rely on these datasets may include finding better ways of curating and managing noise in the assay data.

This is the second year Computational Drug Repositioning has been offered as a track at the Pacific Symposium on Biocomputing, and we are pleased with the results of our call for participation. These papers reflect a trend in the field towards target and off-target prediction of molecules. Understanding how drugs work and could work in human disease is, unsurprisingly, the central challenge in computational repurposing. They also reflect a trend towards integrating data from disparate sources, to make connections that would otherwise be hidden.

In the future, we expect the field will continue to develop these themes. There will continue to be cross-pollination with chemical informatics and further progress towards integrating information from disparate datasets. We believe these challenges and opportunities will continue to stimulate innovative work for years to come.

## Acknowledgments

The session co-chairs are grateful to the numerous reviewers for their help in selecting the best papers among many excellent submissions.

## References

1. A. Krantz, *Nat Biotechnol*, **13**, 1294, (1998)
2. C. P. Adams and V. V. Brantner, *Health Aff (Millwood)*, **2**, 420, (2006)
3. M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal. Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.*, **93**, 335–341, (2013)
4. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, **7270**, 175, (2009)
5. S. J. Swamidass, *Brief Bioinform*, **4**, 327, (2011)
6. M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, **5886**, 263, (2008)



7. L. Yang and P. Agarwal, *PLoS One*, **12**, e28025, (2011)
8. M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, *Sci Transl Med*, **96**, 96ra77, (2011)
9. G. Hu and P. Agarwal, *PLoS One*, **8**, e6536, (2009)
10. P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon and V. Mooser, *Nat Biotechnol*, **4**, 317, (2012)
11. J. Li and Z. Lu, *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, (2012)
12. A. Gottlieb, G. Y. Stein, E. Rupp and R. Sharan, *Mol Syst Biol*, 496, (2011)
13. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte, *Sci Transl Med*, **96**, 96ra76, (2011)
14. N. S. Jahchan, J. T. Dudley, P. K. Mazur, N. Flores, D. Yang, A. Palmerton, A. F. Zmoos, D. Vaka, K. Q. Tran, M. Zhou, K. Krasinska, J. W. Riess, J. W. Neal, P. Khatri, K. S. Park, A. J. Butte, and J. Sage. *Cancer Discov*, (2013)
15. P. Sanseau and J. Koehler, *Brief Bioinform*, **4**, 301, (2011)

## CHALLENGES IN SECONDARY ANALYSIS OF HIGH THROUGHPUT SCREENING DATA

AURORA S. BLUCHER, SHANNON K. MCWEENEY

*Division of Bioinformatics and Computational Biology, Oregon Health & Science University  
Portland, OR 97203 USA*

Emails: [blucher@ohsu.edu](mailto:blucher@ohsu.edu), [mcweeney@ohsu.edu](mailto:mcweeney@ohsu.edu)

Repurposing an existing drug for an alternative use is not only a cost effective method of development, but also a faster process due to the drug's previous clinical testing and established pharmacokinetic profiles. A potentially rich resource for computational drug repositioning approaches is publically available high throughput screening data, available in databases such as PubChem Bioassay and ChemBank. We examine statistical and computational considerations for secondary analysis of publicly available high throughput screening (HTS) data with respect to metadata, data quality, and completeness. We discuss developing methods and best practices that can help to ameliorate these issues.

### 1. Introduction

Despite increasing investment in drug research and development in recent years, the pharmaceutical industry has seen limited results in the form of novel marketable drugs.<sup>1</sup> Attention has recently turned to drug repositioning, or finding new uses for already developed drugs. Drug repurposing is particularly attractive due to its simplified timeline; while the traditional drug discovery process can take between ten and seventeen years to bring a drug to production, repurposing a drug can take as little as three to twelve years depending on the drug's previously established chemical properties.<sup>2</sup> In several cases, repurposing has provided enormous benefit to patients with previously limited treatment options, such as the repositioning of thalidomide to treat multiple myeloma, or bromocriptine for Type 2 diabetes. Other well-known repositioning successes include Wellbutrin as Zyban for a smoking cessation aid, Minoxidil for hair loss, and Viagra (sildenafil) for erectile dysfunction.<sup>1-3</sup>

A potentially valuable resource for drug repositioning efforts is publically available high throughput screening (HTS) data.<sup>4</sup> A primary strategy for drug discovery, the automated high throughput screening process allows for the activity of hundreds of thousands of chemical compounds to be tested simultaneously.<sup>5</sup> Compounds are screened against a particular target compound, typically a receptor or enzyme implicated in a disease, and are declared active if their results differ from the majority of the test compounds. However, it is well known that there are several common sources of variation within high throughput screens, both technological, such as batch, plate, and positional (row or column) effects, and biological, such as the presence of non-selective binders, which can result in false positives and negative bioactivity results.<sup>4-8</sup> These problems can be resolved through pre-processing, standardization and normalization methods, which include the z-score, percent inhibition, and median-based methods among others.<sup>5,9,10</sup>

Results from high throughput screening projects, primarily from academic institutions, are often made available through public databases such as NCBI PubChem Bioassay and ChemBank.<sup>4</sup> The PubChem Bioassay database contains the results of high throughput screens for the biological activities of molecules cross-listed in PubChem Substance and Compound.<sup>11,12</sup> Each PubChem assay has a unique assay identifier (AID). Assay data sets usually contain compound information, accompanying readout (for example, recorded fluorescence emission), activity score, activity outcome, and the mean values of minimum and maximum control wells for each plate in the assay. Activity scores and outcome are defined in the assay description, which typically explains the threshold used to declare a particular compound active.<sup>12</sup> The actual raw HTS data is not included in PubChem, however, and therefore there is no information on batch, plate, or within-plate position for each screened compound.

The Broad ChemBank database also contains the results of small molecule screens, as well as the raw datasets from screening centers. Each assay in ChemBank therefore contains not only compound information and accompanying readout, but also batch, plate, row, and column annotation for each screened compound. Additionally, each assay is conducted twice, so assay datasets contain replicate fluorescence readings.<sup>13</sup>

Given the common sources of variation known to affect high throughput screening data, it is crucial that the quality of a particular bioassay is evaluated before its results are used in further research efforts. For instance, researchers interested in using bioactivity information from databases such as PubChem and ChemBank for computational repositioning methods must first be convinced of the reliability of the screens in these databases.<sup>7</sup> Issues in assay quality can result in false positive or false negative bioactivity results, affecting which compounds are considered for potential repositioning. Here, datasets from both PubChem and ChemBank are evaluated to quantify the advantages and limitations of each repository as well as to investigate common sources of variation such as batch, plate, and positional effects. This analysis is representative of a typical investigation of HTS data that would be conducted before utilizing this data in further computational repurposing efforts. Overall, the problems encountered here illustrate some of the key barriers to effective secondary use of publically available high throughput screening data in order to realize the full potential of these datasets.

## 2. Methods

In this study, exploratory analysis was conducted on representative bioassay datasets from PubChem and ChemBank to examine data completeness, particularly in the context of data pre-processing and addressing technical sources of variation. Additional data was obtained directly from the original screeners of the highlighted PubChem study to complete the exploratory data analysis and allow for comparable assessments to the ChemBank study.

## 2.1 PubChem Example

The PubChem CDC25B (AID 368) dataset contains the results from approximately 65,222 compounds and controls of a primary screen against the target CDC25B. CDC25 is a protein tyrosine phosphatase cell cycle regulator, and of three existing isoforms, two are oncogenic and have been found to be overexpressed in a variety of human tumors. The goal of this screen was to find potential inhibitors for the CDC25B isoform.<sup>14</sup> The CDC25B dataset contained the following attributes: PubChem Substance ID, PubChem Compound ID, activity score, activity outcome, database URL, comment field, raw fluorescence intensity, calculated percent inhibition, mean of minimum control well signals (by plate), mean of maximum control well signals (by plate), calculated z-factor, and assay run date. Exploratory data analysis was conducted to evaluate the overall distribution of fluorescence intensity, percent inhibition, minimum control well means, maximum control well means, and calculated z'-factors. However, no further analysis could be performed for this dataset in the form available from the PubChem database, given the lack of plate level data such as batch number, plate number, and row and column information for each well.

## 2.2 Full PubChem Example

The full CDC25B dataset, including plate-level annotation, was obtained directly from the PMLSC screening center and contained results from approximately 83,711 compounds and controls across 218 384-well microtiter plates. In addition to PubChem Compound ID, raw fluorescence emission, calculated percent inhibition, mean minimum signal, mean maximum signal, calculated z-factor, and run date, this dataset also included assay batch, plate ID, row, column, well number, and well annotation. This information enabled further exploratory data analysis such as evaluation of fluorescence intensity distribution by well type and across plates and batches. Heatmaps were created for individual plates to check for positional effects. The mean signal to background ratio and percent coefficients of variation for the minimum and maximum control wells were also calculated. Based on the exploratory data analysis, percent inhibition was chosen as the most appropriate normalization method, which was also the method chosen by the original screeners when processing the dataset.<sup>5,14</sup>

## 2.3 ChemBank Example

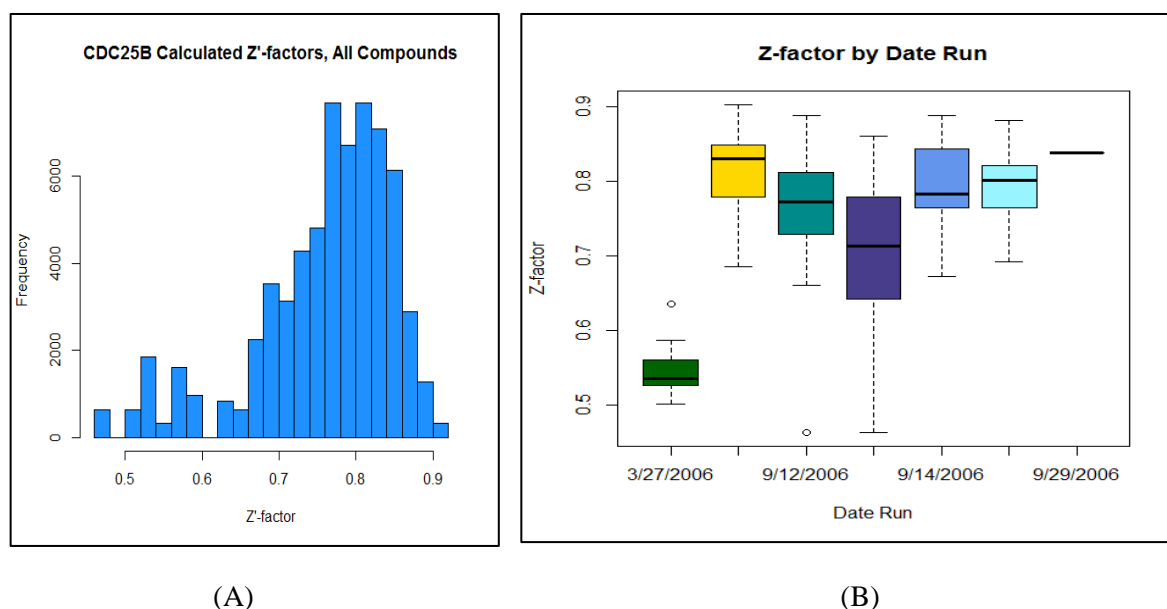
The ChemBank BRAF dataset contains the results from approximately 41,088 compounds and controls of a primary screen to find an inhibitor of the BRAF<sup>V600E</sup> mutant. The BRAF gene plays an important role in the mitogen-activated signaling pathway and in particular, the BRAF<sup>V600E</sup> mutation has been implicated in melanoma, papillary thyroid carcinoma, and colorectal cancer.<sup>15</sup> The BRAF dataset is composed of seven different assays, each with two replicates. Given limited assay description and annotation provided, each of the seven assays was evaluated separately. First, correlation of raw fluorescence intensity between the two replicates was assessed for each of the seven assays, and if present, any outlying data points were investigated at the plate level. Next,

exploratory data analysis was conducted for each assay to assess the overall distribution of fluorescence intensity, background-subtracted values, and calculated z-score. This analysis included histograms, boxplots, and quantile-quantile plots for individual replicates and statistical indices of the combined data, as appropriate.

### 3. Results

#### 3.1 PubChem Example

Overall, the distribution of fluorescence intensity across all compounds in the CDC25B dataset is strongly skewed right, while the distribution of percent inhibition across all compounds is strongly skewed to the left. The distribution for the range between the mean minimum and mean maximum control wells is slightly skewed bimodal (See Supplementary Material S1) The distribution of z'-factors across all compounds is fairly skewed to the left and appears to be slightly bimodal. Boxplots of z'-factor by run date reveal strong variation by date (Figure 1).



**Figure 1. Distribution of Z'-factors for PubChem CDC25B dataset.** (A) Histogram depicting distribution of calculated z'-factors. (B) Boxplots by run date for calculated z'-factors.

It is noted that the compounds run in March 2006 have much lower z'-factors than the remaining compounds, run in August and September 2006. Additionally, the compounds run on September 13th, 2006 exhibit a much wider range of z'-factors than compounds run on any other dates, while compounds run on September 29<sup>th</sup>, 2006 exhibit a much narrower range. Given that the z'-factor is a commonly used measure of assay quality, plates with a such divergent z'-factors should be examined for possible errors and batch effects. Here, however, further investigation into the sources of this variation could not be conducted due to the lack of plate level annotation available through the

PubChem Bioassay database. If the metadata had been available, it would then be possible to attempt to correct for batch and technical sources of variation.

#### Full PubChem CDC25B example

Histograms of fluorescence intensity by well type (compound, 50% inhibition, minimum, and maximum) for the full CDC25B dataset show that the distribution of fluorescence intensity across all wells is somewhat normal with a strong peak. The distributions of fluorescence intensities for compound wells and maximum control wells are slightly skewed right, while the distributions of fluorescence intensities for minimum and 50% inhibition control wells are more strongly skewed to the right (See Supplementary Material S2 Fig 1 and 2). Fluorescence intensity appears to vary widely by both batch and run date as well as by plate within respective batches (See Supplementary Material S2 Fig 3-8). No apparent positional effects were detected by visual examination of heatmaps for each of the 218 plates in the dataset.

Following a recently proposed decision process for HTS data processing, percent inhibition was chosen as the most appropriate method of normalization, due to the fairly normal distribution of fluorescence intensity, lack of row and column biases, a mean signal to background ratio greater than 3.5, and percent coefficients of variation for both the minimum and maximum controls wells less than 20%<sup>5</sup> (See Supplementary Material S2 Table 1). This appeared to successfully normalize the data by batch, date, and across plates within each batch and reproduced the original analysis (See Supplementary Material S2 Fig 9-16). It is important to note that it would not be possible to successfully evaluate this data set with regard to pre-processing and normalization without the plate level annotation.

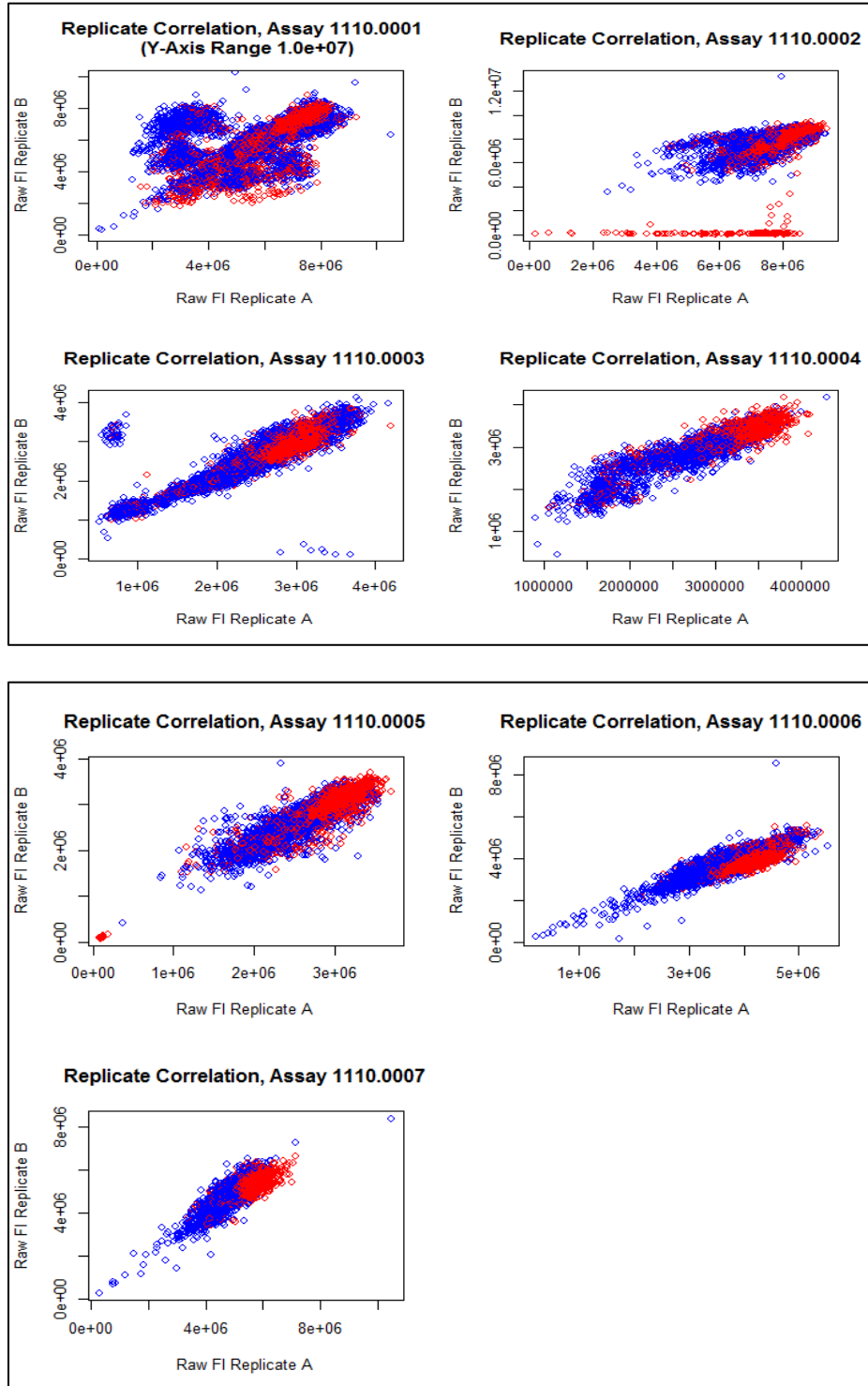
### **3.2 ChemBank Example**

There was a large range with regard to correlation of fluorescence intensity between replicates: 0.436-0.910 (Table 1). Scatterplots further illustrate the high variability among some replicates (Figure 2). This allows easy identification of signal discrepancies. For example, the bottom of the scatterplot for assay 1110.0002, it is easy to detect a set of mock treatment wells (in red) where signal was present in replicate A, but not in replicate B. Similarly, the upper left-hand corner of the scatterplot for assay 1110.0003 shows a replicate specific cluster of compound treatment wells. The outlying data points in assay 1110.0002 were found to be confined to one plate, 1110.0002.Base. The outlying data points in assay 1110.0003 were similarly located on a single plate, 1110.0003.2340.

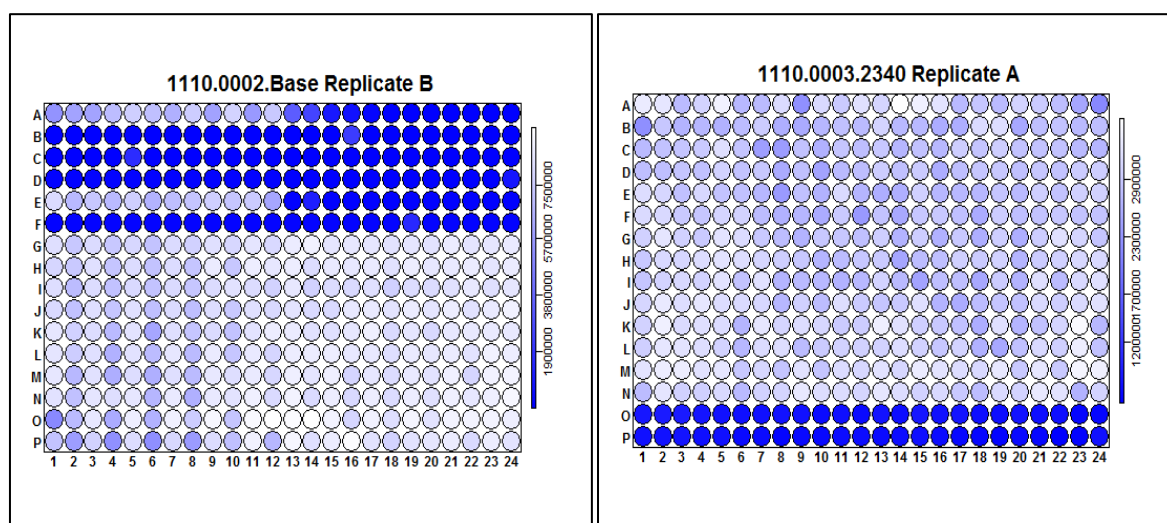
**Table 1. Correlation Coefficients for Fluorescence Intensity Replicate A vs Fluorescence Intensity Replicate B, by Assay, ChemBank BRAF dataset.**

Assay Number	1110.0001	1110.0002	1110.0003	1110.0004	1110.0005	1110.0006	1110.0007
Correlation	0.436	0.536	0.906	0.910	0.902	0.869	0.846

Examination of the well-plate layout for 1110.0002 allowed identification of an obvious positional effect in the upper six rows of the plate (Figure 3). Similarly for 1110.0003, the corresponding well-plate layout illustrated a clear positional effect along the bottom two rows of the plate.



**Figure 2. Scatterplots for Correlation of Fluorescence Intensity Between Replicates A and B.** Correlation between replicates of Assay 1110.0001- 1110.0007. Blue indicates compound-treatment wells, red indicates control wells.

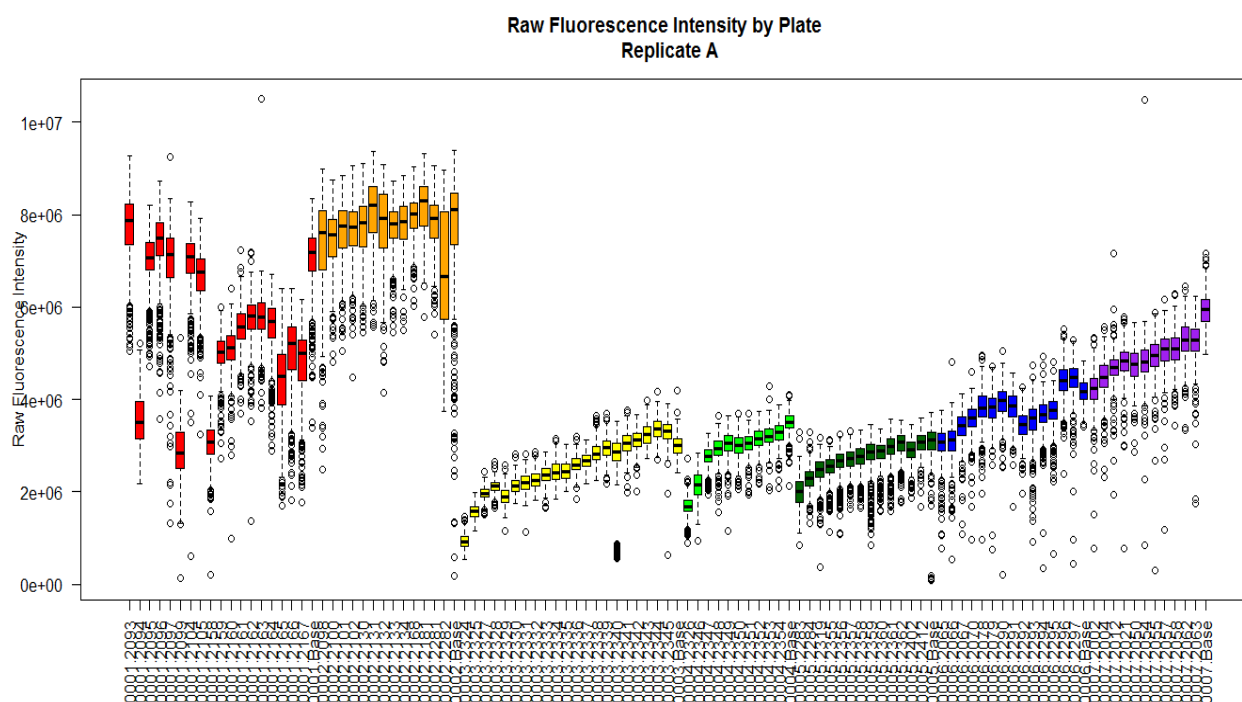


**Figure 3. Well Plate Layouts for Selected BRAF Assays.** (Left) Replicate B of Base Plate for Assay 1110.0003. (Right) Replicate A of Plate 2340 for Assay 1110.0003. Darker wells indicate decreased fluorescence.

Overall, each of the seven assays in the BRAF dataset showed fairly different distributions for fluorescence intensity, background-subtracted values, and calculated z-scores (See Supplementary Material S3), further reiterating the role of exploratory data analysis to examine model assumptions prior to downstream analysis.

Boxplots of the fluorescence intensity by plate were then examined. It was noted that the signal varies considerably across plates, both within and across each of the seven assays. (Replicate A shown in Figure 4). Beginning with assay 1110.0003 in replicate A, it is apparent that within each assay, fluorescence intensities steadily increase with each successive plate that is run before dropping down at the beginning of the next assay. In the absence of timestamps for each plate, it was assumed that increasing plate numbers indicate passage of time. However, without that appropriate metadata, it is not possible to determine the actual source of variation, again limiting the ability to correctly model batch or temporal effects.





**Figure 4. Raw Fluorescence Intensity by Plate, Across All Assays, Replicate A, ChemBank BRAF dataset.**

Each boxplot depicts the fluorescence values of the wells of one plate. Colors indicate assay “Name”, which may or may not be synonymous with batch.

#### 4. Discussion

Both repositories examined provide excellent opportunities for secondary analysis of public HTS data. However, we have noted several issues that need to be addressed in order to realize their full potential. Most notably, the lack of actual raw data, and therefore plate level annotation for bioassays in PubChem BioAssay prevents rigorous analysis of data quality. As illustrated above, initial exploratory analysis of the limited CDC25B dataset (as obtained from PubChem) reveals potential quality issues, such as variation by run date. These issues cannot be fully investigated, however, without knowledge of batch and plate numbers and row and column positioning for each tested compound. The complete CDC25B dataset, obtained directly from the screeners, allowed for more in-depth investigation of sources of variation, which in turn allowed for more appropriate pre-processing and normalization recommendations to be made. It would not have been possible to evaluate the dataset solely from the data and annotation made available through the PubChem database.

Another issue for researchers seeking to extract assay information from PubChem is the lack of description for the particular readouts used in assays. While the PubChem assay discussed in this paper provided a full description of the fluorescence emission readout, many assays do not necessarily include this level of information. It is also important to note that the issues discussed here are likely

extensible to other databases, such as ChEMBL, which contain bioactivity information from selected PubChem Bioassays.<sup>16</sup>

The ChemBank database is currently the only publically available bioassay database that requires the inclusion of plate level annotation in their datasets. While this information is crucial for secondary analysis, the value of the datasets in ChemBank is negatively impacted by the lack of assay annotation and description. For instance, the BRAF dataset was composed of seven different assays, but it was unclear how these differed from one another, if at all. From the assay descriptions, it appeared that only the first assay differs in its biological components, but there was no additional information as to why the remaining six assays were conducted separately. Additionally, while we might expect strong correlation between replicates for each assay, several assays exhibited exceptionally poor correlation, which casts doubt on the overall quality of the screening data. Furthermore, the lack of date or timestamps for the ChemBank data makes it impossible to confirm temporal batch effects, limiting one to data visualization by plate, with an assumption that plate order corresponds with time, as done in Figure 4.

Correspondence with PubChem confirmed that PubChem Bioassay does not require plate level annotation in uploaded datasets to the BioAssay database. It is also noted that there is no way to query for which, if any, datasets include this level of annotation (Personal communication with PubChem). ChemBank also confirmed that the “AssayName” field is used by depositors in different ways: it can be used for biologically different assays or batches of similar assays. Currently, there is no method of querying for datasets to identify those for which particular descriptive information/metadata are included (Personal Communication with ChemBank). These issues affect not only the general usability of the databases, but in particular hinder a larger-scale systematic quality analysis of HTS assays. The analysis presented here was restricted to one assay from each database primarily due to difficulties in accessibility and poor annotation.

Issues such as these in turn stymie the usage of high throughput screening data in further research efforts such as computational repositioning efforts requiring bioactivity information. There is the potential for improved data standards and development of best practices for data dissemination to improve the quality and reusability of the data in these repositories. At a minimum, the inclusion of metadata such as plate and well-level annotation will enable a more thorough secondary analysis of HTS data. Additional oversight to ensure descriptor fields for assays are completed may also encourage assay re-use. With respect to cost-benefit analysis, the potential for re-use of the data via secondary analysis far outweighs any costs due to additional data standards or metadata requirements, as the metadata has already been generated. Further impact in time/resources for depositing additional metadata can easily be mitigated by automation. One example of methods to facilitate the reporting of this metadata is a recently proposed method to first extract workflows directly from screening data in PubChem and then use the workflows to organize data within screening projects.<sup>17</sup>

Addressing these issues in the research community and in the requirements for submission to these repositories could improve the re-use of these data sets. A PubMed search for “PubChem” results in only 263 articles, and the more specific “PubChem BioAssay” pulls up only 51 articles. Querying for “ChemBank” returns even fewer articles, with only 17 results. For perspective, searching “GEO” brings up approximately 8480 results for Gene Expression Omnibus. While both PubChem BioAssay and ChemBank are fairly young databases and more expansive mining efforts using their datasets may still be yet to come, the annotation and data quality issues in both databases cannot be ignored as a potential barrier to dissemination. Expanded datasets as well as more rigorous quality standards are necessary to ensure the public data is truly accessible and re-usable.

## 5. Acknowledgements

Funding for this project was provided by the following grants: NLM (2T15LM007088-21); NIH/NCI (5P30CA069533-13, 4R00CA151457-03); NIH/NCATS (5UL1RR024140). Supplementary data is available at <http://www.biodevlab.org/HTS>

## References

1. Dudley, J. T., Deshpande, T. & Butte, A. J. Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.* **12**, 303–311 (2011).
2. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
3. Pijl, H. *et al.* Bromocriptine: a novel approach to the treatment of type 2 diabetes. *Diabetes Care* **23**, 1154–1161 (2000).
4. Swamidass, S. J. Mining small-molecule screens to repurpose drugs. *Brief. Bioinform.* **12**, 327–335 (2011).
5. Shun, T. Y., Lazo, J. S., Sharlow, E. R. & Johnston, P. A. Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *J. Biomol. Screen.* **16**, 1–14 (2010).
6. Mayr, L. M. & Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **9**, 580–588 (2009).
7. Xie, X.-Q. S. Exploiting PubChem for virtual screening. *Expert Opin. Drug Discov.* **5**, 1205–1220 (2010).
8. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**, 188–195 (2011).
9. Brideau, C. Improved Statistical Methods for Hit Selection in High-Throughput Screening. *J. Biomol. Screen.* **8**, 634–647 (2003).

10. Gribbon, P. Evaluating Real-Life High-Throughput Screening Data. *J. Biomol. Screen.* **10**, 99–107 (2005).
11. Li, Q., Cheng, T., Wang, Y. & Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discov. Today* **15**, 1052–1057 (2010).
12. Wang, Y. *et al.* An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **38**, D255–D266 (2009).
13. Seiler, K. P. *et al.* ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **36**, D351–D359 (2007).
14. Johnston, P. A. *et al.* Cdc25B Dual-Specificity Phosphatase Inhibitors Identified in a High-Throughput Screen of the NIH Compound Library. *ASSAY Drug Dev. Technol.* **7**, 250–265 (2009).
15. Coffee, E. M. *et al.* Concomitant BRAF and PI3K/mTOR Blockade Is Required for Effective Treatment of BRAFV600E Colorectal Cancer. *Clin. Cancer Res.* **19**, 2688–2698 (2013).
16. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2011).
17. Calhoun, B. T., Browning, M. R., Chen, B. R., Bittker, J. A. & Swamidass, S. J. Automatically Detecting Workflows in PubChem. *J. Biomol. Screen.* **17**, 1071–1079 (2012).

## DRUG INTERVENTION RESPONSE PREDICTIONS WITH PARADIGM (DIRPP) IDENTIFIES DRUG RESISTANT CANCER CELL LINES AND PATHWAY MECHANISMS OF RESISTANCE

DOUGLAS BRUBAKER<sup>1,2\*</sup>, ANALISA DIFEO<sup>2</sup>, YANWEN CHEN<sup>2</sup>, TAYLOR PEARL<sup>4</sup>, KAIDE ZHAI<sup>2</sup>, GURKAN BEBEK<sup>1,2,3</sup>, MARK CHANCE<sup>1,2</sup>, JILL BARNHOLTZ-SLOAN<sup>1,2</sup>

<sup>1</sup> Case Center for Proteomics and Bioinformatics, Case Western Reserve University School of Medicine, BRB 932, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA, Email: dkb50@case.edu, \*Corresponding Author; <sup>2</sup>Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, 11100 Euclid Avenue, Cleveland, Ohio 44106, USA; <sup>3</sup>Genomic Medicine Institute, Cleveland Clinic Lerner Research Institute, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA; <sup>4</sup>Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

The revolution in sequencing techniques in the past decade has provided an extensive picture of the molecular mechanisms behind complex diseases such as cancer. The Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Project (CGP) have provided an unprecedented opportunity to examine copy number, gene expression, and mutational information for over 1000 cell lines of multiple tumor types alongside IC50 values for over 150 different drugs and drug related compounds. We present a novel pipeline called DIRPP, Drug Intervention Response Predictions with PARADIGM<sup>7</sup>, which predicts a cell line's response to a drug intervention from molecular data. PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models) is a probabilistic graphical model used to infer patient specific genetic activity by integrating copy number and gene expression data into a factor graph model of a cellular network. We evaluated the performance of DIRPP on endometrial, ovarian, and breast cancer related cell lines from the CCLE and CGP for nine drugs. The pipeline is sensitive enough to predict the response of a cell line with accuracy and precision as high as 80 and 88% respectively. We then classify drugs by the specific pathway mechanisms governing drug response. This classification allows us to compare drugs by cellular response mechanisms rather than simply by their specific gene targets. This pipeline represents a novel approach for predicting clinical drug response and generating novel candidates for drug repurposing and repositioning.

### 1. Introduction

The potential for bioinformatics techniques to bring about transformative results in personalized medicine is just beginning to be realized. Large scale studies such as The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Genome Project (CGP) have provided bioinformaticians with a wealth of –omic and pharmacologic data to interrogate<sup>1-5</sup>. Novel algorithms have been developed to perform detailed signaling pathway analysis<sup>6</sup>, integrate diverse –omic data types<sup>7-11</sup>, and even predict markers of drug sensitivity and resistance<sup>12</sup>. Analytical efforts are also underway to identify candidates for drug repurposing or repositioning and to computationally predict new drug indications for disease<sup>13</sup>.

Despite this wealth of innovation, the complexity for interpretation and translation of results to cancer patients remains challenging. The diversity of computational approaches has made it difficult to identify which of these have the most potential to improve the treatment of patients and improve clinical outcomes<sup>14</sup>. Each algorithm relies on a different type of –omic or combination of –omic data making it difficult to integrate them in a single analytical pipeline<sup>12, 13</sup>.

An important goal of computational bioinformatics pipelines is to provide actionable results to help physicians make optimal therapeutic decisions for a patient. To this end, the patient's likelihood to respond to a specific treatment regimen is of particular interest to clinicians. The typical clinical case includes investigators looking to discover alternative therapies for patients who demonstrate resistance to the primary treatment. Both drug repurposing, the recycling of shelved or failed drugs, and drug repositioning, the use of active therapies for new applications, represent opportunities for the development of second line therapies. In order to maximize the impact of such an analysis pipeline, it should be versatile enough to address a myriad of clinical and scientific questions and easily integrate with existing clinical pipelines to assist physicians.

To address these clinical and analytical challenges we propose an integrative pipeline called DIRPP, Drug Intervention Response Predictions with PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models)<sup>7</sup>. Our pipeline aims to classify a cell line as either sensitive or resistant to a given therapy and to define specific genetic backgrounds represented in the cell line, potentially applicable to specific patients, associated with drug response phenotypes. This classification is performed using an extension of an open source probabilistic graphical model called PARADIGM. Drawing on multiple data types, DIRPP proceeds to integrate the copy number and gene expression data for a cell line into a biological pathway activity score which includes the result of a simulated drug intervention. Once the cell line (which may be a surrogate for a patient of interest) has been classified as sensitive or resistant to a given therapy, downstream gene set enrichment analysis (GSEA) on the pathway activity scores illustrates the underlying biological pathway mechanisms at work driving the drug response phenotype. The method can be applied to assess the impact of a wide variety of therapies on one particular cancer, or multiple cancers at a time to develop precision medicine strategies.

## 2. Materials and Methods

### 2.1. Datasets, Pathway Sources, and Pharmacologic Profile Data

Copy number, gene expression, and drug sensitivity data for 202 cancer cell lines from two recently published preclinical studies, the cancer genome project (CGP)<sup>4</sup> and the cancer cell line encyclopedia (CCLE)<sup>5</sup> were used for analysis. The distribution of cell lines by cancer type was: 20 ovarian, 39 breast, and 6 endometrial cancer cell lines from the CGP for testing of the algorithm and 51 ovarian, 59 breast, and 27 endometrial cancer cell lines from the CCLE for an independent dataset to validate the algorithm. Of the 16 drugs in common between the two studies, 9 inhibitory drugs were selected for analysis based on their clinical potential for treatment of ovarian cancer and repurposing/repositioning in breast and endometrial cancers (Table 1). Genetically similar

sub-types of these cancers represented in this array of cell lines have been the subject of numerous genomic and drug repositioning studies and provide a robust sample set for analysis.

Table 1. Nine (9) anticancer inhibitory drugs analyzed in both the CGP and CCLE with primary clinical relevance to ovarian cancer and secondary clinical relevance to breast, and endometrial cancer.

Drug Name	Target(s)	Class
Erlotinib	EGFR	Kinase Inhibitor
Irinotecan	Topoisomerase	Cytotoxic
AZD0530	Src, ABL/BCR-ABL, EGFR	Kinase Inhibitor
AZD6244	MEK, ERK, MAPK	Kinase Inhibitor
PD0325901	MEK, RAF, MAPK	Kinase Inhibitor
Lapatinib	EGFR, HER2	Kinase Inhibitor
17-AAG	HSP90	Other
Sorafenib	KIT, PDGFRB, FLT3, FLT4, KDR, RAF1, BRAF	Kinase Inhibitor
Paclitaxel	Microtubules	Cytotoxic

All cell line drug sensitivity values were reported as  $IC_{50}$  values, the concentration at which a drug inhibits 50% of cellular growth<sup>4,5</sup>. Gene expression probes were normalized by centering on the gene's median expression across all cell lines and then taking the base 2 log of that value<sup>7</sup>. SuperPathway, a merged biological pathway of 1,441 curated signal transduction, transcriptional, and metabolic pathways, was used to analyze the comprehensive cellular network of activity in the cell lines. This framework captures the global interactions of any perturbation in a cell while removing redundant pathway elements<sup>15</sup>. For each drug of interest, detailed pharmacological information about gene targets and mechanism of action was obtained from the drugbank and selleckchem databases<sup>16-18</sup>.

## 2.2. Analysis Pipeline

The DIRPP<sup>7</sup> pipeline was implemented and tested using the overall scheme and specific steps laid out in Figure 1. Two runs of the PARADIGM algorithm are completed, one with –omic data, the two factor analysis, the other with –omic data and a simulated drug intervention, a three factor analysis. PARADIGM represents each entity in a biological pathway as a node whose value depends upon a defined internal set of “evidence nodes” whose connectivity mirrors the central dogma of molecular biology (Figure 2). These “evidence nodes” enable the integration of patient data into the biological pathway network. After assessing the signaling pathway activity of the cell lines with an initial run of the PARADIGM algorithm, where a DNA node interacts with a mRNA node to propagate biological information to the cellular network<sup>7</sup>, a second run of PARADIGM is performed while including a drug induced re-wiring of the cellular network (Figure 2). The resulting IPLs were then compared on a per-patient-per-gene basis to assess the impact of the drug intervention on perturbing the signaling network of a cancer cell line by computing a paired t-test p-value using the IPLs of the two PARADIGM runs for each cell line. The least perturbed cell lines were deemed the most resistant (least sensitive). All cell lines were then ranked in order of increasing sensitivity. Biological pathways involved in drug sensitivity and resistance were then identified using Gene Set Enrichment Analysis (GSEA)<sup>6</sup>.

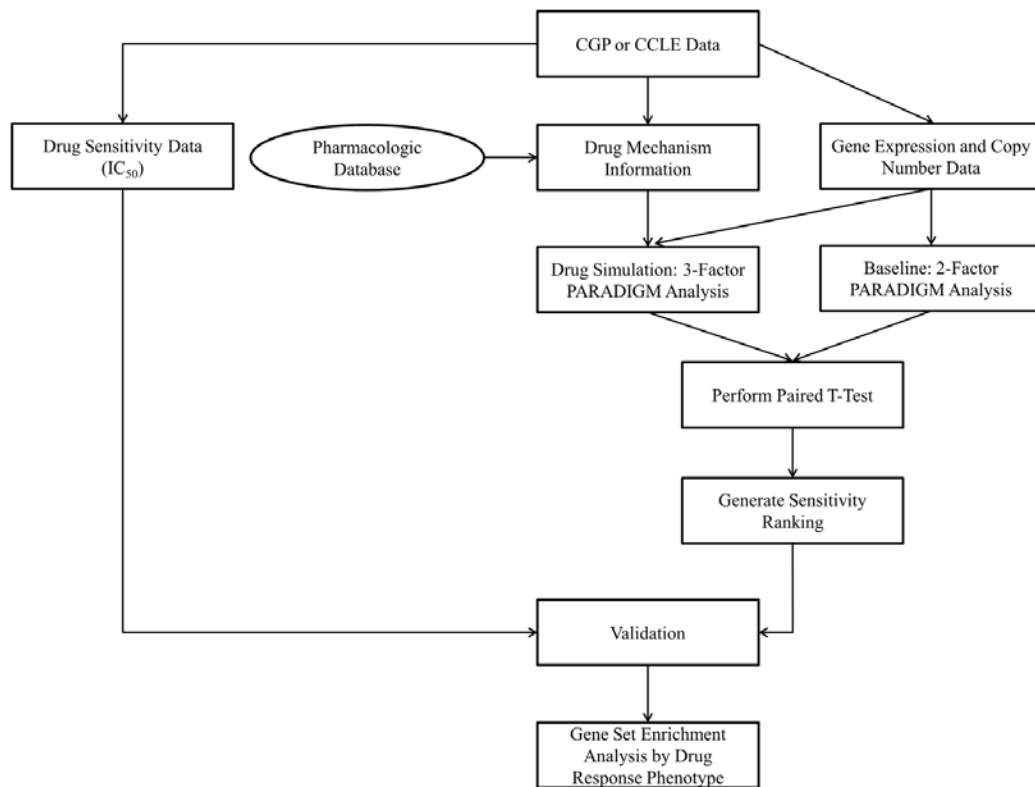


Figure 1. Experimental design of DIRPP. For each cell line dataset, gene expression and copy number data were analyzed in 2-factor PARADIGM analysis. These *inferred pathway levels* IPL's were compared to those from 3-factor PARADIGM analysis with a simulated drug intervention to generate a ranking by drug sensitivity. This ranking was then validated on the CGP and CCLE data. Response mechanisms were classified with GSEA.

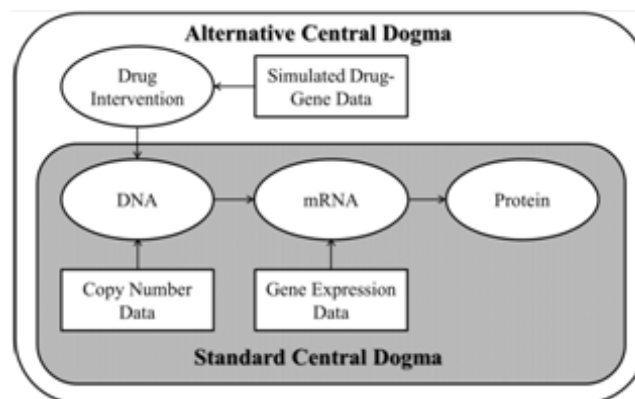


Figure 2. Comparison of the PARADIGM standard central dogma with an alternative dogma which represents a drug induced re-wiring of the network. The drug intervention propagates through the network based on an inferred interaction at the DNA node as a surrogate for its actual influence on protein activity.



### 2.3. PARADIGM Model

Briefly, PARADIGM is a factor-graph-based approach which quantifies the activity of a gene given a pathway diagram and dataset of observations<sup>8</sup>. For the model proposed here, SuperPathway was used to define this pathway diagram where each gene, protein, or process is connected by a series of *factors* which encode the probabilistic constraints between variables<sup>7, 15</sup>. Each entity in the pathway infers its activity from a set of nodes which define an internal set of rules for how these data types interact to assign a value to the pathway entity. Nodes for DNA and mRNA connect to the active protein node which then passes information through the entire pathway diagram via the dependencies encoded in the factors. The DNA and RNA nodes of each gene in the pathway are assigned values as a function of the copy number and gene expression data respectively to include biological information from the cell lines. For each gene, PARADIGM is capable of integrating these diverse –omic data types to compute an *inferred pathway level* (IPL) for each gene in the pathway. These IPL scores were computed using a belief-propagation algorithm on the factor graph diagram of the pathway. Each score represents a log-posterior odds (LPO) ratio of the state of a pathway entity given the observed data. Positive IPLs correspond to an entity being active in a tumor relative to normal tissue and negative to inactivity<sup>7, 8</sup>.

### 2.4. Drug Intervention Simulation

DIRPP exploits a versatile feature of PARADIGM which allows the user to define a drug induced re-wiring for a gene in a pathway. As designed, PARADIGM is capable of integrating DNA methylation data by including an extra node in a gene's normal wiring connected at the DNA node<sup>7</sup>. The current algorithm utilized the DNA methylation feature to encode the action of a drug on that particular gene's regulatory structure (Figure 2). A drug's mechanism of action was retrieved from drugbank and selleckchem databases, which provide a list of genes (proteins) the drug targets<sup>16-18</sup>. A matrix of genes that correspond to a drug intervention was then defined. The edge connecting the intervention node to the DNA node encoded a factor which signaled a downregulation to the gene (similar to the standard use of methylation). Only genes listed in this intervention matrix had the extra node added to their wiring diagrams. In principle, the edge connecting the intervention node to the DNA node could be changed to act in an amplifying manner for an agonist.

To assess the significance of a drug intervention, two runs of the PARADIGM algorithm were completed: one with copy number and gene expression data, the other with the addition of a third data type, the simulated drug intervention, with each run generating a matrix of IPL scores. The two resulting matrices of IPL scores were then compared on a per-cell line-per-gene basis using a paired t-test to calculate a p-value for that cell line. The cell lines were ranked in order of largest to smallest p-value corresponding to a ranking of least to most sensitive cell lines for a given drug.

## 2.5. Validation

To validate our approach, analysis of the CGP and CCLE data were independently performed by calculating the accuracy and precision statistics for each ranking. Accuracy assesses the algorithm's overall performance for distinguishing between sensitive and resistant cancer cell lines while precision is used to assess the positive predictive value of the algorithm at identifying drug resistant cancer cell lines.

$$\text{accuracy} = \frac{\text{number of (true positives+true negatives)}}{\text{number of (true positives+true negatives+false positives+false negatives)}} \quad (1)$$

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives+false positives}} \quad (2)$$

A ranking of cell lines by p-value was first constructed using the results of the t-test. This ranking by p-value was compared to the actual ranking by IC<sub>50</sub> value measuring drug sensitivity. The accuracy and precision statistics were calculated by defining a cutoff in the ranking where the cell lines change from primarily resistant (IC<sub>50</sub>>1) to primarily sensitive (IC<sub>50</sub><0.1), where intermediately sensitive lines (0.1<IC<sub>50</sub><1) were treated as resistant. There were generally more drug-resistant cancer cell lines than sensitive ones and for some drugs; no sensitive cell lines were available for comparison. For validation of these difficult drugs we defined our cutoff for drug resistance detection at an IC<sub>50</sub> value of 8μM, where we considered values greater than 8μM to correspond to highly drug resistant cell lines and everything below to moderately drug resistant cell lines. The CGP did not have any ovarian, breast, or endometrial cancer cell lines with IC<sub>50</sub> less than 8μM for Erlotinib. We then calculated DIRPP's accuracy (1) and precision (2) for each dataset for each of the three cancers studied individually and together as a whole. Previous studies have indicated 78% accuracy as being a very high level, others have used a concordance index and set the cutoff at 0.6 to measure correlations<sup>12,13</sup>. We chose to use accuracy and precision cutoffs at 0.67 to define an "acceptable" level of validation between these two cutoffs.

## 3. Results

### 3.1. Drug Simulations

We simulated drug interventions for each of the drugs in Table 1 by defining mechanism-specific drug intervention files. The drug's mechanism of action, the genes it targets, was propagated through the cancer cell line's network via a drug intervention node coded in the PARADIGM algorithm's rewiring for each effected gene. Four interventions were simulated for each drug in each dataset, one which included all breast, ovarian, and endometrial cancer cell lines as one cohort, and three other simulations for each cancer-type individually.

The ranking of cell lines by p-value was compared to the ranking of cell lines by IC<sub>50</sub> for each drug and the accuracy and precision of that ranking was assessed using the cutoffs for resistance and sensitivity either by IC<sub>50</sub> value, or by the highly-moderately resistant cutoff previously described. Certain ovarian cell lines have been shown to be hypermutated or were potentially mislabeled as they are more similar to other tumor types<sup>19</sup>. These cell lines were excluded to ensure the consistency of this analysis for only breast, endometrial, and ovarian

cancer. Except for AZD0530, the overall response of all drugs across both datasets was predicted within 67% average accuracy or greater, with most being predicted with over 75% accuracy (Table 2). DIRPP predicted the resistance of cell lines with precision of 0.67 or greater for all drugs except for Paclitaxel. Some drugs such as Irinotecan performed distinctly different between datasets (Table 2). DIRPP was able to detect resistant cell lines with an overall precision of 0.81 across all datasets (Table 3). Across all cancers studied combined DIRPP performed with a precision of 0.78 and accuracy of 0.73. Ovarian cancer drug response was predicted better than the other cancers with an overall precision of 0.81 and accuracy of 0.79 (Table 3).

Table 2. Precision and accuracy statistics for each drug across all cancer types combined by dataset and overall.

	CGP Data		CCLE Data		Overall	
	Resistance Precision	Accuracy	Resistance Precision	Accuracy	Average Precision	Average Accuracy
17AAG	0.88	0.77	0.72	0.59	0.80	0.68
AZD0530	0.73	0.57	0.62	0.58	0.67	0.58
AZD6244	0.84	0.76	0.90	0.81	0.87	0.79
Erlotinib	-	-	0.88	0.80	0.88	0.80
Irinotecan	0.44	0.63	0.91	0.88	0.68	0.75
Lapatinib	0.90	0.75	0.72	0.58	0.81	0.67
Paclitaxel	0.80	0.83	0.43	0.60	0.61	0.72
PD0325901	0.79	0.69	0.93	0.86	0.86	0.78
Sorafenib	1.0	1.0	0.71	0.59	0.86	0.80

Table 3. Precision and accuracy statistics by dataset for all cancer types combined and by cancer type individually

	CGP Data		CCLE Data		Overall	
	Resistance Precision	Accuracy	Resistance Precision	Accuracy	Precision	Accuracy
All Cancers	0.81	0.78	0.76	0.70	0.78	0.73
Breast	0.83	0.80	0.73	0.67	0.78	0.73
Ovarian	0.75	0.81	0.84	0.78	0.81	0.79
Endometrial	0.83	0.83	0.74	0.65	0.76	0.70

### 3.2. Mechanisms of Drug Resistance

Once the cell lines were classified as either sensitive or resistant to a drug we performed gene set enrichment analysis (GSEA) by drug response phenotype to uncover the biological pathway mechanisms driving drug resistance. For this analysis we required cell lines with IC50 values greater than 1 or less than 0.1 to define, resistant and sensitive, respectively. Only 17AAG, Irinotecan, Paclitaxel, and PD0325901 had sufficiently diverse drug sensitivity profiles to classify cell lines using the above definition in order to perform GSEA. Each of these drugs has a distinct mechanism of action and no overlapping molecular targets. Despite this, we were able to identify several signaling pathway mechanisms that these cell lines shared related to drug resistance.

We ran GSEA using the IPL values generated by PARADIGM using the simulation that combined copy number and gene expression data. Permutation analysis of the phenotypes (sensitive or resistant) was used to judge significance. Pathways which had nominal p-values less than 0.05 were selected for further comparison across drugs. There was a common activation of PDGF signaling associated with resistance to PD0325901, Paclitaxel, and Irinotecan in the resistant endometrial, breast, and ovarian cancer cell lines. This confirms previous work which associates PDGF upregulation with Paclitaxel resistance in breast and ovarian cancer<sup>20, 21</sup> and suggests that the genetically similar endometrial cancer<sup>1</sup> may also share this mechanism of drug resistance. Irinotecan and Paclitaxel shared 9 mechanisms of resistance with each cancer.

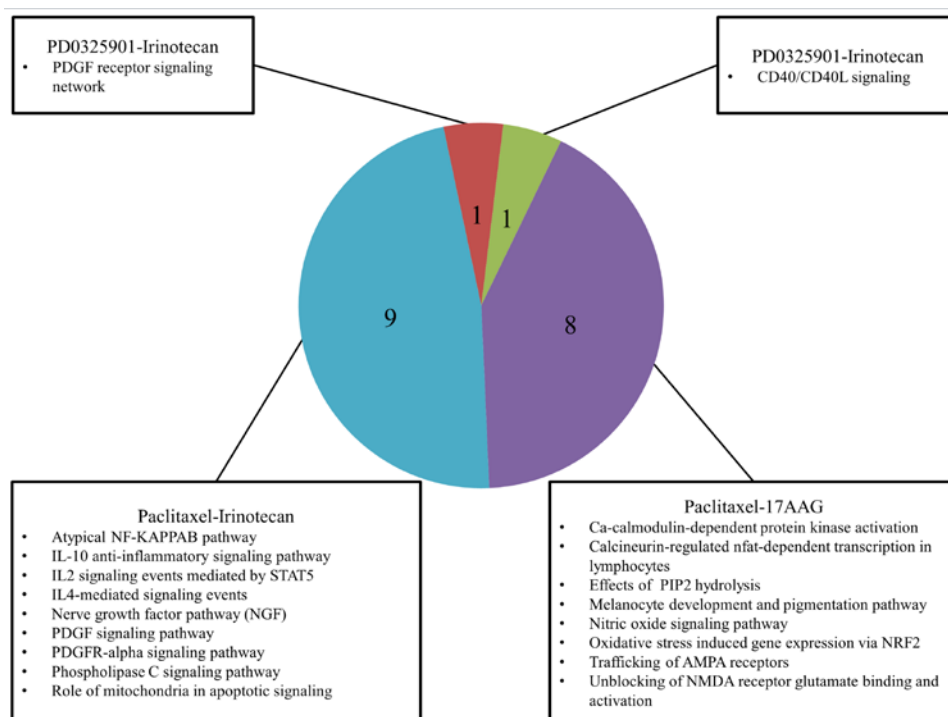


Figure 3. Number of common pathways implicated in drug resistance between 17AAG, Irinotecan, Paclitaxel, and PD0325901 (1,441 total pathways tested).

Paclitaxel also shared 8 mechanisms of drug resistance with 17AAG when comparing all; however none of these pathways overlapped between 17AAG and Irinotecan. Our results suggest that resistance to Paclitaxel is closely tied to that of Irinotecan and 17AAG.

We were able to identify common pathways which confer drug sensitivity in all three cancers to multiple drugs. 17AAG shared 7 sensitivity based biological pathways with Irinotecan and 4 with Paclitaxel. This is contrasted by the single biological pathway Paclitaxel and Irinotecan share associated with drug sensitivity. We can then begin to compare drugs on the basis of which biological pathways play a role in conferring drug sensitivity or resistance. Hierarchical schemas of drug similarity are illustrated in Figures 3 and 4.

Our results suggest that cancer cell resistant to Paclitaxel is likely to also resist 17AAG and Irinotecan. As Irinotecan and 17AAG appear to have quite distinct biological pathway mechanisms of action for drug resistance, it is less likely that a cancer cell line resistant to one will be resistant to another (Figure 3). On the other hand, as sensitivity to Irinotecan has some pathway similarities to sensitivity to 17AAG it is more likely that a cell line that is sensitive to one is sensitive to another (Figure 4). These results may suggest that a good starting point for the repurposing and repositioning of drugs is to classify them by their impact on the biological network of a cancer cell rather than by their distinct mechanism of action.

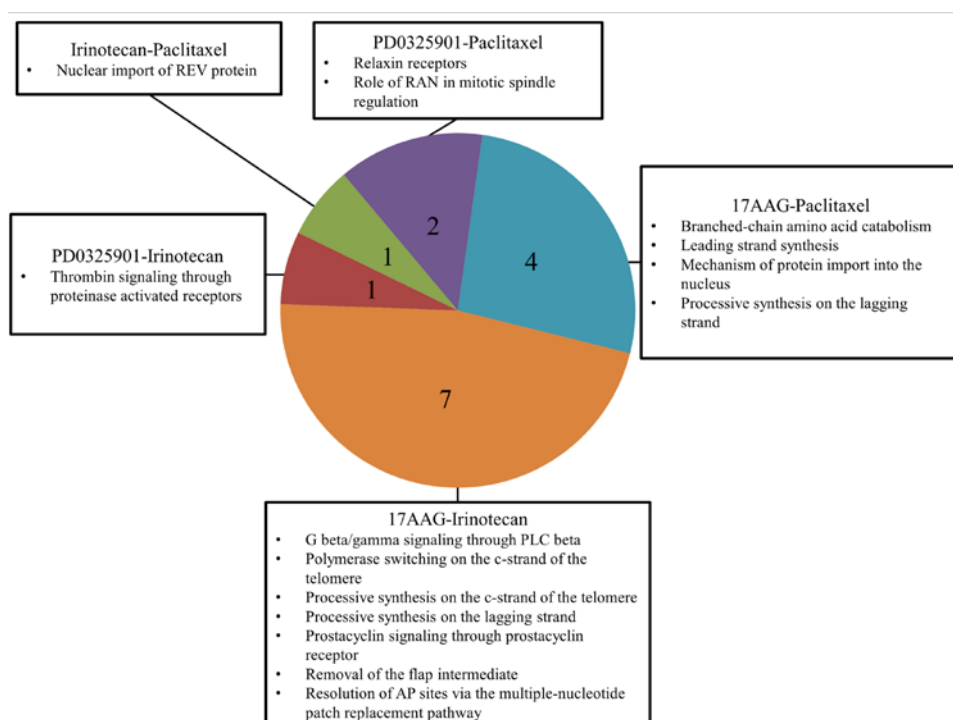


Figure 4. Number of common pathways implicated in drug sensitivity between 17AAG, Irinotecan, Paclitaxel, and PD0325901 (1,441 total pathways tested).

## 4. Discussion

Though there are some success stories, many clinical biomarkers have had limited impact<sup>7</sup>, and a shift is needed to more global explanations of disease and drug response phenotypes. Since a single gene is often involved in multiple pathways, it is difficult to assess the significance of a given genetic aberration without considering the broader context in which the dysregulation occurs<sup>7</sup>. In addition, many cancer patients have multiple genetic aberrations and multiple signaling pathways may be dysregulated and associated with drug resistance. However, the current analysis suggests that these signaling pathways related to drug resistance are shared by four drugs with completely different mechanisms of action. These results suggest that grouping drugs for treatment, repurposing, and repositioning by shared mechanisms which govern resistance and sensitivity may be more accurate than grouping them by the specific genes they appear to target. Such classification allows for simplification of the drug repurposing and repositioning process by making it a simple matter of counting and comparing biological pathway mechanisms.

DIRPP is a novel pipeline for classifying cell lines by drug sensitivity and for elucidating biological pathway mechanisms that drive drug response. PARADIGM forms the foundation of DIRPP and thus its scalability and comparability to other pathway based ones will be similar to that of PARADIGM. PARADIGM has been utilized in the hallmark TCGA papers and is an integral part of their pipeline easily scaling up to over 400 patient samples<sup>1-3</sup>. When compared to other pathway based methods, PARADIGM was shown to perform better compared to other methods<sup>7,9</sup>. Though PARADIGM has been used to compare separate groups of patients known to respond better to a selected therapies than others, it has never been used in a discovery manner as presented here. The DIRPP pipeline thus represents a novel extension of PARADIGM's capabilities. Though we chose to connect the drug-intervention node to the DNA entity in PARADIGM's central dogma, many drug targets are proteins. This could be reflected in future refinements of the method by modifying the connecting point for the drug-intervention node.

DIRPP performs comparably well on two independent datasets and is generalizable to other datasets with gene expression, copy number data, or both. The high predictive power of DIRPP across multiple drugs and cancers makes it a versatile tool to aid pre-clinical research. Further work to assess the utility of DIRPP is required. The CCLE and CGP datasets contain cell lines for ovarian cancer which were not screened for drug response and do not have IC<sub>50</sub> values. Once a robust ranking of cell lines with known drug response is built and the accuracy is validated, DIRPP can be used to classify the unknown cell line(s) as either sensitive or resistant to a particular drug. Further analysis will utilize the -omic data for tumor samples from TCGA and other publically available datasets to predict drug response phenotypes by applying the knowledge learned and methods developed from the current analysis.

The complexity of cancer presents many challenges to predicting therapeutic effectiveness if using individual biomarkers alone. Pathway level approaches such as DIRPP bring us one step closer to the goal of personalized medicine by utilizing complex -omic data and knowledge on biological pathways in order to robustly identify drug sensitivity.

## 5. Acknowledgements

We would like to thank the investigators of the cancer genome project and cancer cell line encyclopedia for their comprehensive data which enabled this analysis. We especially thank Dr. Steve Benz for his assistance with running PARADIGM and the entire team at Five3genomics for their support. This work was supported in part by the Case Comprehensive Cancer Center Support Grant NCI 5P30 CA043703.

## References

1. TCGA. *Nat.* **497**:67-73 (2013)
2. TCGA. *Nat.* **474**:609-615 (2011)
3. TCGA. *Nat.* **490**:61-70 (2012)
4. Garnett, M.J., et al., *Nat.* **483**(7391):570-575. (2012).
5. Barretina, J. et al., *Nat.* **483**(7391):603-607 (2012)
6. Subramanian, A. et al., *PNAS.* **102**:15545-15550 (2005)
7. Vaske, C.J. et al., *Bioinf.* **26**: i237-i247 (2010)
8. Ng, S. et al., *Bioinf.* **28**:i640-i646 (2012)
9. Varadan, V. et al., *IEEE Sig. Proc.* 35-49 (2011)
10. Nibbe, R.K. et al., *PLOS Comp. Bio.* **6**(1) (2009)
11. Nibbe, R.K. et al., *Journal of Comp. Bio.* **18**(3)263-281 (2010)
12. Papillon-Cavanagh, S. et al., *J Am Med Inform Assoc.* **20**:597-602 (2013)
13. Napolitano, F. et al., *Journal of Cheminformatics.* **5**:30 (2013)
14. McCarthy, J.J. et al., *Science Translational Med.* **5**(189):1-17 (2013)
15. Heiser, L.M. et al., *PNAS.* **109**:2724-2729 (2011)
16. Knox C. et al., *Nucleic Acids Res.* **39**(Database issue):D1035-41. PMID: 21059682 (2011)
17. Wishart DS. et al., *Nucleic Acids Res.* **36**(Database issue):D901-6. PMID: 18048412 (2008)
18. Wishart DS. et al. *Nucleic Acids Res.* **34**(Database issue):D668-72. PMID: 16381955 (2006)
19. Domcke, S. et al. *Nat. Commun.* **4**:2126 doi: 10.1038/ncomms3126 (2013)
20. Isonishi, S. et al. *Oncology Reports.* **18**(1):195-201 (2007)
21. Weigel, M.T. et al. *Annals of Oncology.* **24**(1): 126-133. doi:10.1093/annonc/mds240 (2013)

## ANTI-INFECTIOUS DRUG REPURPOSING USING AN INTEGRATED CHEMICAL GENOMICS AND STRUCTURAL SYSTEMS BIOLOGY APPROACH

CLARA NG

*Department of Computer Science, Hunter College, the City University of New York,  
695 Park Avenue, New York City, NY 10065, U. S. A.  
Email: cng0003@hunter.cuny.edu*

RUTH HAUPTMAN

*Department of Computer Science, Hunter College, the City University of New York,  
695 Park Avenue, New York City, NY 10065, U. S. A.  
Email: rhauptma@hunter.cuny.edu*

YINLIANG ZHANG

*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093, U. S. A.  
Email: yiz071@ucsd.edu*

PHILIP E. BOURNE

*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093, U. S. A.  
Email: pbourne@ucsd.edu*

LEI XIE

*Department of Computer Science, Hunter College, The Graduate Center, the City University of New York  
695 Park Avenue, New York City, NY 10065, U. S. A.  
Email: lei.xie@hunter.cuny.edu*

The emergence of multi-drug and extensive drug resistance of microbes to antibiotics poses a great threat to human health. Although drug repurposing is a promising solution for accelerating the drug development process, its application to anti-infectious drug discovery is limited by the scope of existing phenotype-, ligand-, or target-based methods. In this paper we introduce a new computational strategy to determine the genome-wide molecular targets of bioactive compounds in both human and bacterial genomes. Our method is based on the use of a novel algorithm, ligand Enrichment of Network Topological Similarity (ligENTS), to map the chemical universe to its global pharmacological space. ligENTS outperforms the state-of-the-art algorithms in identifying novel drug-target relationships. Furthermore, we integrate ligENTS with our structural systems biology platform to identify drug repurposing opportunities via target similarity profiling. Using this integrated strategy, we have identified novel *P. falciparum* targets of drug-like active compounds from the Malaria Box, and suggest that a number of approved drugs may be active against malaria. This study demonstrates the potential of an integrative chemical genomics and structural systems biology approach to drug repurposing.



## 1. Introduction

Treatment of infectious diseases is under threat. The emergence of multi-drug resistance and extensively drug resistant microbes to antibiotics calls for new treatment regimes.<sup>1</sup> Yet, at the same time, the drug discovery process, characterized by a one-drug-one-genome-one-disease paradigm, has yielded few successes in combating drug resistance and is hampered by a high failure rate leading to soaring costs.<sup>2</sup> Fortunately, the cause of that failure is also cause for optimism. Since the failure is due, in part, to drug promiscuity there is also the opportunity to repurpose existing drugs to treat infectious diseases.<sup>3</sup> However, there are several unique challenges in anti-infectious drug repurposing. First, successful phenotype-based methods which compare the genome-wide molecular signature of repositioned drugs to a disease-induced phenotype,<sup>4</sup> have limitations when applied to anti-infectious drug discovery. Second, recent efforts in cell-based antibiotics screening produce thousands of active compounds, but gives few hints as to their molecular targets as well as their *in vivo* activities and toxicities.<sup>5-6</sup> Finally, due to the bias in high-throughput screening, existing chemical genomics databases only collect several thousand targets, most of which are from human and model organisms, not pathogens. Taken together these limitations hinder the application of state-of-the-art computational methods to anti-infectious drug repurposing.

These limitations can be addressed through *chemical genomics* - the construction of genome-scale drug-target interaction networks. Creating such networks requires that we address the question, given a chemical entity, how do we accurately identify its targets on a genome scale based on its structural similarity with known ligands and reliably determine the significance of those putative targets? Several data mining techniques have recently been developed to predict drug-target interactions.<sup>7-15</sup> However, few of them can assess the statistical significance of ranked targets. A notable advance was the development of Similarity Ensemble Approach (SEA) statistical model,<sup>16-17</sup> which is comparable to the state-of-the-art machine learning algorithms.<sup>18</sup> However, SEA and most of the existing machine learning techniques only consider local neighborhoods for relevance between chemicals.<sup>19</sup> Thus it remains a big challenge to find the global relationships between chemicals so that an expanded target space can be established.<sup>20-27</sup> In this paper, we introduce a fundamentally new methodology, ligand Enrichment of Network Topological Similarity (ligENTS), which integrates graph mining algorithms and random set theory to begin to address the above challenges. ligENTS considerably improves the performance of existing methods for drug-target prediction. Thus, ligENTS may open new doors to the next generation of chemical genomics algorithms.

The integration of chemical genomics and structural genomics is needed since current chemical genomics methods have only identified targets for a small portion of the human (<10%) and pathogen genomes (often <1%), respectively.<sup>28</sup> In other words the molecular targets of a large number of active compounds against bacteria are still unknown. Complementary to the knowledge of existing drug targets, the structural information of proteins has increased rapidly.<sup>29</sup> Previously,

starting from a known drug-target set, we have developed a *structural systems biology* approach for linking drug molecules to pathogen structural genomes through target binding site similarity, thereby reconstructing high-resolution 3D drug-target physical interaction models.<sup>30</sup> However, these structural systems biology methods are not scalable to millions of chemicals. To address these limitations, we combine ligENTS with the structural systems biology approach to link entire bioactive chemical space to the pathogen structural genome. The innovative integration of chemical genomics with structural systems biology will not only greatly expand the scope of both ligand- and target-based methods, but also considerably improve the quality of predicted drug-target interaction models. Consequently, it may provide new opportunities for drug discovery.

To demonstrate the utility of this integrated approach, we apply it to identify molecular targets of drug-like compounds from the Malaria Box, and suggest drug repurposing opportunities for anti-malaria chemotherapies. Malaria is one of the most devastating and widespread tropical parasitic diseases and is the most prevalent in developing countries.<sup>31</sup> The Malaria Box includes 200 drug-like and 200 probe-like compounds that are active against the blood stage of *P. falciparum*, one of the most dangerous pathogen causing malaria. Although the compounds have desirable ADMET properties, their molecular targets in bacteria and human, as well as *in vivo* activity and toxicity, remain unknown. We use ligENTS to identify their target profiles in the chemical genomics databases, and their mapping to the *P. falciparum* genome. Using the target profile of active compounds as a proxy, we link approved drugs with active compounds against *P. falciparum*. Our results provide abundant testable hypothesis for further experimental validation.

## 2. Results and Discussion

### 2.1. Ligand Enrichment of Network Topological Similarity (ligENTS) method

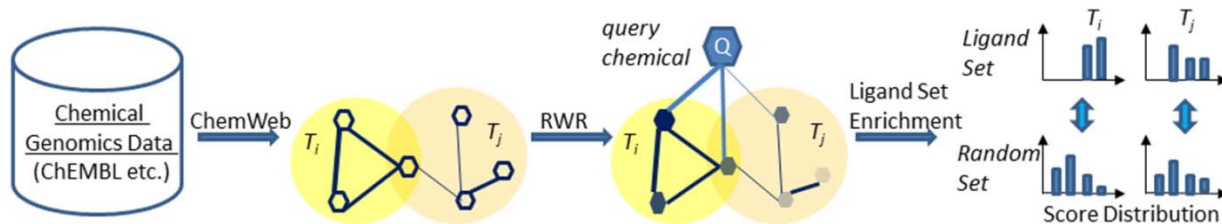


Fig. 1. Scheme of ligENTS. Hexagons represent chemicals. Two similar chemicals are connected. The more similar a chemical is to the query, the darker the hexagon. The chemicals in the colored sphere bind to corresponding targets  $T_i$  and  $T_j$ .

We have developed a new algorithm, ligand Enrichment of Network Topological Similarity (ligENTS), to assess the statistical significance of chemical-target associations based on the network topological similarity. As shown in Fig. 1, ligENTS consists of three key steps. (1) We connect around half a million chemicals in ChEMBL<sup>32</sup> into a chemical similarity network (termed ChemWeb). (2) Given a query, we apply a Random Walk with Restart (RWR) algorithm to define the network topological similarity between the query and other chemicals in ChemWeb. (3) To assess the statistical significance of the topological rank derived from the RWR, we apply random set theory to estimate the enrichment of a ligand set that is associated with a protein target in terms

of the distribution of its network topological similarity scores. The final output of ligENTS is the false discovery rate (FDR) of a list of targets in the database, which may interact with the query chemical.

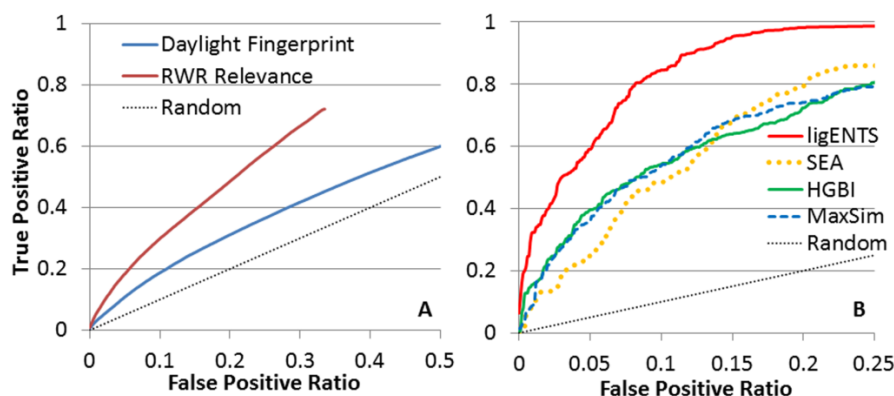


Fig. 2. Performance comparison of (A) global RWR relevance and Daylight fingerprint in detecting a pair of chemicals that share the same target, and (B) ligENTS (ligand Enrichment of Network Topological Similarity), SEA (Similarity Ensemble Approach)<sup>16</sup>, HGBI (Heterogeneous Graph Based Inference)<sup>15</sup>, and MaxSim (maximum similarity score in a set of ligands)<sup>33</sup> in ranking targets given a query chemical.

## 2.2. Graph mining improves the performance of detecting pairwise chemical similarity

State-of-the-art algorithms such as SEA, TurboSim,<sup>34</sup> MaxSim,<sup>33</sup> and IRV<sup>19</sup> only consider the similarity between the nearest neighbors, but ignores the global structure similarity relationships among all entities in a database. To overcome this limitation, we apply graph mining algorithms to define global relationships between chemicals. Given a query chemical, we first link the query to all nodes in ChemWeb, if edge weights between the query and any node are above a predefined threshold. Then, we use a Random Walk with Restart (RWR) algorithm to perform a probabilistic traversal of ChemWeb across all paths leading away from the query. The probability of choosing a path depends on the edge weight. The output of the algorithm is the list of all nodes (chemicals) in the network, ranked by the probability  $p_i$  for the query to reach node  $i$ . In this way, the query may detect related chemicals that are missed by the direct neighbors through intermediate nodes. As shown in Fig. 2A, a RWR transversal of ChemWeb improves the sensitivity and specificity of pair-wise chemical similarity search over a Daylight fingerprint similarity (<http://www.daylight.com>). When the Tanimoto Coefficient (TC) is 0.57 (approximately false positive ratio of 0.1), the Daylight fingerprint only identifies around 20% of all ligand pairs that bind to the same target. Using the same threshold to construct ChemWeb, the sensitivity of RWR is approximately 0.30, 50% more than that of the Daylight fingerprint. Thus the exploration of global community structures within the chemical similarity network allows us to detect novel protein-ligand interactions.

## 2.3. Ligand Enrichment of Network Topological Similarity (ligENTS) considerably improves the performance of detecting novel drug-target associations

Conventional ligand-based virtual screening focuses on ranking putative active compounds to a specific target. The issue that we need to address here is a reverse screening problem. Given a query chemical, how can we reliably rank all protein targets in a database by their likelihood to interact with the query chemical? To detect novel protein-chemical interactions, we developed a new algorithm, Ligand Enrichment of Network Topological Similarity (ligENTS). ligENTS combines RWR/ChemWeb with a ligand set enrichment framework. We compare the performance of ligENTS with three state-of-the-art algorithms: Similarity Ensemble Approach (SEA),<sup>16</sup> Heterogeneous Graph Based Inference (HGBI),<sup>15</sup> and the target assignment based on the most similar chemical in a ligand set (MaxSim).<sup>33</sup> SEA normalizes the sum of similarity scores between two sets of ligands known to bind to their targets, based on an empirical extreme value distribution model, and in an extensive benchmark study, SEA outperforms a state-of-the-art machine learning method.<sup>18</sup> SEA is the most relevant comparison to ligENTS in terms of statistical models for evaluating the chemical-target association. MaxSim is found to be the best performing method for ligand-based virtual screening when multiple ligands are used as a profile.<sup>33</sup> For comparison we modified the MaxSim algorithm to rank targets based on the maximum similarity score when comparing their ligands to the query. HGBI applies RWR on a heterogeneous drug-drug, drug-target, and target-target network to infer drug-target interactions, and outperforms other network inference algorithms for drug-target prediction.<sup>15</sup> As shown in Fig. 2B, HGBI is slightly better in the low false positive region than MaxSim. Consistent with a recent study in evaluating the performance of ligand profiles,<sup>33</sup> MaxSim outperforms SEA when the false positive rate is less than 0.15. Although HGBI is one of the best performers of the three of existing methods, HGBI does not provide a statistical significance assessment for predicted interactions.

LigENTS outperforms the above three methods in identifying novel chemical-target relationships, as shown in Fig. 2B. ligENTS identifies 200% and 50% of true positives more than that of HGBI at a false positive ratio of 0.01 and 0.05, respectively. The superior performance of ligENTS comes from its combination of the RWR search and global set statistics. The RWR captures the global structure of chemical space. However, conventional statistics models such as SEA fail when applied to global similarity problems. Global set statistics is more powerful than the fitted parametric statistical model. However, it is less useful when only the nearest neighbors are considered, as the scores of most ligands in the set are zeros, providing no information for the hypothesis testing. Enrichment of Network Topological Similarity (ENTS) by integrating RWR and global set statistics provides a general framework to enhance similarity search and association detection. Although this paper focuses on its application to chemical-target prediction, we have shown that ENTS improves the performance of protein fold recognition, RNA structure prediction, and disease gene identification. These results will be published elsewhere.

#### ***2.4. Prediction of molecular targets of Malaria Box in the chemical genomics database***

To demonstrate the application of ligENTS to drug repurposing, we first use it to identify molecular targets of drug-like compounds from the Malaria Box, which are annotated in ChEMBL.

At a false discovery rate of 0.05, we associate 161 out of 200 drug-like active compounds from the Malaria Box with more than 577 proteins annotated in ChEMBL. The majority of these hits (~80%) are proteins from human and animal models. This reflects the screening and annotation bias in the chemical genomics databases. Nevertheless, enriched biological processes for these genes may provide valuable information on potential side effects (e.g., regulation of blood pressure, and muscle contraction) of these compounds, or their impact on pathogen-host interactions (e.g., response to molecule of bacterial origin), as shown in Table 1.

Table 1. Enriched biological processes of molecular targets of human and animal models for drug-like compounds from the Malaria Box.

Biological process	False Discovery Rate
second-messenger-mediated signaling	1.171e-58
positive regulation of lipase activity	3.028e-23
calcium ion homeostasis	1.923e-22
oxidoreductase activity	6.961e-17
regulation of blood pressure	1.117e-15
inflammatory response	1.593e-10
phosphoric diester hydrolase activity	3.745e-10
smooth muscle contraction	2.888e-07
regulation of apoptosis	2.353e-05
response to molecule of bacterial origin	3.870e-03

## 2.5 Prediction of molecular targets of drug-like compounds in *P. falciparum*

To identify the *P. falciparum* targets of drug-like compounds from the Malaria Box, we mapped the targets identified from the chemical genomics databases to the *P. falciparum* genome using both sequence similarity and ligand binding site similarity. Most of the mapped targets are essential genes in *P. falciparum*. Some of them (e.g., dihydroorotate dehydrogenase, beta-hydroxyacyl-ACP dehydratase, cysteine protease falcipain-3, and type II DNA topoisomerase) are novel targets under investigation.<sup>35-38</sup> When we rank the targets by the number of binding compounds, the top ranked targets include several proteins that bind to quinine, one of the most efficient drugs to treat malaria, providing support for our predictions. Other proteins include the JmjC domain containing protein, 3-oxoacyl-acyl-carrier protein reductase, and several putative transporters. The JmjC domain containing protein is particularly interesting. Twelve compounds are predicted to interact with JmjC. JmjC plays a key role in chromatin remodeling and histone posttranslational modifications that is fundamentally important in the developmental program of *P.*

*falciparum*.<sup>39</sup> However, this protein has not been explored as a drug target. Because the human homolog of JmjC exists, the detailed analysis of the drug binding site features may provide critical information on developing selective anti-malaria chemotherapy targeting JmjC. This analysis is ongoing.

## 2.6 Repurposing approved drugs to target *P. falciparum*

We apply ligENTS to 1,484 approved small molecule drugs in DrugBank to identify their molecular targets in ChEMBL. If the target profile of a drug is similar to that of the active compounds from the Malaria Box, we hypothesize that the drug is active against malaria. We term this strategy Target Similarity Profiling (TSP). Based on TSP, Table 2 shows the top ranked drugs that have the potential to treat malaria. The top hit sirolimus is a macrolide compound, targeting the FK506 binding protein. It has been used as an anti-fungal and an anti-neoplastic agent. FK506 binding protein in *P. falciparum* has been suggested as a novel target to fight malaria infection.<sup>40</sup> Several other drugs are predicted to target phosphodiesterase, dihydrofolate reductase, protease, carbonic anhydrase, somatostatin receptor, and ion channels. All these proteins are novel targets for anti-malaria therapeutics.<sup>41-46</sup> Doxycycline is a known anti-malaria agent, providing putative validation to TSP predictions. Thus, TSP provides abundant testable hypotheses for anti-malaria drug repurposing.

Table 2. Top 10 ranked drugs by TSP and their predicted target by ligENTS

Drug	Target(s)	Primary indication
Sirolimus	FK506 binding protein	anti-fungal and anti-neoplastic
Acitretin	Lyase, Nitric oxide synthase, DNA methyltransferase, Collagenase	treatment of psoriasis
Roflumilast Ph	osphodiesterase (PDE)	chronic obstructive pulmonary disease
Trimetrexate	dihydrofolate reductase (DHFR)	Antibiotics
Metaxalone Prot	ease	muscle relaxant
Piperazine C	arbonic anhydrase	Anthelmintic
Doxycycline dem	ethylase, hydrolase, dehydrogenase	Anti-malaria
Octreotide Som	atostatin receptor	treatment of acromegaly and reduction of side effects from cancer chemotherapy
Benazepril	Sodium channel subunit alpha, Voltage-dependent calcium channel subunit alpha	Hypertension

### 3. Conclusion

In this paper, we introduce a new chemical genomics algorithm, ligENTS, to map the chemical universe to its global pharmacological space, as well as an integrated chemical genomics and structural systems biology approach for anti-infectious drug repurposing. Although the detailed implementation of the algorithm needs to be improved, its prototype outperforms existing state-of-the-art methods, and demonstrates the potential for use in anti-infectious drug repurposing. The further development of this new strategy may consolidate phenotype-, ligand-, and target-based drug discovery, thereby facilitating the transformation of the conventional drug discovery process to a new paradigm of systems pharmacology.

## 4. Methods

### 4.1. Benchmark

We extract positive and negative cases from the bioactivity database ChEMBL<sup>32</sup>. To reduce the chance of including false positive hits, we only include those pairs with  $IC_{50} < 10.0 \mu M$  as positive cases. The negative cases include those pairs in which no binding is detected. We define the benchmark using the intersection of ligand sets in the positive and negative cases. After removing the chemical redundancy (Tanimoto Coefficients (TC) of 0.85, a common threshold in virtual screening), the final benchmark includes 390 chemicals, which involve 803 true and 1,336 false chemical-target interactions, respectively. We evaluate the sensitivity and specificity of the ranked target for a benchmark chemical when querying ChemWeb in which all benchmark chemicals are excluded.

### 4.2. Construction of similarity matrix of ChemWeb

Using a Daylight fingerprint representation of each chemical and TC as a similarity measure, we connect 415,975 chemicals that have high confidence annotation to targets in ChEMBL into a pairwise chemical similarity network. We represent ChemWeb as a weighted graph, in which nodes are chemicals. An edge is formed between two chemicals if they share the same activity and their chemical similarity is above a certain threshold. With a TC larger than 0.57, a threshold used by SEA but not optimized for ligENTS, ChemWeb consists of more than 10 million edges. We represent the ChemWeb weighted graph as a similarity matrix  $W$ .

### 4.3. Implementation of Random Walk with Restart (RWR) algorithm

We modified the RankProp algorithm,<sup>47</sup> a variant of RWR, and implemented it using a boost library (<http://www.boost.org>). The pseudo code of the algorithm is shown as follows.

Input: A graph representation of ChemWeb, with  $i = 1, \dots, N$  chemicals and their chemical similarity matrix  $W$  with the instance of  $w_{ji}$ ; a diffusion vector  $A$  with the instance of  $a_i$ , and a query chemical  $q$ .

Initialization:  $p_q(0) = 1; p_i(0) = 0$   
 while  $t = 0, 1, 2, \dots$  do  
   for  $i = 1$  to  $N$  do  
      $p_i(t+1) = w_{qi} + a_i \sum_{j=1}^N w_{ji} p_j(t)$   
   end for  
 until convergence  $t = T^*$   
 output: a ranked list of  $p_i(T^*)$

$a_i$  corresponds to the restart probability in the RWR and determines how far the query will propagate through ChemWeb. In this study,  $a_i$  was set as a constant of 0.65.

#### 4.4. Implementation of set statistics

Inspired by Gene Set Enrichment Analysis, we adapted the random set method<sup>48</sup> to estimate the enrichment of a ligand set that is associated with a protein target. For the RWR output  $p_i(T^*)$ ,  $i = 1, \dots, N$ , an unnormalized score for a ligand set  $S$  consisting of  $m$  chemicals is calculated as the average of the RWR outputs of these chemicals

$$\bar{X} = \frac{\sum_{p_j \in S} p_j}{m}$$

To compare the enrichment in a ligand set  $S$  with that of all other ( $N, m$ ) distinct randomly drawn ligand sets of size  $m$ , the ligand set  $S$  is now considered as a random collection of  $m$  ligands whose score  $p_j$  are fixed. The exact distribution of  $\bar{X}$  is intractable, but can be approximated with the normal distribution with mean and variance as follows:

$$\mu = \frac{1}{N} \sum_{j=1}^N s_j$$

$$\sigma^2 = \frac{1}{m} \left( \frac{N-m}{N-1} \right) \left[ \left( \frac{1}{N} \sum_{j=1}^N s_j^2 \right) - \left( \frac{1}{N} \sum_{j=1}^N s_j \right)^2 \right]$$

The enrichment score is then normalized with

$$Z = \frac{\bar{X} - \mu}{\sigma}.$$

The false discovery rate (FDR) is estimated by fitting the enrichment score  $Z$  with the false positive ratio from the benchmark.

#### 4.5. Target identification of active compounds from the Malaria Box in the ChEMBL database and *P. falciparum* genomes

LigENTS was first used to identify potential molecular targets of active compounds from the Malaria Box found in the ChEMBL database. Because most of the targets in ChEMBL are from



human or model organisms, SMAP<sup>49-51</sup> and PSI-Blast<sup>52</sup> are applied to map the targets identified by ligENTS, which are not from *P. falciparum* genome, to *P. falciparum* proteins.

#### 4.6. Functional Enrichment Analysis

Functional Enrichment Analysis of human targets is carried out using the DAVID functional annotation tool ( <http://david.abcc.ncifcrf.gov/>). The whole genome of Homo sapiens is used as background.

#### Acknowledgments

This research was supported, in part, under National Science Foundation Grants CNS-0958379 and CNS-0855217 and the City University of New York High Performance Computing Center at the College of Staten Island. C.N. and R.H. were supported by the John P. McNulty Scholars Program.

#### References

1. World Health Organization. Fact sheet N°194 (2012).
2. S. Nwaka and A. Hudson. *Nat Rev Drug Discov* **5**, 941-955 (2006).
3. A. P. Chiang and A. J. Butte. *Clin Pharmacol Ther* **86**, 507-510 (2009).
4. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, *et al.* *Sci Transl Med* **3**, 96ra76 (2011).
5. W. A. Guiguemde, A. A. Shelat, D. Bouck, S. Duffy, *et al.* *Nature* **465**, 311-315 (2010).
6. F. J. Gamo, L. M. Sanz, J. Vidal, C. de Cozar, *et al.* *Nature* **465**, 305-310 (2010).
7. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, *et al.* *Bioinformatics* **24**, i232-240 (2008).
8. N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, *et al.* *PLoS Comput Biol* **5**, e1000397 (2009).
9. D. Vina, E. Uriarte, F. Orallo and H. Gonzalez-Diaz. *Mol Pharm* **6**, 825-835 (2009).
10. A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan. *Mol Syst Biol* **7**, 496 (2011).
11. F. Cheng, C. Liu, J. Jiang, W. Lu, *et al.* *PLoS Comput Biol* **8**, e1002503 (2012).
12. J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li, *et al.* *Bioinformatics* **29**, 238-245 (2013).
13. T. van Laarhoven and E. Marchiori. *PLoS One* **8**, e66952 (2013).
14. S. Alaimo, A. Pulvirenti, R. Giugno and A. Ferro. *Bioinformatics* **29**, 2004-2008 (2013).
15. W. Wang, S. Yang and J. Li. *Pac Symp Biocomput*, 53-64 (2013).
16. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, *et al.* *Nat Biotechnol* **25**, 197-206 (2007).
17. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, *et al.* *Nature* **462**, 175-181 (2009).
18. J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea, *et al.* *J Chem Inf Model* **48**, 755-765 (2008).

19. S. J. Swamidass, C. A. Azencott, T. W. Lin, H. Gramajo, *et al.* *J Chem Inf Model* **49**, 756-766 (2009).
20. V. Namasivayam, P. Iyer and J. Bajorath. *Chem Biol Drug Des* **79**, 22-29 (2012).
21. H. Sun, G. Tawa and A. Wallqvist. *Drug Discov Today* **17**, 310-324 (2012).
22. R. Guha. *J Chem Inf Model* **52**, 2181-2191 (2012).
23. J. J. Irwin. *Nat Chem Biol* **5**, 536-537 (2009).
24. S. Renner, W. A. van Otterlo, M. Dominguez Seoane, S. Mocklinghoff, *et al.* *Nat Chem Biol* **5**, 585-592 (2009).
25. R. D. Cramer. *J Comput Aided Mol Des* **25**, 197-201 (2011).
26. G. Schneider. *Nat Rev Drug Discov* **9**, 273-276 (2010).
27. G. Hu, G. Kuang, W. Xiao, W. Li, *et al.* *J Chem Inf Model* **52**, 1103-1113 (2012).
28. L. Xie, L. Xie and P. E. Bourne. *Curr Opin Struct Biol* **21**, 189-199 (2011).
29. H. M. Berman, B. Coimbatore Narayanan, L. D. Costanzo, S. Dutta, *et al.* *Febs Lett* (2013).
30. S. L. Kinnings, L. Xie, K. Fung, L. Xie, *et al.* *PLoS Comp Biol* **6**, e100976 (2010).
31. World Health Organization. *World malaria report 2008* (2008).
32. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, *et al.* *Nucleic Acids Res* **40**, D1100-1107 (2012).
33. R. J. Nasr, S. J. Swamidass and P. F. Baldi. *J Cheminform* **1**, 7 (2009).
34. J. Hert, P. Willett, D. J. Wilton, P. Acklin, *et al.* *J Med Chem* **48**, 7049-7054 (2005).
35. M. A. Phillips and P. K. Rathod. *Infectious disorders drug targets* **10**, 226-239 (2010).
36. D. Kostrewa, F. K. Winkler, G. Folkers, L. Scapozza, *et al.* *Protein Sci* **14**, 1570-1580 (2005).
37. C. Teixeira, J. R. Gomes and P. Gomes. *Curr Med Chem* **18**, 1555-1572 (2011).
38. C. Garcia-Estrada, C. F. Prada, C. Fernandez-Rubio, F. Rojo-Vazquez, *et al.* *Proc Biol Sci* **277**, 1777-1787 (2010).
39. L. Cui and J. Miao. *Eukaryot Cell* **9**, 1138-1149 (2010).
40. N. Bharatham, M. W. Chang and H. S. Yoon. *Curr Med Chem* **18**, 1874-1889 (2011).
41. K. Yuasa, F. Mi-Ichi, T. Kobayashi, M. Yamanouchi, *et al.* *Biochem J* **392**, 221-229 (2005).
42. Y. Yuthavong, B. Tarnchompoo, T. Vilaivan, P. Chitnumsub, *et al.* *Proc Natl Acad Sci U S A* **109**, 16823-16828 (2012).
43. C. Wegscheid-Gerlach, H. D. Gerber and W. E. Diederich. *Curr Top Med Chem* **10**, 346-367 (2010).
44. S. Reungprapavut, S. R. Krungkrai and J. Krungkrai. *J Enzyme Inhib Med Chem* **19**, 249-256 (2004).
45. J. X. Pan, R. B. Mikkelsen, D. F. Wallach and C. R. Asher. *Mol Biochem Parasitol* **25**, 107-

111 (1987).

46. S. A. Desai. *Curr Drug Targets Infect Disord* **4**, 79-86 (2004).
47. I. Melvin, J. Weston, C. Leslie and W. S. Noble. *Bioinformatics* **25**, 121-122 (2009).
48. M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, *et al.* *Ann. Appl. Stat.* **1**, 85-106 (2007).
49. L. Xie and P. E. Bourne. *BMC Bioinformatics* **8 Suppl 4**, S9 (2007).
50. L. Xie and P. E. Bourne. *Proc Natl Acad Sci U S A* **105**, 5441-5446 (2008).
51. L. Xie and P. E. Bourne. *Bioinformatics* **25**, i305-312 (2009).
52. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, *et al.* *Nucleic Acids Res.* **25**, 3389-3402 (1997).

# DRUG-TARGET INTERACTION PREDICTION BY INTEGRATING CHEMICAL, GENOMIC, FUNCTIONAL AND PHARMACOLOGICAL DATA

FAN YANG<sup>†</sup>, JINBO XU<sup>‡</sup>, JIANYANG ZENG<sup>§,\*</sup>

<sup>†</sup>*Department of Mathematical Sciences  
Tsinghua University  
Beijing, 100084, P. R. China  
E-mail: f-yang10@mails.tsinghua.edu.cn*

<sup>‡</sup>*Toyota Technological Institute at Chicago  
6045 S. Kenwood Ave.  
Chicago, IL 60637, USA  
E-mail: j3xu@ttic.edu*

<sup>§</sup>*Institute for Interdisciplinary Information Sciences  
Tsinghua University  
Beijing, 100084, P. R. China  
E-mail: zengjy321@tsinghua.edu.cn  
\*Corresponding author*

*In silico* prediction of unknown drug-target interactions (DTIs) has become a popular tool for drug repositioning and drug development. A key challenge in DTI prediction lies in integrating multiple types of data for accurate DTI prediction. Although recent studies have demonstrated that genomic, chemical and pharmacological data can provide reliable information for DTI prediction, it remains unclear whether functional information on proteins can also contribute to this task. Little work has been developed to combine such information with other data to identify new interactions between drugs and targets. In this paper, we introduce functional data into DTI prediction and construct biological space for targets using the functional similarity measure. We present a probabilistic graphical model, called *conditional random field* (CRF), to systematically integrate genomic, chemical, functional and pharmacological data plus the topology of DTI networks into a unified framework to predict missing DTIs. Tests on two benchmark datasets show that our method can achieve excellent prediction performance with the area under the precision-recall curve (AUPR) up to 94.9. These results demonstrate that our CRF model can successfully exploit heterogeneous data to capture the latent correlations of DTIs, and thus will be practically useful for drug repositioning. Supplementary Material is available at [http://iiis.tsinghua.edu.cn/~compbio/papers/psb2014/psb2014\\_sm.pdf](http://iiis.tsinghua.edu.cn/~compbio/papers/psb2014/psb2014_sm.pdf).

**Keywords:** Drug-Target Interaction; Drug Repositioning; Conditional Random Field; Functional Similarity.

## 1. Introduction

In recent years, *drug repositioning* or *drug repurposing* has become an increasingly popular trend in drug discovery.<sup>1-4</sup> The main goal of drug repositioning is to reuse existing or abandoned drugs and identify their new therapeutic functions. Recent literature reveals that drugs often possess the so-called *promiscuity* property,<sup>5,6</sup> that is, individual drugs can act on other off-target proteins in addition to the original target. This property provides a strong theoretical support for drug repositioning.

*In silico* prediction of drug-target interactions (DTIs) has been widely applied in drug

repositioning, since it can significantly reduce time and cost of drug development. Molecular docking methods have been commonly used in predicting new DTIs if structure coordinates of both proteins and drugs are available.<sup>7–10</sup> When three-dimensional (3D) structures of molecules are absent, we need to depend on other approaches to perform DTI prediction. The structure-free approaches can be roughly divided into two categories: *ligand-based* and *network-based* methods. Ligand-based methods exploit ligand similarity to identify new targets that can interact with a query drug.<sup>11,12</sup> Although with some successful stories, ligand-based approaches have difficulty in identifying new interactions associated with novel binding scaffolds.<sup>13</sup> Network-based methods<sup>14–20</sup> detect the latent correlation features of DTIs to predict new interactions, and recently have become a popular tool for drug repositioning and drug development. A key challenge in network-based prediction approaches lies in integrating heterogeneous data for accurate DTI prediction. Traditional DTI prediction approaches often relate genomic and chemical data with DTI networks to perform new prediction.<sup>21</sup> Recently, pharmacological data such as drug side-effects have also been taken into consideration,<sup>18,20,22–24</sup> and the results suggest that incorporating more data into DTI prediction can further improve prediction accuracy. Most existing network-based approaches mainly rely on the sequence similarity to measure the closeness of two targets. The sequence similarity, however, is not necessarily sufficient enough to characterize the shared patterns of DTI profiles between two targets.

Functional similarity enables us to compare two proteins with respect to their molecular and biological functions.<sup>25</sup> It is defined mainly based on Gene Ontology (GO) terms, which indicate the biological roles of gene products. This measure can identify functionally-related proteins regardless of homology, and hence provide additional information about the similarity of two targets aside from their genomic data. Based on functional similarity, we can construct biological space for proteins and analyze their DTI patterns from a different angle.

Although numerous approaches<sup>18,20,23,24,26</sup> have been proposed to integrate genomic (i.e., protein sequences), chemical (i.e., chemical substructures of drugs) and pharmacological (i.e., drug side-effects) data for predicting unknown DTIs, functional information has not been well exploited in DTI prediction. To our knowledge, little work has been developed to systematically integrate functional information on proteins with the aforementioned data to predict missing interactions between drugs and targets. In this paper, we present a new approach to address the DTI prediction problem by systematically integrating large-scale chemical, pharmacological, genomic and functional data and DTI network information into a unified framework. Our method applies a probabilistic graphical model, called *conditional random field* (CRF), to encode the complicated network associated with drugs and targets, and predict new DTIs. We apply a *stochastic gradient ascent* approach plus the *contrastive divergence* (CD) algorithm<sup>27</sup> to train our graphical model and capture the hidden correlations between drugs and targets. Tests on two benchmark datasets derived from multiple publicly-available databases show that our CRF model can effectively integrate multiple sources of information and achieve excellent prediction performance, with the area under the precision-recall curve (AUPR) up to 94.9. These results indicate that our approach can have potential applications in drug repositioning.

In summary, the following contributions are made in this paper: (1) Introduction of func-

tional data into DTI prediction and construction of biological space for proteins using the functional similarity measure; (2) Development of a new machine learning approach that can systematically integrate heterogeneous data into a unified framework to predict unknown DTIs; and (3) Promising testing results on two benchmark datasets.

## 2. Methods

### 2.1. Conditional Random Field Framework

Conditional random field (CRF) is a probabilistic graphical model or a variant of Markov random field<sup>28–30</sup> that was first proposed for object recognition and image segmentation.<sup>31</sup> Now it has been widely used in many fields such as shallow parsing,<sup>32</sup> named entity recognition,<sup>33</sup> topic distillation,<sup>34</sup> social recommendation<sup>35</sup> and molecular structural modelling.<sup>36</sup> We apply a binary CRF model<sup>34,35</sup> to formulate our DTI prediction problem.

Let  $\{d_i\}, 1 \leq i \leq n_d$ , be the set of known drugs and  $\{t_j\}, 1 \leq j \leq n_t$ , be the set of targets, where  $n_d$  and  $n_t$  represent the total numbers of drugs and targets respectively. We use  $X$  to denote observed data, including known DTIs and various similarity scores, such as sequence similarity scores for proteins and chemical similarity scores for drugs. In other words,  $X$  stands for a set of binary indicators representing known drug-target interactions, and positive variables representing observed similarity scores. For each drug  $d_i$ , we construct a CRF on an undirected graph  $G = (V_t, E_t)$ , where  $V_t = \{t_j\}$  is the set of targets and each edge in  $E_t$  represents the similarity between a pair of targets. Let vector  $Y = (y_1, y_2, \dots, y_{n_t})$  denote the prediction, where each  $y_j$  is a binary random variable representing the prediction of target  $t_j$ , that is,  $y_j = 1$  if the predicted interaction between drug  $d_i$  and target  $t_j$  is true, and  $y_j = 0$  otherwise. We call this model the *target-based CRF*. Similarly, for each target  $t_j$ , we construct a CRF on an undirected graph  $G = (V_d, E_d)$ , where  $V_d = \{d_i\}$  is the set of drugs and each edge in  $E_d$  represents the similarity between a pair of drugs. We call the second model the *drug-based CRF*. For the convenience of description, next, we will mainly use the target-based model as an example to illustrate the learning and prediction procedures of our CRF model unless otherwise specified.

For each target-based CRF, we define a joint probability distribution conditioning on observation  $X$ . In the underlying graph, each node represents a target  $t_i$  or its associated binary random variable  $y_i$ , and each edge connecting two nodes represents the dependency between these two nodes. Hereinafter, we will slightly abuse the notation and use terms ‘node’ and ‘random variable’ interchangeably. The undirected graphical model possesses the so-called *conditional independence property*,<sup>37</sup> which states that the conditional distribution of node  $y_i$  is independent of all other nodes given its ‘neighbors’ (i.e., all other nodes that  $y_i$  is connected to). By connecting similar proteins together, we indeed assume that the conditional state of a target depends only on the states of other proteins with high similarity. Details about how to construct edges between targets will be described in Section 3.1.

In a CRF model, the energy of a joint configuration  $Y$  given  $X$  can be defined as follows:

$$E(Y|X) = \sum_i a_i f(y_i|X) + \sum_{i,j} b_{ij} g(y_i, y_j|X) \quad (1)$$

where  $f(y_i|X)$  is a *local node feature function* defined based on the state of  $y_i$ ,  $g(y_i, y_j|X)$  is a *relational edge feature function* defined based on states of both  $y_i$  and  $y_j$ , and  $a_i \geq 0$  and  $b_{ij} \geq 0$  are weight parameters that need to be learned from training data. In our DTI prediction framework, we let all target-based or drug-based CRFs share the same parameters  $a_i$  and  $b_{ij}$ . Then the joint probability density function of  $Y$  given  $X$  can be defined as

$$p(Y|X) = \frac{1}{Z(X)} \exp(-E(Y|X)) \quad (2)$$

where  $Z(X) = \sum_Y \exp(-E(Y|X))$  is the *normalizing constant*, also called *partition function*. We define functions  $f(\cdot)$  and  $g(\cdot)$  as followings:

$$f(y_i|X) = -(y_i - H_{x_i}(y_i))^2 \quad (3)$$

$$g(y_i, y_j|X) = -H_{x_i, x_j}(y_i - y_j)^2 \quad (4)$$

where  $H_{x_i}(y_i)$  represents the observed feature of target  $t_i$ , and  $H_{x_i, x_j}(y_i - y_j)$  represents the relational feature measure of  $y_i$  and  $y_j$  given observation  $X$ . In our framework, we let  $H_{x_i}(y_i)$  be the average number of observed drug interactions for target  $t_i$ , and let  $H_{x_i, x_j}(y_i - y_j)$  be the difference between binary variables  $y_i$  and  $y_j$ . By defining the above two feature functions, we indeed add a penalization when (1) predictions for two connected nodes are different, and (2) the prediction of a given node deviates from its average state. Unlike in Ref. 35, which assumes that all nodes share the same parameter  $a$  and all edges share the same parameter  $b$ , here in our model all weight parameters  $a_i$ ,  $b_{ij}$  are set to be different values for individual nodes and edges. This parameter setting is more flexible to capture information from data and can avoid potential improper assumptions about weight parameters. Our test results (details are not shown in the paper) suggest that this new parameter setting can yield better performance than the original version<sup>35</sup> which chooses a relatively rigid parameter setting.

## 2.2. Parameter Training

In the training process, we aim to learn parameters  $a_i$  and  $b_{ij}$  from training data. We use *stochastic gradient ascent*<sup>38</sup> as an optimization method to maximize the conditional log-likelihood of training data. To simplify the notation, we use vector  $\theta$  to denote parameters  $(a_i, b_{ij})$ , and function vector  $h$  to denote  $(f, g)$ . Then the probability density function in Eq. (2) can be rewritten as

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} \exp(\theta \cdot h) \quad (5)$$

Thus we can derive the following conditional log-likelihood:

$$L_\theta = \sum_{i=1}^{n_t} \log(p(y_i|X)) = \sum_{i=1}^{n_t} [\theta \cdot h(y_i|X) - \log(Z_\theta(X))] \quad (6)$$

Since each component of  $\theta$  is non-negative, we let  $\theta = (\exp(\theta'_1), \dots, \exp(\theta'_{n_t}))$ . For simplicity, we use  $\exp(\theta')$  to represent  $(\exp(\theta'_1), \dots, \exp(\theta'_{n_t}))$ . Then we have

$$L_\theta = \sum_{i=1}^{n_t} \left[ e^{\theta'} \cdot h(y_i|X) - \log(Z_\theta(X)) \right] \quad (7)$$

The gradient in Eq. (7) is

$$\frac{\partial L_\theta}{\partial \theta'} = \theta \cdot \sum_{i=1}^{n_t} [h(y_i|X) - E_\theta(h(Y|X))] \quad (8)$$

where  $E_\theta(h(Y|X))$  is the expectation of  $h(Y|X)$  and  $Y|X$  follows the distribution  $p_\theta$  defined in Eq. (5).

To apply the gradient ascent method, we need to deal with the expectation term in Eq. (8). It is algebraically intractable to directly calculate this expectation, and one possible solution is to employ some simulation techniques such as Markov Chain Monte Carlo (MCMC) to approximate its value. A Gibbs sampling method was used in Ref. 35 to sample a sequence of  $Y$  following the current distribution  $p_\theta$  and then approximate  $E_\theta(h(Y|X))$  by

$$E_\theta(h(Y|X)) = \frac{1}{L} \sum_{i=1}^L h(\tilde{y}_i|X) \quad (9)$$

where  $\{\tilde{y}_i\}$ ,  $1 \leq i \leq L$ , is the sampled sequence, and  $L$  is the total number of sampling iterations. Sampling such sequence often proceeds as follows: We first randomly pick some initial value  $y_0$ , and then sample each variable using the current value according to its conditional distribution. Normally, after some burn-in period, the distribution of  $y_i$  can approximate distribution  $p_\theta$ .

Although Gibbs sampling is a popular method to approximate the expectation, it suffers from heavy computational cost, which is impractical in our case. Here we apply another sampling algorithm, called *contrastive divergence* (CD), which was first proposed in Ref. 27. The CD algorithm has been successfully used to train restricted Boltzmann machines<sup>39</sup> and it can be easily implemented. The basic idea of the CD algorithm is to substitute  $E_\theta(h(Y|X))$  in Eq. (8) by  $E_{p_T}(h(Y|X))$ , where  $p_T$  represents the distribution of data transformed by  $T$  cycles of Gibbs sampling.<sup>27</sup> In practice,  $T$  is often chosen to be one. Although the CD algorithm may lead to biased estimates, the bias is small in general.<sup>40</sup> In practice, the CD algorithm can provide an efficient method to approximate the log-likelihood function.<sup>27,39,40</sup>

### 2.3. Predicting New Drug-Target Interactions

To predict unknown drug-target interactions for a query drug given observation  $X$ , we compute the conditional probability distribution  $p(y_k|y_{-k}, X)$  for each target  $t_k$ , where  $y_{-k}$  denote the all other targets except  $t_k$ . For  $i \neq k$ ,  $y_i = 1$  if target  $y_i$  is known to interact with the query drug, and  $y_i = 0$  otherwise. We then calculate the conditional expectation of  $y_k$  as the prediction score of the interaction between target  $y_k$  and the query drug.



### 3. Results

#### 3.1. *Constructing Conditional Random Field*

In our CRF model, an edge connecting two nodes indicates the relational dependency between them, and we assume that two connected nodes should share high similarity. One natural approach for constructing edges in the underlying graph is to connect two nodes if their similarity score is above a threshold. By choosing different threshold values we should be able to tune the number of edges in the graph. This construction method, which we call the *threshold-based approach*, could yield an unbalanced graph in which some nodes may have much fewer neighbors than others. This situation would make it difficult for inferring the states of those neighbor-free nodes. To avoid this problem, we used another approach to construct the underlying graph. For each node  $t_i$ , let  $N_i$  be the set of top  $K$  nodes that have the highest similarity scores with  $t_i$ , and we connect two nodes  $t_i$  and  $t_j$  if  $t_i \in N_j$  or  $t_j \in N_i$ . We refer to this new approach as the *degree-based approach*, which ensures that the degree of each node in the underlying graph is at least  $K$  and roughly balanced, and thus can prevent the existence of ‘isolated’ nodes. In practice, we should not choose a large value of  $K$  in order to train our CRF model efficiently on a large-scale dataset. Our sensitivity analysis shows that our results did not vary much for different  $K$  values (Supplementary Material S2). We can also combine the above two approaches to get an *integration-based approach* for constructing edges, that is, we connected two nodes mainly based on a similarity score threshold but also added more connections to a node if its degree is less than  $K$ . The comparison results show that different construction approaches did not influence much on prediction performance when choosing  $K \geq 2$  (Supplementary Material S3). In the following analysis, the underlying graph of our CRF model was constructed mainly based on the degree-based approach, unless otherwise specified. We chose  $K = 4$  when a single similarity measure was used and  $K = 2$  when multiple similarity measures were used. This parameter was fixed throughout all our tests.

We tested the following six different approaches in our conditional random field framework:

- Genomic approach (GEN): The target-based CRF was constructed using the sequence similarity measure.
- Functional approach (FUN): The target-based CRF was constructed using the functional similarity measure.
- Integrated Genomic-Functional approach (IGF): The target-based CRF was constructed using the sequence and functional similarity measures simultaneously. In other words, two nodes were connected if they satisfied the sequence or functional similarity criterion.
- Chemical approach (CHEM): The drug-based CRF was constructed using the chemical similarity measure.
- Pharmacological approach (PHAR): The drug-based CRF was constructed using the side-effect similarity measure.
- Integrated Chemical-Pharmacological approach (ICP): The drug-based CRF was constructed using the chemical and side-effect similarity measures simultaneously. In other words, an edge was constructed if it was valid under the chemical or side-effect similarity

measure criterion.

In addition, we investigated the combination of two independent predictions from target-based and drug-based CRFs respectively. For any given drug-target pair, let  $S_d$  denote the prediction score using the drug-based CRF model and  $S_t$  denote the prediction score using the target-based CRF model. Then our final score for this query drug-target pair is

$$S = \alpha S_d + (1 - \alpha) S_t \quad (10)$$

In the current version of our program, we fixed  $\alpha = 0.5$ . By fine-tuning the parameter  $\alpha$ , we may achieve better results than our current tests. Our final approach integrated chemical, pharmacological, genomic and functional data simultaneously:

- Full Integration approach (FI): The final prediction was the simple linear combination of both integrated chemical- pharmacological (ICP) and integrated genomic-functional (IGF) approaches using Eq. (10).

Our program was implemented in Matlab (2010 b) based on the UGM package developed by Mark Schmidt (<http://www.di.ens.fr/~mschmidt/Software/UGM.html>). UGM is a Matlab toolbox that implements various tasks in discrete undirected graphical models with pairwise potentials. We used the default parameters of functions in the UGM package throughout all our tests.

### 3.2. Datasets

To demonstrate the predictive power of our approach, we first tested it on a dataset derived from the KEGG database<sup>41,42</sup> which contains experimentally-verified drug-target interactions. We call this dataset the *first dataset*. All drugs in the first dataset have molecular weight more than 100. In order to obtain pharmacological information we only included those drugs that also have side-effect records in the SIDER database.<sup>43</sup> As a consequence, in total 875 drugs and 249 proteins with 2596 drug-target interactions were obtained in the first dataset.

To compare with other existing approaches, we tested our algorithm on another dataset that has been published in Ref. 24, where all drugs have records in SIDER, JAPIC and AERS. JAPIC and AERS are two public databases about drug side-effects. More details about these two databases can be found in Ref. 24. The data we tested here is slightly different from the original data which contains 359 drugs and 226 proteins with 1188 drug-target interactions. We excluded six proteins that do not have any GO annotation and two drugs that have no interaction with the remaining proteins. Thus the new dataset includes 357 drugs and 220 proteins with 1174 drug-target interactions. We call this new dataset the *second dataset*. Descriptive statistics about the first and second datasets are provided in Supplementary Material S1.

Chemical similarities between drugs were calculated using the graph kernel approach,<sup>44</sup> where chemical structure information of drugs was taken from the KEGG database. Side-effect similarities between drugs were calculated using the same method as in Ref. 24, where pharmacological information was obtained from the SIDER database. Sequence similarities between proteins were computed using local alignment kernel approach.<sup>45</sup> Functional similarities

between proteins were calculated using online software FunSimMat,<sup>46,47</sup> in which functional similarity scores were derived from GO terms annotated with biological process and molecular function. In both datasets that we have tested, most pairs of proteins or drugs were dissimilar. In the first dataset, less than 3% of all drug pairs had chemical similarity score greater than 0.85 (all similarity scores were normalized to 1), and less than 1% of all protein pairs had sequence similarity score greater than 0.85. In the second dataset, less than 2% of all drug pairs had chemical similarity score greater than 0.85, and less than 1% of all protein pairs had sequence similarity score greater than 0.85.

### 3.3. Performance Evaluation

We used the Receiver Operator Characteristic (ROC) curve and the Precision-Recall (PR) curve to evaluate the performance of our algorithm. In addition, we also computed the AUC (area under ROC curve) and AUPR (area under PR curve) scores. In our performance evaluation, true positives were those correctly predicted interactions, while false positives were those predicted interactions that were not present in the tested dataset. For highly-unbalanced data, the PR curve is usually considered to be a better criterion to assess the prediction performance, since it can punish more false positive examples.<sup>16,19,48</sup> Thus our analysis mainly focused on AUPR, although in many cases AUC and AUPR were positively correlated. Our tests were performed mainly using a 10-fold cross-validation procedure. In this procedure, all DTIs were randomly partitioned into 10 equal size subsamples. Each subsample was in turn used as validation data to test our algorithm, and the remaining nine subsamples were used as training data.

Table 1. Prediction results on the first dataset using 10-fold cross-validation. Both AUC and AUPR scores are normalized to 100. The best result is shown in bold.

Approach		Evaluation Criterion	
		AUC	AUPR
Target-based CRF	GEN	97.3	80.7
	FUN	97.7	80.9
	IGF	98.0	83.9
Drug-based CRF	CHEM	96.0	81.5
	PHAR	96.6	79.9
	ICP	98.1	85.9
Full Integration Approach (FI)		<b>99.2</b>	<b>94.9</b>

Table 1 summarizes the test results on the first dataset using the 10-fold cross-validation procedure. Under the target-based CRF framework, integrating both genomic and functional data achieved better performance than other two approaches, with the AUPR score improved by > 3%. When both chemical and pharmacological data were integrated into the drug-based CRF framework, the results outperformed each single-similarity based approach with the AUPR score improved by > 4%. When integrating all available information, the FI approach achieved the best performance with AUPR > 94. Figure 1 shows the AUPR curves for different approaches tested on the first dataset. These results demonstrate that incorporating additional

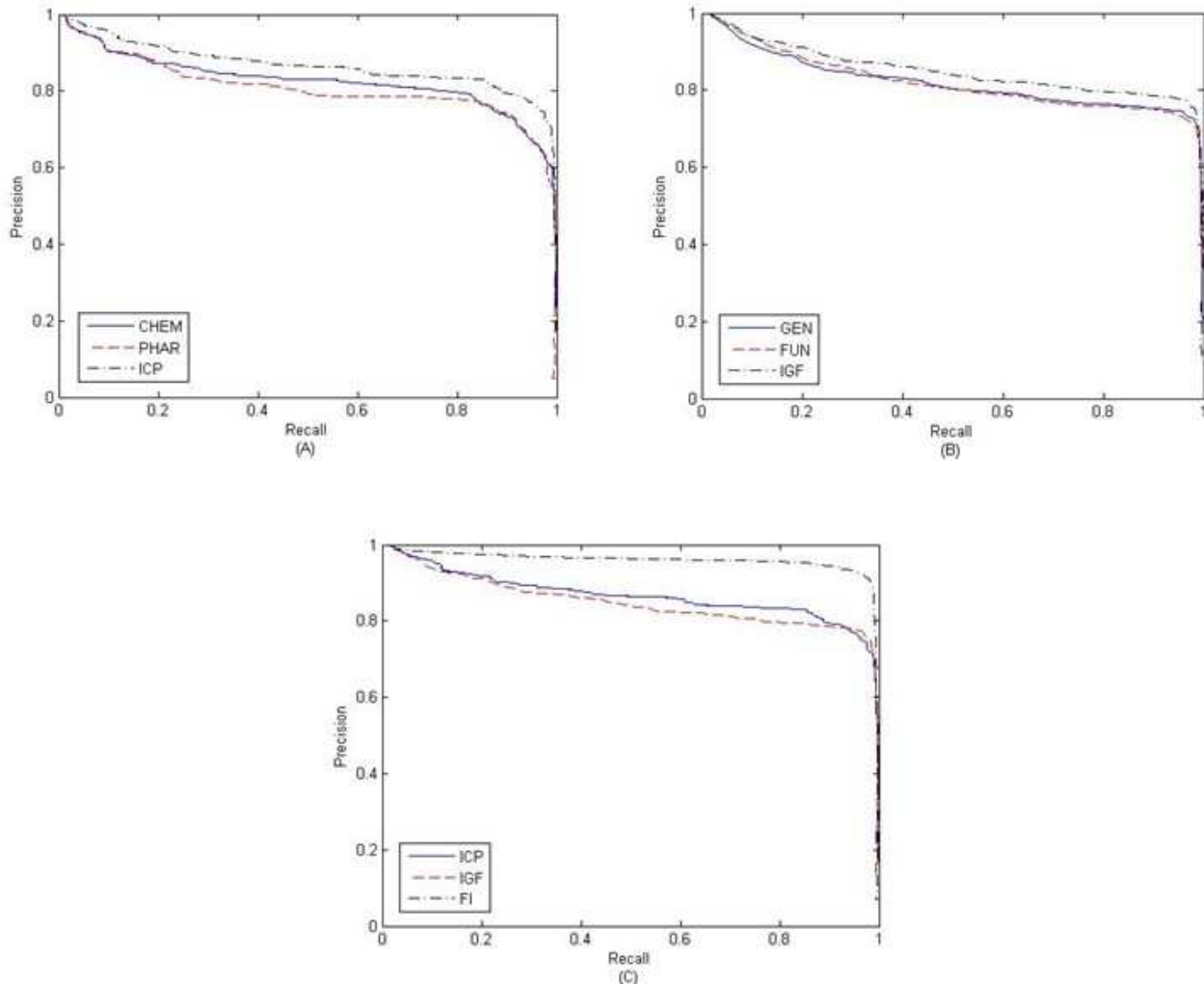


Fig. 1. PR curves for different approaches on the first dataset. (A) PR curves for drug-based CRFs. (B) PR curves for target-based CRFs. (C) PR curves for the FI approach.

information about drugs and proteins can further improve prediction accuracy. To check the robustness of our model, we also performed a 5-fold cross-validation test, and only observed a slight decrease in AUC and AUPR values (Supplementary Material S4).

### 3.4. Comparison Results

To compare with other existing approaches, we tested our algorithm on the second dataset, i.e., the benchmark dataset published in Ref. 24. Here we mainly compared our approach with the *pairwise kernel regression* (PKR) method proposed in Ref. 24, which claimed that PKR outperformed many other state-of-the-art methods on the same data. As in Ref. 24, we also tested seven different approaches, including AERS-freq-based pharmacogenomic approach (AERS-freq), AERS-bit-based pharmacogenomic approach (AERS-bit), SIDER-based pharmacogenomic approach (SIDER), JAPIC-based pharmacogenomic approach (JAPIC), chemogenomic approach (CHEM), integrated pharmacogenomic approach (INTEG-P) and

integrated pharmaco-chemogenomic approach (INTEG-PC). These different methods, as suggested by their names, are defined mainly based on input data, and more details about them can be found in Ref. 24 or Supplementary Material S5 of this paper. In addition, we tested an additional approach that combines chemical, side-effect, sequence and functional data together. This approach was not included in Ref. 24 and we referred to it as ‘INTEG-ALL’. Table 2 shows the comparison results between our conditional random field (CRF) model and the pairwise kernel regression (PKR) model.

As shown in table 2, our method outperformed the PKR model over all different tests. In particular, our approach can improve the AUPR score by up to 10.5 when only SIDER-based information was used. Furthermore, the results produced by CRF were not as sensitive to different input data as those produced by PKR. For example, the AUPR score of PKR based on JAPIC was about 10% larger than that based on SIDER, whereas the test of our algorithm on SIDER-based data can still yield decent performance. These comparison results indicate that our method is more robust to input data than PKR, and may have a better capacity to handle noise in data.<sup>a</sup>

Table 2. The comparison results between our CRF and PKR methods. The second dataset was tested in our CRF model using 3-fold cross-validation. The results for PKR were taken from Ref. 24 in which pair-wise cross-validation corresponds to our 3-fold cross-validation test here. Note that the INTEG-ALL approach was absent in Ref. 24. The best score is shown in bold.

Approach	AUPR	
	CRF	PKR
AERS-freq	85.7	80.6
AERS-bit	85.4	81.3
SIDER	87.3	76.8
JAPIC	91.2	87.7
CHEM	87.7	79.7
INTEG-P	90.7	87.4
INTEG-PC	90.4	88.5
INTEG-ALL	<b>91.5</b>	\

#### 4. Conclusion

In this article, we introduced functional data into DTI prediction and developed a probabilistic graphical model to predict new drug-target interactions using known drug-target interactions and various similarity scores for both drugs and targets. Our model can integrate chemical, pharmacological, genomic and functional data systematically, and predict new DTI interactions with high accuracy. We demonstrated that incorporating functional information of targets can further improve prediction performance.

Currently, our algorithm uses a simple linear combination of independent predictions from

<sup>a</sup>Although our dataset were slightly differently from the original data tested in the PKR model (six proteins and two drugs were excluded from the original dataset), the tiny difference between two datasets should not change the conclusions that we draw here.

drug-based and target-based CRFs respectively. In the future, we will extend our model into a more sophisticated framework that can better integrate both drug-based and target-based CRF models. In addition, we will incorporate other data such as drug-drug interaction (DDI) and protein-protein interaction (PPI) information into DTI prediction. We hope that by incorporating these additional information our model can reveal mechanism of drug action to a greater extent. Currently we only evaluated our approach based on benchmark data. We will explore the practical applications of our prediction algorithm, e.g., identifying novel drug-target interactions for drug repositioning.

## 5. Acknowledgements

This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61061130540. We thank the anonymous reviewers for their helpful comments.

## References

1. J. T. Dudley, T. Deshpande and A. J. Butte, *Brief Bioinform* **12**, 303 (Jul 2011).
2. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte, *Sci Transl Med* **3**, p. 96ra76 (Aug 2011).
3. Y. A. Lussier and J. L. Chen, *Sci Transl Med* **3**, p. 96ps35 (Aug 2011).
4. L. Xie, L. Xie, S. L. Kinnings and P. E. Bourne, *Annu Rev Pharmacol Toxicol* **52**, 361 (2012).
5. S. Ekins, A. J. Williams, M. D. Krasowski and J. S. Freundlich, *Drug discovery today* **16**, 298 (2011).
6. J. Blatt and S. J. Corey, *Drug discovery today* **18**, 4 (2012).
7. A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg and E. S. Huang, *Nature biotechnology* **25**, 71 (2007).
8. 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).
9. S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie and P. E. Bourne, *J Chem Inf Model* **51**, 408 (Feb 2011).
10. B. R. Donald, *Algorithms in structural molecular biology* (The MIT Press, 2011).
11. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nature biotechnology* **25**, 197 (2007).
12. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran *et al.*, *Nature* **462**, 175 (2009).
13. H. Yabuuchi, S. Nijima, H. Takematsu, T. Ida, T. Hirokawa, T. Hara, T. Ogawa, Y. Minowa, G. Tsujimoto and Y. Okuno, *Mol Syst Biol* **7**, p. 472 (Mar 2011).
14. F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang and Y. Tang, *PLoS Computational Biology* **8**, p. e1002503 (2012).
15. X. Chen, M.-X. Liu and G.-Y. Yan, *Molecular BioSystems* **8**, 1970 (2012).
16. T. van Laarhoven, S. B. Nabuurs and E. Marchiori, *Bioinformatics* **27**, 3036 (2011).
17. J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li and J. Zheng, *Bioinformatics* **29**, 238 (2013).
18. Y. Shi, X. Zhang, X. Liao, G. Lin and D. Schuurmans, *Pac Symp Biocomput* **18**, 41 (2013).
19. Y. Wang and J. Zeng, *Bioinformatics* **29**, i126 (2013).
20. W. Wang, S. Yang and J. Li, *Pac Symp Biocomput* **18**, 53 (2013).
21. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics* **24**, i232 (2008).

22. M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science* **321**, 263 (2008).
23. Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics* **26**, i246 (2010).
24. M. Takarabe, M. Kotera, Y. Nishimura, S. Goto and Y. Yamanishi, *Bioinformatics* **28**, i611 (2012).
25. A. Schlicker, F. S. Domingues, J. Rahnenführer and T. Lengauer, *BMC bioinformatics* **7**, p. 302 (2006).
26. S. Zhao and S. Li, *PloS one* **5**, p. e11764 (2010).
27. G. E. Hinton, *Neural computation* **14**, 1771 (2002).
28. D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques* (The MIT Press, 2009).
29. J. Zeng, P. Zhou and B. R. Donald, A markov random field framework for protein side-chain resonance assignment, in *Research in Computational Molecular Biology*, 2010.
30. J. Zeng, P. Zhou and B. R. Donald, *Journal of biomolecular NMR* **50**, 371 (2011).
31. J. Lafferty, A. McCallum and F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, 2001.
32. F. Sha and F. Pereira, Shallow parsing with conditional random fields, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003.
33. B. Settles, Biomedical named entity recognition using conditional random fields and rich feature sets, in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004.
34. T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang and H. Li, *Global ranking of documents using continuous conditional random fields*, tech. rep., Technical Report MSR-TR-2008-156, Microsoft Corporation (2008).
35. X. Xin, I. King, H. Deng and M. R. Lyu, A social recommendation framework based on multi-scale continuous conditional random fields, in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009.
36. Z. Wang and J. Xu, *Bioinformatics* **27**, i102 (Jul 2011).
37. C. M. Bishop *et al.*, *Pattern recognition and machine learning* (Springer New York, 2006).
38. D. P. Bertsekas, A. Nedić and A. E. Ozdaglar, *Convex analysis and optimization* (Athena Scientific Belmont, 2003).
39. R. Salakhutdinov, A. Mnih and G. Hinton, Restricted boltzmann machines for collaborative filtering, in *Proceedings of the 24th international conference on Machine learning*, 2007.
40. G. E. H. M. A. Carreira-Perpignan, On contrastive divergence learning, in *Artificial Intelligence and Statistics*, 2005.
41. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, *Nucleic acids research* **34**, D354 (2006).
42. M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu *et al.*, *Nucleic acids research* **36**, D480 (2008).
43. M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen and P. Bork, *Molecular systems biology* **6** (2010).
44. P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret and J.-P. Vert, *Journal of chemical information and modeling* **45**, 939 (2005).
45. H. Saigo, J.-P. Vert, N. Ueda and T. Akutsu, *Bioinformatics* **20**, 1682 (2004).
46. A. Schlicker and M. Albrecht, *Nucleic acids research* **36**, D434 (2008).
47. A. Schlicker and M. Albrecht, *Nucleic acids research* **38**, D244 (2010).
48. K. Bleakley and Y. Yamanishi, *Bioinformatics* **25**, 2397 (2009).

# PREDICTION OF OFF-TARGET DRUG EFFECTS THROUGH DATA FUSION

EMMANUEL R. YERA, ANN E. CLEVES, and AJAY N. JAIN<sup>†</sup>

*Bioengineering and Therapeutic Sciences, University of California, San Francisco,  
San Francisco, CA 94143, USA*

<sup>†</sup>*E-mail: [ajain@jainlab.org](mailto:ajain@jainlab.org)  
[www.jainlab.org](http://www.jainlab.org)*

We present a probabilistic data fusion framework that combines multiple computational approaches for drawing relationships between drugs and targets. The approach has special relevance to identifying surprising unintended biological targets of drugs. Comparisons between molecules are made based on 2D topological structural considerations, based on 3D surface characteristics, and based on English descriptions of clinical effects. Similarity computations within each modality were transformed into probability scores. Given a new molecule along with a *set* of molecules sharing some biological effect, a single score based on comparison to the known set is produced, reflecting either 2D similarity, 3D similarity, clinical effects similarity or their combination. The methods were validated within a curated structural pharmacology database (SPDB) and further tested by blind application to data derived from the ChEMBL database. For prediction of off-target effects, 3D-similarity performed best as a single modality, but combining all methods produced performance gains. Striking examples of structurally surprising off-target predictions are presented.

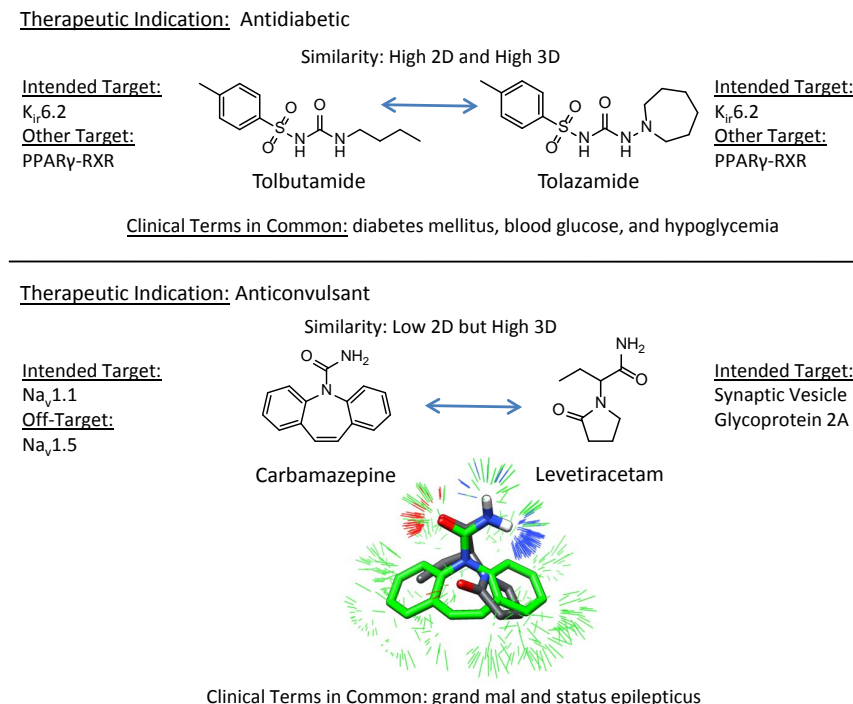
*Keywords:* Molecular similarity; Surflex-Sim; Patient Package Inserts; Off-Target Predictions.

## 1. Introduction

In prior work, we introduced a methodological approach for data fusion which was used to predict the protein targets of small molecules based on molecular similarity.<sup>1</sup> Given a test molecule and a set of small molecules with a known shared biological effect, the method produces a score corresponding to the likelihood that the test molecule will share the same activity. We showed that for predicting primary targets (i.e. targets modulating intended therapeutic effects) the performance advantage of a 3D similarity method over a 2D method was relatively small, due to the dominating effects of human 2D bias in drug design (i.e. “me-too” drugs).<sup>1,2</sup> However, for predicting *secondary* targets (i.e. sources of side-effects) 3D similarity was much more effective than 2D topological comparisons. We also showed that clinical effects of drugs could be used as a surrogate for biochemical characterization,<sup>1</sup> making use of common side effects of muscarinic antagonism as markers for the biochemical protein-ligand effect. It was possible using 3D chemical similarity to achieve strong separation of likely muscarinic modulators from those with no evidence of such effects.

In the current work, we expand the analysis to a much larger set of small molecule drugs, again making use of 2D and 3D chemical similarity computations. Additionally, computations involving structural similarity are augmented with clinical effects similarity, made possible by *automating* the extraction and weighting of relevant textual terms from drug package inserts. The top row of Figure 1 shows two highly similar first generation sulfonylureas, tolbutamide and tolazamide, each having highly similar pharmacological effects,<sup>3</sup> with their therapeutic benefits deriving from identical mechanisms.<sup>4</sup> Clinical effects similarity coincides here with



**Figure 1.**

Relationship between molecular similarity methods, the primary modulation, and clinical effects. The top row shows two antidiabetic drugs, tolbutamide (top left) and tolazamide (top right), which are very structurally similar, interact with similar proteins, and have similar clinical effects. The bottom row shows two anti-epileptic drugs, carbamazepine (bottom left) and levetiracetam (bottom right), that have different primary targets but similar clinical effects and 3D molecular similarity. Surflex-Sim's 3D overlay is shown at the bottom where carbamazepine is colored by green carbons and levetiracetam is colored by blue carbons. Green sticks correspond to regions of significant hydrophobic similarity and blue/red sticks correspond to regions of significant polar similarity.

Fig. 1. Relationship between molecular similarity methods, the primary modulation, and clinical effects. The top row shows two antidiabetic drugs, tolbutamide (top left) and tolazamide (top right), which are very structurally similar, interact with similar proteins, and have similar clinical effects. The bottom row shows two anti-epileptic drugs, carbamazepine (bottom left) and levetiracetam (bottom right), that have different primary targets but similar clinical effects and 3D molecular similarity. Surflex-Sim's 3D overlay is shown at the bottom where carbamazepine is colored by green carbons and levetiracetam is colored by blue carbons. Green sticks correspond to regions of significant hydrophobic similarity and blue/red sticks correspond to regions of significant polar similarity.

high structural 2D and 3D similarity. Next, consider the two structurally dissimilar anticonvulsants on the levetiracetam of Figure 1. Carbamazepine was one of the first anticonvulsants (approved in 1968), and its therapeutic benefit is attributed to stabilizing the inactivated state of voltage-gated sodium channels (Nav1.1).<sup>5</sup> Levetiracetam is a newer anticonvulsant, believed to act through interaction with synaptic vesicle glycoprotein 2A (SV2A).<sup>6</sup> As expected, the two package inserts have clinical effect terms in common due to shared indications. Given the high 3D structural similarity, our expectation is that these drugs do in fact share some molecular targets, as will be discussed later.

The present study establishes a computational method to draw relationships between drugs based on the clinical effects present in Patient Package Inserts (PPI), whose utility for predicting drug target interactions has been shown previously.<sup>7</sup> The present study makes three primary contributions. First, we introduce a method to extract and weight medically relevant terms from English clinical effects information. Second, we show that drug similarity computed from package inserts is *directly correlated* with drug similarity computed by molecular structure comparison. Third, we established that the combination of 2D, 3D, and PPI similarity yielded better off-target predictive performance over any single similarity computation. Recovery of roughly 40–50% of off-target annotations was possible with false positive rates of about 1–3%. The approach is generalizable to other computational modalities (e.g. docking of ligands to protein structures), and it is our hope that broad application of the methods will aid in identifying unexpected interactions between drugs and biological targets.

## 2. Methods and Data

The following describes the molecular data sets, computational methods, and specific computational procedures (see <http://www.jainlab.org> for additional details on software, data, and protocols).

### 2.1. *Molecular Data Sets*

In the present study two molecular data sets are used. The Structural Pharmacology Database (SPDB) is a deeply curated drug target database that is used as the basis to make predictions. A set of drug target annotations from ChEMBL that were not annotated in our database were used as a blind test set.

The details of the SPDB and its relationship to other databases has been extensively described elsewhere.<sup>1,2,8</sup> It has two features that are particularly important for the present study. First, “targets” are *specific* binding sites on proteins or protein complexes. This is a critical distinction in order to make inferences about small molecule activity based on structural similarity. Second, primary targets (those that are believed to be therapeutically beneficial) are distinguished from secondary targets (which mediate pharmacologically relevant off-target effects). By making this distinction, it is possible to explicitly quantify performance of methods for prediction of *surprising* effects. Of the roughly 1000 drugs within the SPDB, 602 met our criteria for inclusion based on PPI information (see below). Of the 257 primary and secondary targets of these 602 drugs, 91 had at least 5 annotated drugs and formed the basis of cross-validation experiments. These 91 targets were comprised of 83 human proteins, including 28 aminergic GPCRs, 19 ligand and voltage gated ion channels, 13 human enzymes, 7 nucleotide and short peptide GPCRs, 5 tyrosine kinases, 5 steroid receptors, 3 reuptake transporters, 2 ion transporters, and 1 transcription factor. The remaining 8 targets were bacterial, fungal, and viral proteins. To test the methodology, we employed ChEMBL version 14, which curates linkages between chemicals and biological targets.<sup>9</sup> For each of the 602 drugs, corresponding ChEMBL compounds were identified based on direct structural equivalence. Equating the 91 SPDB target binding sites to ChEMBL bioactivities was done manually, yielding 65 corresponding ChEMBL targets. Significant bioactivity was defined as  $K_d$ ,  $K_i$ , or  $IC_{50}$  values less than or equal to  $1\mu M$ . There were 380 drug-target interactions present in ChEMBL that were missing from the SPDB matrix of 602 drugs and 91 targets. This set served as a blind test set and will be referred to as the ChEMBL set in what follows.

### 2.2. *Patient Package Insert Similarity*

We employed the well established vector space information retrieval approach<sup>10,11</sup> to model patient package inserts (PPIs). Text documents are modeled as vectors in high dimensional space where each dimension corresponds to a term with an associated weight. Coincidence of terms with high weight leads to high computed similarity between documents. The process to transform PPIs into weighted term vectors requires four steps. First, relevant sections are extracted, including: Indication, Contraindications, Precautions, Adverse Reactions, Drug Interactions, and Clinical Pharmacology. Second, term lists (up to five

words each) are generated, with punctuation and short words like prepositions and articles removed. Third, to eliminate artifactual terms and enhance relevance, terms are identified that are part of two controlled vocabularies: Medical Subject Headings (MeSH, <http://www.ncbi.nlm.nih.gov/mesh>) and the low-level Medical Dictionary for Regulatory Activities (MedDRA, <http://www.meddra.org>). Last, term weights are assigned based on information richness (e.g. “generalized seizures” > “seizures”). Word frequencies from the Google Web 1T 5-gram Corpus (<http://www ldc.upenn.edu/Catalog/index.jsp>, catalog number LDC2006T13) were used to compute term weights, with rare terms producing higher scores than common ones. For example, “seizures” produced a log odds weighting of 4.74, but the more specific term “generalized seizures” yielded 6.89. The final output for each drug is a vector composed of 6,591 term weights (the weight of the term if present and zero otherwise). From the PPI for carbamazepine, the Indication Section includes: “patients with the following seizure types: partial seizures with complex symptomatology (psychomotor, temporal lobe).” The unfiltered bigrams include both sensible ones such as “partial seizures” and useless ones such as “patients with” with the filtering process eliminating the latter. For carbamazepine, the two most heavily weighted terms were “failure liver” (8.83) and “syncope and collapse” (8.62). The term “partial seizures” scored 6.37, with many related terms (e.g. “grand mal”) scoring similarly.

$$PPI_{Similarity}(A, B) = \frac{\sum_{i=1}^n A_i * B_i}{\sqrt{\sum_{i=1}^n A_i^2} * \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

Comparison of a pair of drug PPI vectors is quantified using the cosine similarity metric (Eq. 1). The metric has a range of 0–1, but its units are both arbitrary and counterintuitive. To employ such values in our data fusion framework, the raw similarity scores were normalized to *p*-values by generating a distribution of PPI similarity scores for *unrelated* molecule pairs. The unrelated pairs were identified based on having low 2D and low 3D similarity, quantified as described below with pairwise *p*-value comparisons  $\geq 0.5$  (we have previously shown that structurally unrelated drug pairs very infrequently share targets<sup>1</sup>). So, given a PPI similarity score *S* between a drug pair, the *p*-value is simply the proportion of occurrences of *S* or greater in the background set. For example, the raw PPI similarity between carbamazepine and levetiracetam was 0.286 (see Figure 1), and this corresponded to a *p*-value of 0.044. The most heavily weighted terms in the comparison included the following: pancytopenia (6.6), cytochrome p450 (6.6), grand mal (6.5), antiepileptic drugs (6.5), and partial seizures (6.4).

### 2.3. Target Prediction using Patient Package Insert Similarity

We have previously reported a framework for data fusion which allows for the integration of similarity scores into a single value.<sup>1</sup> Briefly, given a molecule *A* and a set of molecules with a shared biological effect, *B<sub>n</sub>*, the similarity between molecule *A* and each molecule *B<sub>i</sub>* is computed. The similarity scores are normalized to *p*-values as detailed above by assessing score magnitude against score from a random background set. The multinomial distribution is then used to compute the likelihood, *M*, of observing the set of *p*-values and of the converse probabilities, *M*\*. The log-odds score *L* is then computed by taking the log of the ratio of *M*

and  $M^*$  and inverting the sign. A detailed discussion of the computation and corresponding 2D and 3D similarity example can be found in the original publication.<sup>1</sup> An attractive feature of our methodology is that it is able to integrate the results of different similarity computations into a single value. For example, the log-odds calculation for tolazamide interacting with PPAR $\gamma$ -RXR yields single-modality values of 11.35 for PPI, 7.57 for 3D, and 5.49 for 2D. Combining the similarity methods gives a stronger prediction compared to using any single method alone with 3D+2D+PPI log-odds = 23.43.

#### 2.4. *Similarity and p-value Computation with Surflex-Sim*

The Surflex-Sim 3D molecular similarity method and its use for virtual screening and off-target prediction has been extensively described in multiple publications.<sup>2,8,12,13</sup> Briefly, given two molecules in specific poses, a value from 0 to 1 is computed that reflects the degree to which their molecular surfaces are congruent with respect to both shape and polarity. The function is based on the differences in distances from observer points surrounding the molecules to the closest points on their surfaces, including both the closest hydrophobic surface points and the closest polar surface points. So, two molecules that may have very different underlying chemical scaffolds may exhibit nearly identical surfaces to the observer points. These points are analogous to a protein binding pocket, which also “observes” ligands from the outside. Additional details regarding the theory and underlying algorithmic details can be found in the previously published work. In order to produce a log-odds value for a molecule against a list of molecules with a shared annotation, 3D similarity values must be computed against each annotated molecule, and these values must then be transformed into probabilities. Given the particulars of the conformational sampling density, 3D similarity optimization thoroughness, and empirical conversion of raw scores to  $p$ -values, the overall process required many hours for each comparison of one molecule to a typical set of annotated molecules.

In the current work, two improvements were made to support large-scale application of the methods. First, a new mode of pose optimization was developed in which diverse conformations of molecules are pre-generated prior to molecular comparison. Using this new mode, the optimal pose for one molecule onto a specific pose of another can be done quickly enough to process roughly 2 million drug-like molecules per day on a single computing core (compared with roughly 10,000 previously). Second, rather than using explicit computation of 1000 background similarity values for each molecule (as previously), we made use of the observation that these distributions were essentially always normally distributed. Given a molecule pair, only the particular mean and standard deviation for each need be estimated in order to derive a  $p$ -value rather than making use of the full empirical computation. Estimation of the distributional parameters was accomplished using simple linear regression models that made use of “molecular imprints” for each molecule.<sup>8</sup> A molecular imprint is a vector of similarity values for a particular molecule against a fixed basis set of molecules (one pose each). Such vectors have precedent in predicting many molecular properties,<sup>14,15</sup> and the conformational pre-search procedure was augmented to produce standard molecular imprints. So, given two pre-searched molecules, their mutual maximal 3D similarity can be rapidly calculated, and the  $p$ -value conversion is immediately derived from the estimated distributional parameters

for each molecule. Taken together, the two improvements allow for typical 3D log-odds computations to be made in a few minutes for a given molecule against a target characterized by twenty known ligands. To test the accuracy of the faster method, we recomputed the  $p$ -values and log-odds values from our previous work. An all-by-all similarity of the 358 drugs from the original study yielded a Pearson’s correlation of 0.947 and Kendall’s tau of 0.814, both highly statistically significant. The full log-odds computation of 358 drugs against 44 targets yielded a Pearson’s correlation of 0.955 and Kendall’s Tau of 0.761 (again highly statistically significant).

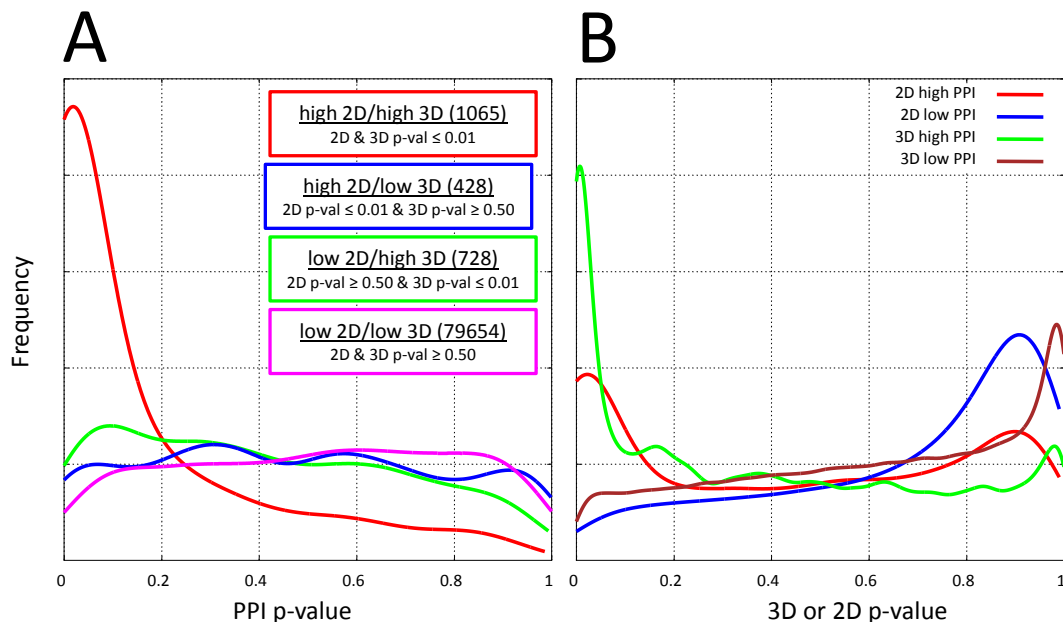
For 2D molecular similarity computations, which make purely topological comparisons between molecules, we employed the previously described GSIM-2D method.<sup>1,2</sup> This method is sufficiently efficient that empirical conversion of raw scores into  $p$ -values is possible, as we have previously described.<sup>1</sup> For this method to yield high similarity, two molecules must be roughly the same size and contain similar subgraph compositions, especially for those subgraphs rooted at heteroatoms.

### 3. Results and Discussion

#### 3.1. *Relationship between Structural Novelty and Clinical Effects*

Previously, we quantified the effect of me-too drugs by showing that drug pairs with high 2D and high 3D similarity had four times more likelihood of having identical primary and secondary targets than drugs pairs where one was structurally novel.<sup>1</sup> Here, this analysis has been extended to clinical effects by making use of the lexical similarity of package inserts. Both to establish the relevance of the PPI similarity metric and to quantify the degree to which structural novelty is related to changes in clinical effects, we computed the pairwise 2D, 3D, and PPI similarity of all 602 drugs. The drug pairs were separated into four categories based on chemical structural similarity: high 2D and 3D similarity, low 2D but high 3D, high 2D but low 3D, and low 2D and 3D. High similarity included pairs with  $p$ -values  $\leq 0.01$  and low similarity were those with  $p$ -values  $\geq 0.5$ .

Figure 2A shows the histogram of the PPI  $p$ -value distributions for each of the four structural categories. It is clear that the “me-too” drug distribution (red line, drug pairs with high 2D and high 3D similarity) is different than the others. Toward the left side of the plot, where clinical effects similarity was high (PPI  $p$ -values  $\leq 0.05$ ), a large fraction of the me-too drug pairs had highly similar clinical effects. Structurally novel drug pairs (high 3D but low 2D similarity, green line) exhibited a significantly smaller fraction with highly concordant clinical effects but still showed some relationship between structural similarity and therapeutic profile. The high 2D and low 3D pairs had little signal (blue), and only a very small portion of structurally dissimilar drug pairs (low 2D and low 3D, magenta) shared clinically similar effects. Clearly, drug pairs with very high structural similarity (both by 2D and 3D methods) were much more likely to have closely shared clinical effects than molecule pairs of any other category, even those sharing high 3D similarity but low 2D similarity. The converse observations paralleled these observations. Figure 2B shows the corresponding histograms of 3D and 2D  $p$ -value distributions where molecule pair segregation was made based on clinical effect similarity. The 2D and 3D similarity  $p$ -value distributions for drug pairs with high PPI



**Figure 7.** Relationship between structural similarity and clinical effects similarity.

**Quantifying the effect of me-too drugs based on PPI similarity.** Panel A shows the PPI p-value distribution of drug pairs that were segregated based on 2D and 3D p-values into the four bins shown above (number of pairs per quadrant are shown in parentheses). Drug pairs with high 2D and high 3D have a higher likelihood of having significant phenotypic effects than molecules with low 2D but high 3D. Panel B shows the 3D p-values distribution of drug pairs that were segregated based on high and low PPI p-values. `len(high_ppi): 3968`  
`len(low_ppi): 88539`

### 3.2. Internal SPDB Validation: Off Target Effects

An attractive aspect of the log-odds framework is that it allows us to combine different types of similarity computations into a single value. For each of the 602 drugs in our dataset, we computed the 2D, 3D, PPI, and combination log-odds scores of interacting with each of the 91 targets that had at least 5 drugs as ligands in the SPDB. In each case, any self/self comparisons were omitted from the calculations, making this exercise a leave-one-out cross-validation of the log-odds predictive methodology. The three methods were used independently and in combination to predict the log-odds of known primary and secondary target interactions. As we observed in our previous study, primary target predictions were dominated by the presence of me-too drugs, limiting the differences between any methods (data not shown). However, for prediction of secondary targets, i.e. those that mediate side-effects, significant differences appeared. Table 1 summarizes the true-positive rates observed for difference log-odds computations for secondary target prediction at different score thresholds.

Yera/Cleves/Jain: Ligand Structure/Function 26

Table 1. SPDB Secondary Target Performance

Log-Odds	3D	2D	PPI	3D+2D	3D+PPI	2D+PPI	3D+2D+PPI
0	97	90	96	95	98	97	97
10	43	7	14	55	61	33	64
20	16	0	0	23	26	1	38

For all single methods and combinations of methods, the information present in the annotated drugs yielded positive information, evidenced by high true-positive rates at a log-odds threshold of 0. However, substantial differences among the methods appeared as higher log-odds thresholds were considered. At a threshold of 10, the 3D similarity approach showed a much higher retrieval rate than either of the other two single-mode methods. All *combinations* of methods showed synergy, with the most effective retrieval occurring with a combination of all three similarity methods to produce a single log-odds score. Roughly 60% of the true secondary target annotations could be recovered using the log-odds score from 3D+2D+PPI similarity computations. Note, however, that true positive rates without the context of false positive rates can be very misleading. The issue of estimating false positive rates is not straightforward though. In our SPDB, a missing annotation between a drug and a target does not mean that the interaction *does not* occur. Authentic interactions within our 602 drug/91 target set may have been published after our curation or have yet to be biochemically characterized. Nonetheless, we expect that the large majority of unannotated interactions, in fact, represent true negative data. So, as a surrogate for a measurement of false positive rates for our similarity methods, we determined the number of drug/target predictions for interactions that were unannotated. At log-odds thresholds of 5, 10, and 20, predictions for non-existent SPDB annotations for both 3D similarity alone and 3D+2D+PPI were 3%, 1%, and 0.2%. These are *upper limits* of false positive predictions. As will be described below, the false positive rate was actually lower since many of the new predictions were validated as true by incorporating annotations from the ChEMBL database.

### 3.3. Prediction of New Drug-Target Pairs within ChEMBL

As discussed above, a missing annotation within the SPDB between a drug and a target does *not* necessarily mean that the interaction does not occur. For example, the drugs orphenadrine and mesoridazine showed high 3D log-odds against the muscarinic receptor but the interactions had been unannotated in the SPDB. Careful inspection of the literature revealed that the drugs were known to antagonize muscarinic receptors.<sup>1</sup> Therefore, drug target annotations that are known but missing from our SPDB can serve as a blind set to test our methodology. To supplement annotations within the SPDB with a blind set for methodological testing, we searched ChEMBL and found 380 biochemically characterized drug/target interactions not present in the SPDB. We then investigated how well the methodology could identify the new ChEMBL annotations based only upon information within in the SPDB as the basis to compute the log-odds.

Table 2 shows the proportions correctly predicted at various log-odds using different methods and combinations. In general, the trends observed for the SPDB leave-one-out experiments were borne out. Among individual methods, 3D similarity strongly outperformed 2D- or PPI-

Table 2. ChEMBL Prediction Performance

Log-Odds	3D	2D	PPI	3D+2D	3D+PPI	2D+PPI	3D+2D+PPI
5	43	14	13	42	41	19	41
10	16	3	3	20	18	8	22
20	2	0	0	3	1	0	4

based similarity, with the latter two having similar performance. However, the combination of the three methods, overall, yielded better performance than 3D alone. At log-odds thresholds of 10 and 20, using the full combination of methods, the percentage of recovered annotations within the SPDB test set was 22% and 4%, respectively. This compared with 16% and 2% using 3D similarity alone, and 3% and 0% using either 2D or PPI similarity alone. The enrichment ratios for the combination approach, using the upper-bound false positive rates discussed above, corresponded to 22-fold and 40-fold, respectively, at log-odds thresholds of 10 and 20.

Figure 3 shows a typical example of a drug/target interaction not annotated in the SPDB where the combination similarity approach confidently identified a pharmacologically relevant target. Sibutramine is an anorexic annotated in the SPDB as a ligand of the serotonin and norepinephrine reuptake transporters. However, it has been shown that sibutramine also interacts with the dopamine reuptake transporter and that this interaction contributes to the therapeutic benefit (indicated in the Meridia package insert, <http://www.rxabbott.com/pdf/meridia.pdf>). Computing the similarity between sibutramine and 11 other dopamine reuptake transporter inhibitors (two are shown in Figure Figure 3), the log-odds were 2.3, 4.2, and 6.9 using 2D, 3D, and PPI, respectively. These predictions were strengthened by combining all three methods, with corresponding log-odds of 9.4. The pairwise PPI similarities between sibutramine and bupropion and nefazodone are highly significant as are the individual 3D similarities. Clinical effects can be sufficient to infer off-targets,

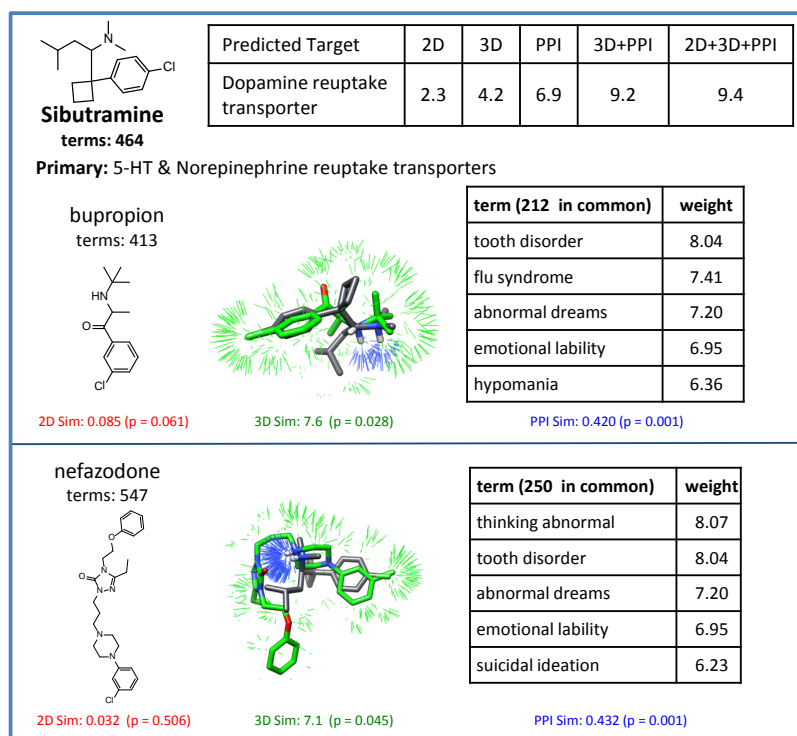


Fig. 3. ChEMBL example showing that combination similarity effectively predicts a drug target interaction not covered within the SPDB. Shown are the 2D structures, 3D overlays, and common clinical terms between sibutramine and two dopamine reuptake transporter inhibitors, bupropion and nefazodone.



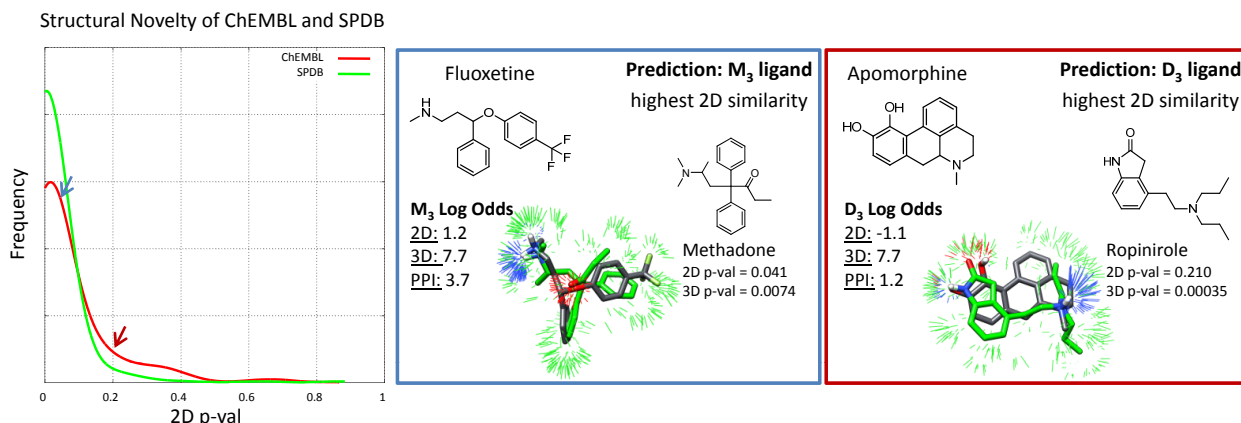


Fig. 4. A near-neighbor analysis for each of 602 drugs (SPDB in green, ChEMBL in red) based on target annotations from the SPDB.

but combining similarity methods generally adds confidence to predictions.

Note, however, that the numerical performance on the ChEMBL set was lower than for the SPDB set in terms of pure true positive recovery rates (see Tables 1 and 2). This stemmed from an increase in structural diversity for molecules within ChEMBL compared to those molecules within the SPDB for the target identified by the ChEMBL annotation. To quantify structural novelty, we performed a nearest-neighbor analysis. For each drug within ChEMBL, the most similar 2D representative from the SPDB was identified (based on  $p$ -value) from within the collection of drugs having the same target annotation. An analogous leave-one-out computation was performed for each drug target annotation within the SPDB. Figure 4 shows a histogram of the distributions of  $p$ -values for the ChEMBL (red line) and SPDB (green line) sets. Within the SPDB set, there were substantially more cases with extremely low  $p$ -values than for the ChEMBL set. The nearest structural neighbor for each ChEMBL test molecule were generally more divergent. Two examples are highlighted from the ChEMBL set where the nearest neighbor had poor 2D  $p$ -values relative to the much more significant 3D  $p$ -values which provided support for high log-odds scores.

Fluoxetine (blue box) is a selective serotonin reuptake inhibitor which mediates its therapeutic benefit through inhibition of the 5-HT reuptake transporter. The ChEMBL data indicated that fluoxetine also interacts with the muscarinic M<sub>3</sub> receptor. The nearest-neighbor molecule sharing this annotation was methadone (2D *p*-value = 0.041). Considering all of the muscarinic M<sub>3</sub> receptor ligands (38 total), the 2D, 3D, and PPI log-odds were 1.2, 7.7, and 3.7 respectively. Combining all of the methods gave a score of 8.2.

Apomorphine (red box) is indicated to treat Parkinson’s disease and its therapeutic benefit is thought to be primarily due to activating dopamine D<sub>2</sub> receptors. However, apomorphine was indicated within ChEMBL to also interact with the dopamine D<sub>3</sub> receptor (which is also known to play a role in the beneficial effects for other anti-Parkinsonian drugs). The nearest-neighbor drug within the D34 ligands was ropinirole (2D *p*-value = 0.210), which is structurally distinct in a topological sense in Figure 4. As in the previous case, when considering all 11 dopamine D<sub>3</sub> ligands, the 3D comparisons provide primary support for a positive log-odds

score. The 2D, 3D, and PPI log-odds were -1.1, 7.7, and 1.2 respectively. The combination of all three comparison types yielded a score of 3.3. Here, the 3D molecular similarity information was the most reliable predictor.

#### 4. Conclusion

In the present study, we report a means to combine chemical similarity between molecules with information derived from computing similarity based upon lexical analysis of patient package inserts (PPI). As expected based on our prior work, drugs that were highly structurally similar (both by 2D and 3D comparison) were much more likely to have significant overlap of their clinical effects compared to drugs that were structurally different (low 2D similarity but high 3D similarity). Our prior work illustrated a similar effect with respect to specifically annotated molecular targets: me-too drugs tend to have nearly identical target profiles.<sup>1</sup> The correlation between lexical and chemical similarity also served to validate the lexical comparison methodology.

We extended a probabilistic data fusion method to include observations from both molecular and clinical effects similarity and reported performance on predicting protein targets of small molecules. This was done both by leave-one-out cross-validation on our internal database of drug-target interactions (the SPDB) as well as on a blind test on new interactions present in ChEMBL. For off-target prediction within the SPDB, 3D similarity was the most effective single information source. However, combining the methods predicted a larger proportion of secondary targets than any of the individual methods, while maintaining a similar nominal false positive rate. On the test against previously unseen ChEMBL drug-target linkages, again 3D similarity was the single most effective predictor, but gains were derived from combining the different data sources. We note that the method supports the integration of *any* method that produces scores relating molecules to targets (e.g. docking), and that inclusion of additional information sources is likely to produce further benefits. It is also important to understand that this framework is similar in character to virtual screening methods, in that while enrichment for compounds with the predicted effects occurs, the actual potencies of the effects are not predicted. This point is discussed at length in a prior study.<sup>16</sup>

In contemplating the problem of off-target prediction for drugs, the problem of molecular design ancestry can confuse the issue of methodological validation. For example, ligands of aminergic GPCRs offer troublesome test case, owing to the established promiscuity of such drugs among numerous targets.<sup>17</sup> Returning to Figure 1 (bottom), we see the example of levetiracetam, an anticonvulsant believed to have a unique mechanism of action when compared with most existing anticonvulsants. The established CNS targets of the major classes of anticonvulsant drugs include the GABA<sub>A</sub> receptor (for barbiturates such as pentobarbital) and neuronal voltage-gated sodium channels (for drugs such as carbamazepine and phenytoin). These drugs have been recently shown to modulate voltage-gated potassium channels as part of their anti-epileptic effects.<sup>18–21</sup> Levetiracetam, having a novel scaffold, has been proposed to work through an entirely new mechanism of action due to high binding affinity to the synaptic vesicle protein SV2A (which is not a known therapeutic target of any drug).<sup>6,22,23</sup> Our methods strongly predict that levetiracetam is a voltage-gated sodium channel modulator

with 3D log-odds alone of 14.5 (the combination log-odds was 21.4). Levetiracetam has been shown to inhibit voltage-gated potassium currents,<sup>22</sup> leading to the suggestion that this drug, like other anti-epileptics, acts at least in part through potassium channels. Considering that many antiepileptics modulate *both* sodium and potassium channels,<sup>23</sup> our prediction supports the notion that levetiracetam shares a similar mechanism of action, perhaps in addition to the interaction with SV2A.

Identification of off-target activities of drugs is a difficult problem, particularly in cases where the drug in question has a non-obvious structural relationship with the known ligands of a given target. Our hope is that methods that make use of multiple information sources will help to identify clinically important and unexpected effects.

## References

1. E. Yera, A. Cleves and A. Jain, *Journal of Medicinal Chemistry* **54**, 6771 (2011).
2. A. E. Cleves and A. N. Jain, *Journal of Computer-Aided Molecular Design* **22**, 147 (2008).
3. J. Wright and R. Willette, *Journal of Medicinal Chemistry* **5**, 815 (1962).
4. K. Nagashima, A. Takahashi, H. Ikeda, A. Hamasaki, N. Kuwamura, Y. Yamada and Y. Seino, *Diabetes Research and Clinical Practice* **66**, S75 (2004).
5. D. S. Ragsdale and M. Avoli, *Brain Research* **26**, p. 16 (1998).
6. B. Lynch, N. Lambeng, K. Nocka, P. Kensel-Hammes, S. Bajjalieh, A. Matagne and B. Fuks, *PNAS* **101**, 9861 (2004).
7. M. Campillos, M. Kuhn, A. Gavin, L. Jensen and P. Bork, *Science* **321**, 263 (2008).
8. A. E. Cleves and A. N. Jain, *Journal of Medicinal Chemistry* **49**, 2921 (2006).
9. A. Gaulton, L. Bellis, A. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich and B. Al-Lazikani, *Nucleic Acids Research* **40**, D1100 (2012).
10. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, Inc., New York, NY, USA, 1986).
11. J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques, Second Edition* (Morgan Kaufmann Series in Data Management Systems), 2 edn. (Morgan Kaufmann, 2006).
12. A. N. Jain, *Journal of Computer-Aided Molecular Design* **14**, 199 (2000).
13. A. N. Jain, *Journal of Medicinal Chemistry* **47**, 947 (2004).
14. A. Ghuloum, R. Carleton and A. Jain, *Journal of Medicinal Chemistry* **42**, 1739 (1999).
15. J. Mount, J. Ruppert, W. Welch and A. Jain, *Journal of Medicinal Chemistry* **42**, 60 (1999).
16. A. Jain and A. Cleves, *J Comput Aided Mol Des* **26**, 57 (2012).
17. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nature Biotechnology* **25**, 197 (2007).
18. C. Zona, V. Tancredi, E. Palma, G. Pirrone and M. Avoli, *Canadian Journal of Physiology and Pharmacology* **68**, 545 (1990).
19. F. Bloom, D. Kupfer and B. Bunney, *Psychopharmacology: The Fourth Generation of Progress* (Raven Press, 1995).
20. M. Nobile and P. Vercellino, *British Journal of Pharmacology* **120**, 647 (1997).
21. A. Ambrósio, P. Soares-da Silva, C. Carvalho and A. Carvalho, *Neurochem. Res.* **27**, 121 (2002).
22. M. Madeja, D. Georg Margineanu, A. Gorji, E. Siep, P. Boerrigter, H. Klitgaard and E. Speckmann, *Neuropharmacology* **45**, 661 (2003).
23. R. Surges, K. Volynski and M. Walker, *Therapeutic Advances in Neurological Disorders* **1**, 13 (2008).

**EXPLORING THE PHARMACOGENOMICS KNOWLEDGE BASE (PHARMGKB) FOR  
REPOSITIONING BREAST CANCER DRUGS BY LEVERAGING WEB ONTOLOGY LANGUAGE (OWL)  
AND CHEMINFORMATICS APPROACHES\***

QIAN ZHU

*Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA*

*Email: zhu.qian@mayo.edu*

CUI TAO

*School of Biomedical Informatics, University of Texas Health Science Center at Houston, TX 77030, USA*

*Email: cui.tao@uth.tmc.edu*

FEICHEN SHEN

*School of Computing and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA*

*Email: fsm89@mail.umkc.edu*

CHRISTOPHER G. CHUTE

*Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA*

*Email: chute@mayo.edu*

Computational drug repositioning leverages computational technology and high volume of biomedical data to identify new indications for existing drugs. Since it does not require costly experiments that have a high risk of failure, it has attracted increasing interest from diverse fields such as biomedical, pharmaceutical, and informatics areas. In this study, we used pharmacogenomics data generated from pharmacogenomics studies, applied informatics and Semantic Web technologies to address the drug repositioning problem. Specifically, we explored PharmGKB to identify pharmacogenomics related associations as pharmacogenomics profiles for US Food and Drug Administration (FDA) approved breast cancer drugs. We then converted and represented these profiles in Semantic Web notations, which support automated semantic inference. We successfully evaluated the performance and efficacy of the breast cancer drug pharmacogenomics profiles by case studies. Our results demonstrate that combination of pharmacogenomics data and Semantic Web technology/Cheminformatics approaches yields better performance of new indication and possible adverse effects prediction for breast cancer drugs.

---

\* This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS-U19 GM61388) and the Cancer Prevention & Research Institute of Texas (CPRIT R1307).

## 1. Introduction

Traditional drug development is costly and labor-intensive, and scientists are devoted to finding an alternative way to facilitate the drug discovery process. Drug repositioning, finding new therapeutic uses for existing drugs, is one of the most efficient and efficacious approaches to speed drug discovery. With the advance in computational technology, computational drug repositioning has shown its advantage as many studies been published recently. Ye et al. [1] explored a disease-oriented strategy for evaluating the relationship between drugs and disease on the basis of their pathway profile; Napolitano et al. [2] investigated machine-learning algorithms to predict drug repositioning; Li and Lu[3] presented an approach for identifying potential new indications of an existing drug through its relation to similar drugs. Butte's lab has reported their efforts on computational drug repurposing by exploring gene expression data [4, 5]. These studies drew on different technologies to address the problem of computational drug repositioning. However, none of them attempted to leverage data from emerging pharmacogenomics (PGx) studies in an integrated and transformable manner and explore Semantic Web technology as core implementation tool to address drug repositioning, which is our proposed aim for this study. PGx study investigates how genetic variations affect drug responses for the individual patient, consequently high volume of PGx information including relations among drugs, genes, single nucleotide polymorphisms (SNPs), etc. has been accumulated. The overarching goal of this study was to provide PGx profiles for FDA approved breast cancer drugs (BCDs) by leveraging informatics approaches and Semantic Web technologies, and ultimately to facilitate oncology-relevant biomedical and clinical studies and to support breast cancer drug repositioning.

Currently in the PGx world, different formats are being used for different data resources, which is the main obstacle to integration of PGx data to support development of relevant applications. Different formats might be preferred to represent scientific data, based on the nature of the source, the way the data are to be queried or visualized, or the type of analyses to be performed. Traditionally, investigators have relied heavily on tools such as Excel spreadsheets and relational databases to store and represent their research findings. However, these tools lack interoperability and capability to make inferences. In contrast, Semantic Web technology can manage scientific data in a more integrative and intelligent way. It is "a rigorous mechanism for defining and linking data using Web protocols in such a way that the data can be used by machines not just for display, but also for automation, integration, and reuse across various applications"[6]. Web Ontology Language (OWL), as a Semantic Web standard, can formally represent domain knowledge; it "organizes concepts or entities within classification (specialization or "is-a") hierarchies that provide for inheritance of attributes"[7]. Reusing existing resources in an integrative manner is essential, but exploring new associations is much more challenging. A Semantic Web reasoner enables identification of new BCD PGx associations, with an ultimate goal of repositioning BCDs. Dumontier [10] has demonstrated some advantages by expressing PGX data, PharmGKB in OWL for personalized medicine purpose.

Additionally, novel PGx information may be detected from a chemical perspective. Drugs with chemical structure similar to that of cancer drugs or genes associated with drugs with similar chemical structure can be identified using cheminformatics approaches[8]. Cheminformatics, a suite of computational technologies to solve a range of chemical problems, can be used to identify and evaluate new PGx associations. More precisely, we implemented a similar-structure searching algorithm to identify drugs similar to BCDs and find potential new uses for these drugs.

The paper is organized into the following sections. First, we introduce materials being used in this study; second, in the Methods section, we introduce details about PGx OWL profiles generation for BCDs and case study; third, we illustrate our results generated from each step in the Results section, which is followed by Discussion and Conclusion.

## 2. Materials

### 2.1. PharmGKB

The PharmGKB contains genomic, phenotype and clinical information collected from PGx studies. PharmGKB provides information regarding variant annotations, drug-centered pathway, pharmacogene summaries, clinical annotations, PGx-based drug-dosing guidelines, and drug labels with PGx information[9].

In this study, we used PGx information extracted from a relationship file received from PharmGKB by May 8, 2013, to generate the PGx profile for FDA-approved BCDs. Figure 1 shows some concrete PGx related association examples from the PharmGKB relationship file. Particularly, we extracted “Entity id”, “Entity name”, and “Entity type” for this study. Other fields, such as PubMed IDs (PMIDs), will be explored and integrated in a future study to support selection of the best PGx associations with publications as evidence.

A	B	C	D	E	F	G	H	I	J	K
Entity1_id	Entity1_name	Entity1_type	Entity2_id	Entity2_name	Entity2_type	Evidence	Association	PK	PD	PMIDs
PA150481189	taxanes	Drug	PA162387925	FAM82A1	Gene	ClinicalAnnotation	associated		PD	17224914;
PA150481189	taxanes	Drug	PA267	ABCB1	Gene	ClinicalAnnotation,VariantAnnotation	associated	PK	PD	12684679;
PA150481189	taxanes	Drug	PA27094	CYP1B1	Gene	ClinicalAnnotation	associated		PD	17224914;
PA150481189	taxanes	Drug	PA29028	GSTP1	Gene	VariantAnnotation	associated		PD	19203783
PA150481189	taxanes	Drug	rs1045642	rs1045642	VariantLocation	VIP,VariantAnnotation	associated		PD	12684679
PA150481189	taxanes	Drug	rs1056836	rs1056836	VariantLocation	ClinicalAnnotation	associated		PD	17224914;
PA150481189	taxanes	Drug	rs1695	rs1695	VariantLocation	VariantAnnotation	associated		PD	19203783
PA150481189	taxanes	Drug	rs2032582	rs2032582	VariantLocation	ClinicalAnnotation,VariantAnnotation	associated	PK	PD	16467099;
PA449412	doxorubicin	Drug	CYP2C8 *1A, CYP2C8 *3	CYP2C8 *1A, CYP2C8 *3	Haplotype	VariantAnnotation	not associated		PD	22527101
PA449412	doxorubicin	Drug	PA116	ABCC2	Gene	ClinicalAnnotation,Pathway,VariantAnnotation	associated	PK	PD	16330681,2
PA449412	doxorubicin	Drug	PA123	CYP2B6	Gene	ClinicalAnnotation,VariantAnnotation	associated		PD	20179710
PA449412	doxorubicin	Drug	PA124	CYP2C19	Gene	ClinicalAnnotation,VariantAnnotation	associated		PD	20179710
PA449412	doxorubicin	Drug	PA125	CYP2C8	Gene	VariantAnnotation	not associated		PD	22527101
PA449412	doxorubicin	Drug	PA130	CYP3A4	Gene	VariantAnnotation	associated		PD	20459744
PA449412	doxorubicin	Drug	PA134911502	SLC22A16	Gene	ClinicalAnnotation,Pathway,VariantAnnotation	associated	PK	PD	17559346,2
PA449412	doxorubicin	Drug	PA134956204	C18orf56	Gene	VariantAnnotation	not associated		PD	19159907
PA449412	doxorubicin	Drug	PA142671588	KLC3	Gene	VariantAnnotation	associated		PD	21826087
PA449412	doxorubicin	Drug	PA164724093	NO52	Gene	Pathway	associated		PD	21048526

Fig. 1. Examples of PGx relations available in PharmGKB

In addition to the PGx information from the PharmGKB relationship file shown in Figure 1, PharmGKB also provides pathway information, which includes associations between pathway and

drug, pathway and gene, and pathway and disease. Overall ten associations among drugs, genes, diseases, pathways, SNPs are available from PharmGKB. Table 1 shows these associations from two PharmGKB data files. Haplotype related associations are beyond the scope of this study.

Table 1. PGx related associations available from PharmGKB

Associations Resources	Drug- Drug	Drug- Gene	Drug- Pathway	Drug- SNP	Gene- Pathway	Gene- Disease	Disease- Pathway	Disease- SNP	Gene- Disease	Gene- Gene
PharmGKB Relationship file	√	√		√		√		√	√	√
PharmGKB Pathway data			√		√		√			

## 2.2. FDA approved BCDs

The National Cancer Institute (NCI) maintains cancer drugs approved by the FDA for breast cancer[11]. In this study, we did not consider drug combinations that are not approved by the FDA, even though the individual drugs are approved. Of 23 BCDs from NCI, a total of 18 BCDs have been manually mapped to the PharmGKB relationship file. The PGx profiles have been generated for these 18 BCDs, as described in the following sections. Table 2 shows the 23 BCDs from NCI vs 18 BCDs mapped to PharmGKB.

Table 2. BCDs from NCI and PharmGKB<sup>a</sup>

BCDs available from NCI	BCDs identified in PharmGKB relationship file
<b>ado-trastuzumab emtansine</b> anastrozole capecitabine cyclophosphamide docetaxel doxorubicin hydrochloride epirubicin hydrochloride everolimus exemestane fluorouracil fulvestrant gemcitabine hydrochloride <b>ixabepilone</b> lapatinib ditosylate letrozole <b>megestrol acetate</b> methotrexate paclitaxel <b>paclitaxel albumin-stabilized nanoparticle formulation</b> pertuzumab tamoxifen citrate trastuzumab <b>toremifene</b>	anastrozole capecitabine cyclophosphamide docetaxel doxorubicin epirubicin everolimus exemestane fluorouracil fulvestrant gemcitabine lapatinib letrozole methotrexate paclitaxel pertuzumab tamoxifen trastuzumab  <sup>a</sup> Drugs that failed to map to PharmGKB are shown in bold.

### 2.3. Semantic Web Technologies

Emerging Semantic Web technologies provide a formal mechanism to represent domain knowledge and data and to perform semantic reasoning on top of this knowledge. Semantic Web technology supports flexible, extensible, and evolvable knowledge transfer and reuse. It has been widely used in biomedical domains to formalize and model medical and biological systems. The Resource Description Framework (RDF)[12] is a World Wide Web Consortium (W3C) standard that specifies a graph-based data model for representing Semantic Web data. Each piece of information is represented in three parts (a triple): subject, predicate, and object. The RDF representations allow efficient querying and visualization of relationships between important biomedical entities. OWL [13] is a standard ontology language for the Semantic Web. A distinguishing characteristic of RDF and ontologies compared with the conventional relational database is “their degree of connectedness, their ability to model coherent, linked relationships”[14]. Representing the associations using OWL will enable powerful data integration among heterogeneous data sets, which is a well-known challenge in the translational science study community.

## 3. Methods

In this study, we focused on FDA approved BCDs and generated PGx OWL profiles by leveraging PharmGKB data and semantic web technologies. The OWL profiles explicitly capture BCD concepts and relationships and enable the semantic inference for novel drug associations. The overall architecture of the proposed project is shown in Figure 2. The details about each step are described in the following sections.

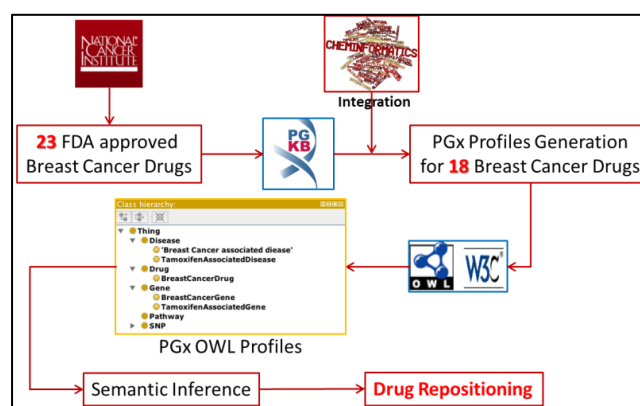


Fig. 2. Building blocks for the overall architecture

### 3.1. Generation of Integrative Breast Cancer PGx Profiles

#### 3.1.1. BCD PGx related association extraction

The PGx related associations shown in Table 1 were explored in this study for generation of PGx profiles. We programmatically extracted the PGx related associations from the relationship file that is tab delimited. In addition, we manually identified associations among pathways, drugs, genes and diseases for 18 BCDs from the PharmGKB pathway file that is a plain text file. Additional associations were inferred by invoking a rule-based OWL reasoner described in section 3.2.



### 3.1.2. *Chemical structure based similarity calculation*

To identify inferred associations for BCDs from a chemical perspective, two steps were involved: retrieval of chemical representations (by the simplified molecular-input line-entry system [SMILES] [15] or the IUPAC International Chemical Identifier InChI [16]) and structural similarity calculation. Except for the drugs with SMILES annotated by PharmGKB, we first converted active ingredient names to chemical representations through publically accessible services, such as the PubChem Entrez web service [17] and the NCI Chemical Identifier Resolver [18]. We then translated such chemical representations to chemical fingerprints and compared chemical structure similarity between BCDs and drugs from the PharmGKB by calculating the Tanimoto coefficient [19]. A cheminformatics toolkit, the Chemical Development Kit [20], has been explored to automate these two steps. Finally, PharmGKB drugs with similarity scores higher than 0.7 compared with BCDs were marked as structurally similar BCDs. Thus, more PGx related associations were transformable to BCDs via similar PharmGKB drugs. Appropriate properties for describing the similar structural relationships have been defined and used for inference in PGx OWL profiles for BCDs.

## 3.2. *BCD PGx OWL profile construction and semantic inference*

We captured and integrated PGx related associations for BCDs as PGx profiles. These integrated PGx profiles can then serve as a knowledge base to further infer new drug targets or associations. We established an OWL ontology-based approach for this purpose. More specifically, we developed an OWL ontology that captures 1) comprehensive BCDs' PGx profiles and 2) rules to infer drug targets or other associations based on the profiles. We used the Protégé system[21] for OWL ontology development.

### 3.2.1. *Meta-ontology model definition*

We first defined a meta-ontology model to describe base classes and relationships for the BCD profiles. Base classes include "Drug," "Gene," "Disease," "SNP," and "Pathway." Specific subclasses of these base classes, such as "Breast Cancer Drug" or "Breast Cancer Drug Associated SNP," can also be defined. Relationships between these classes, such as "associatedwithDrug," "associatedwithDisease," "associatedwithSNP," and "associatedwithPathway," have also been defined as object properties with appropriate domains and ranges.

### 3.2.2. *PGx profile representation*

Specific BCDs, SNPs, genes, and pathways were represented as OWL individuals with appropriate types. For example, line 1 in Figure 3 defines Tamoxifen as an instance of the Drug class. Lines 2-5 further represent additional PGx profile information about the Drug Tamoxifen. Similarly, information about particular genes, SNPs, diseases, and pathways can also be stored using RDF

triples. For example, lines 8-10 and 13-14 represent a partial profile of SNP rs2234693 and the drug clomifene, respectively.

### 3.2.3. Identifying new indications for BCDs via semantic inference

New indication candidates identification for BCDs is built on the

basis of PGx related associations and predefined axioms. We used Description Logic (DL)[22] to define axioms shown in Figure 4. For instance, we defined that a disease *di* may associate with a drug *dr* if *di* is either directly associated with *dr* or associated with any gene, pathway, or SNP that is associated with *dr*. For example, we can find tamoxifen-associated diseases using the first axiom listed in Figure 4. Similarly, we can define a tamoxifen-associated SNP, gene, and pathway using OWL DL. Another way to find tamoxifen-associated disease is to search on the basis of its chemical structure. Our method is based on the fact that drugs with the similar structure (isStructuralSimilarto) are very likely to share the same biological properties, which would likely lead to the same disease profile. The second axiom in Figure 4 defines this feature.

```

1. Tamoxifen rdf:type Drug
2. Tamoxifen associatedwithGene ESR1
3. Tamoxifen associatedwithGene BRCA1
4. Tamoxifen associatedwithDisease Breast_Cancer
5. Tamoxifen associatedwithPathway Aromatase_Inhibitor_Pathway_(Breast_Cell)_Pharmacodynamics
6. Tamoxifen isStructuralSimilarto Clomifene
7. ....
8. rs2234693 rdf:type SNP
9. rs2234693 associatedwithGene ESR1
10. rs2234693 associatedwithDisease Rheumatoid_Arthritis
11. ESR1 associatedwithDisease Ovarian_Neoplasms
12. ....
13. Clomifene rdf:type Drug
14. Clomifene associatedwithDisease Polycystic_Ovary_Syndrome

```

Fig. 3. RDF representation for PGx profiles

#### tamoxifen-associated disease:

1. Disease and (associatedwithDrug value Tamoxifen or associatedwithSNP some TamoxifenSNP or associatedwithGene some TamoxifenGene or associatedwithPathway some TamoxifenPathway)
2. Disease and associatedwithDrug some (Drug and isStructuralSimilarto value Tamoxifen)

Fig. 4. Rule representation for PGx OWL profiles.

## 4. Case Study

Using the above semantic definitions, we can infer more information about a particular BCD. We chose tamoxifen, as a use case testbed. “Tamoxifen treats advanced breast cancer in men and women, and early breast cancer in women. And it may prevent breast cancer in women who are at a high risk because of age, family history, or other factors”[23]. We did not invite domain experts to evaluate our inference results for this study, hence, we attempted to validate the performance and usability of PGx OWL profiles by detecting existing hints from the literature as evidence.

Tamoxifen is associated with the BRCA1 gene (a TamoxifenGene, in Figure 3) and BRCA1 is associated with the disease “Ovarian Neoplasms”.

The reasoner can infer ovarian cancer might be associated with tamoxifen via the first axiom listed in

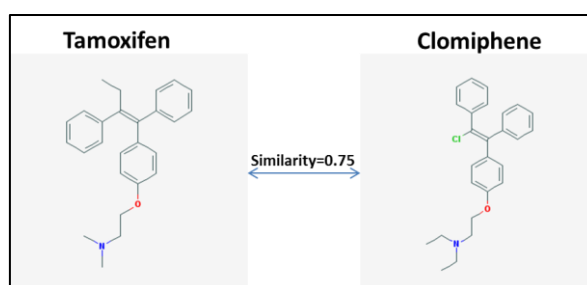


Fig. 5 Structural comparison between tamoxifen and clomifene.

Figure 4. That is to say, tamoxifen can not only treat breast cancer, but also may be used to treat ovarian cancer. Several publications and clinical trials have reported this use of tamoxifen.[24, 25]

“Clomifene treats ovulation problems in women who want to become pregnant”[26]. There are no explicit hints to tie together an ovulation drug and a BCD. However, PGx OWL profiles identified a possible linkage between these two agents. As shown in Figure 5, clomifene and tamoxifen are structurally similar with a similarity score 0.75, which is higher than the threshold 0.7 that we setup. Then the reasoner can infer that tamoxifen may be associated with diseases associated with clomifene (eg, Polycystic\_Ovary\_Syndrome) via the second axiom shown in Figure 4. In 2011, Dhaliwal et al [27] reported that tamoxifen can be prescribed as an alternative to clomifene in women with polycystic ovary syndrome.

In addition to repositioning tamoxifen with other therapeutic uses, we also can identify potential adverse effects by running our PGx OWL profiles based reasoner. From our OWL profiles, as shown in Figure 3, we identified that tamoxifen is associated with the *ESR1* gene as a “TamoxifenGene.” Since the SNP rs2234693 is associated with *ESR1* (a “TamoxifenGene”), rs2234693 is classified as a “TamoxifenSNP” by the reasoner. Furthermore, since rs2234693 is “associatedwithDisease” Rheumatoid Arthritis, then rheumatoid arthritis is identified as a disease that might be associated with tamoxifen by the reasoner. In the real world, as of June 24, 2013, a total of 7,947 people have been reported to have adverse effects when taking tamoxifen citrate. Among them, 35 people (0.44%) have rheumatoid arthritis. [28]

## 5. Results

We generated and presented PGx profiles for 18 breast cancer drugs from NCI by exploring PGx information from PharmGKB. To enable semantic reasoning and to identify more novel PGx associations for BCDs, we created OWL ontology to capture and represent the concepts and relations from PGx profiles.

### 5.1. BCD PGx profile generation

We identified 955 associations for 18 BCDs from the PharmGKB relationship file, which include associations among drugs, genes, and SNPs. We manually identified 287 associations for 18 BCDs from the PharmGKB pathway file, which include associations among pathways, drugs, genes, and diseases.

### 5.2. Chemical structural similarity calculation

To integrate structural similarity, we calculated drug pairs between BCDs and drugs from the PharmGKB. Of 679 unique PharmGKB drugs (including drug classes) extracted from the PharmGKB relationship file, 339 are without SMILES. We invoked NCI chemical resolver to generate SMILES for

these 339 drugs by given drug names, 193 have retrieved SMILES. For the rest of 146 drugs and drug classes without SMILES, we ran PubChem entrez web service to generate SMILES and 37 drugs assigned with SMILES. In total 78 drug classes and 31 drugs were excluded from similarity calculation because no SMILES were generated. For pathway file, we have identified another 71 unique drugs. Among these drugs, there are 65 drugs assigned SMILES via PubChem Entrez web service. Total 5 drugs and 26 drug classes without SMILES were excluded for similarity calculation.

### 5.3. PGx OWL profile generation

BCDs relevant PGx profiles were converted to OWL representation, the drugs, genes, diseases, SNPs from the PharmGKB relationship file and pathway file were also imported into the OWL ontology for inference purpose. A snapshot of the PGx OWL ontology is shown in Figure 6. This ontology includes 294 diseases, 750 drugs including 18 breast cancer drugs, 4277 genes including 215 breast cancer associated genes, 1,426 pathways including 15 breast cancer drugs involved pathways, and 1744 SNPs including 346 breast cancer associated SNPs. It also includes the similarity scores of 10,159 pairs of drugs.

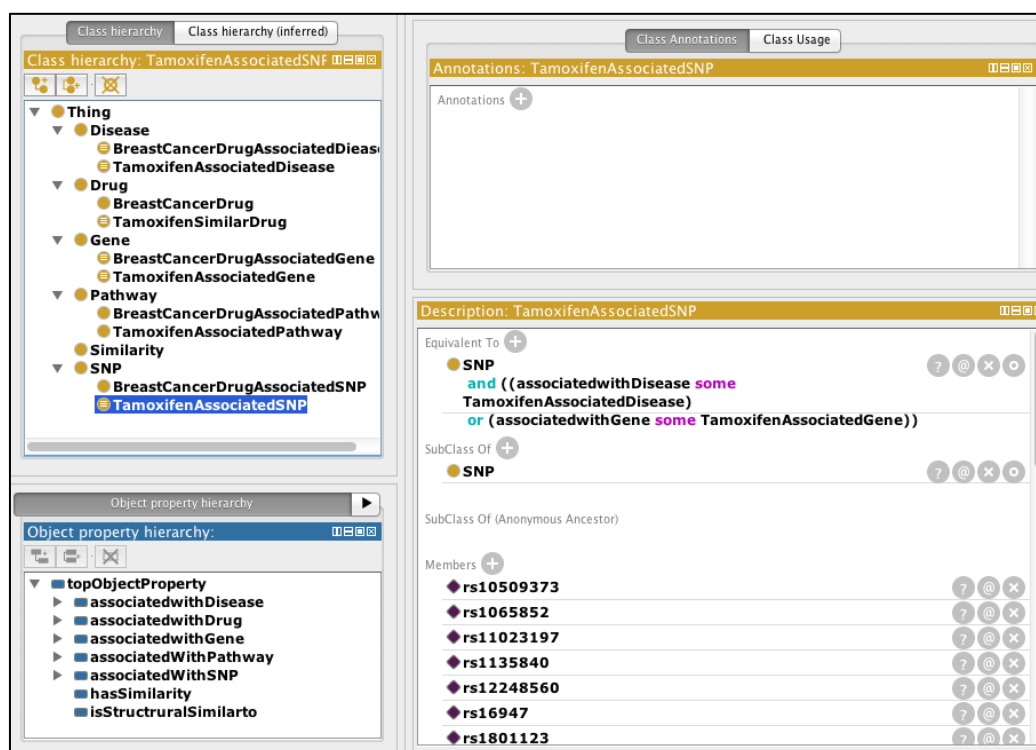


Fig. 6. PGx OWL ontology snapshot

## 6. Discussion and Conclusion

This report presents our preliminary work focusing on computational drug repositioning application development leveraging PGx information integration and Semantic Web technology exploration for

FDA approved BCDs. We have successfully demonstrated the utility of this application to reposition existing BCDs with new uses, and detect potential adverse effects. Our work illustrates that PGx data provides sufficient information to support drug repositioning and, furthermore, that Semantic Web technology provides technical support for formal representation and semantic inference of data.

This is our first attempt to use a PGx resource and Semantic Web technology to address drug repositioning in a computational way. With the promising results of this study, we will expand this investigation in several directions: 1) In the current study, we explored only PharmGKB as a PGx resource, which is not enough to identify more novel associations for BCDs. We will integrate additional PGx-related resources, such as an FDA biomarkers table, the DrugBank database, the Comparative Toxicogenomics Database, and the Kyoto Encyclopedia of Genes and Genomes. 2) Once more PGx resources are integrated, one drug might be inferred to multiple PGx associations. Then we will propose to define some “gold standards” for prioritizing the relevance of these associations to particular drugs. The standards might be built on the number of co-occurrences of the PGx associations, as supported by publications, etc. 3) We worked only on BCDs in this study. In future studies, we will extend our effort to other cancer drug categories or other categories of drugs, such as antidepressants, using the same strategy that we applied in this study.

## 7. Acknowledgments

This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS-U19 GM61388) and the Cancer Prevention & Research Institute of Texas (CPRIT R1307).

## References

1. Ye, H., et al., *A pathway profile-based method for drug repositioning*. Chinese Science Bulletin, 2012. **57**(17): p. 2106-2112.
2. Napolitano, F., et al., *Drug Repositioning: A Machine-Learning Approach through Data Integration*. Journal of Cheminformatics, 2013. **5**(1): p. 30.
3. Li, J. and Z. Lu. *A new method for computational drug repositioning using drug pairwise similarity*. in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. 2012: IEEE.
4. Dudley, J.T., et al., *Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease*. Science Translational Medicine, 2011. **3**(96): p. 96ra76.
5. Sirota, M., et al., *Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data*. Science Translational Medicine, 2011. **3**(96): p. 96ra77.
6. Pathak, J., R.C. Kiefer, and C.G. Chute, *Using Semantic Web Technologies for Cohort Identification from Electronic Health Records for Clinical Research*. AMIA Summits on Translational Science Proceedings, 2012. **2012**: p. 10.
7. Searls, D.B., *Data integration: challenges for drug discovery*. Nature Reviews Drug Discovery, 2005. **4**(1): p. 45-58.

8. Olsson, T. and T.I. Oprea, *Cheminformatics: a tool for decision-makers in drug discovery*. Current opinion in drug discovery & development, 2001. **4**(3): p. 308.
9. Hewett, M., et al., *PharmGKB: the pharmacogenetics knowledge base*. Nucleic acids research, 2002. **30**(1): p. 163-165.
10. Dumontier, M. and N. Villanueva-Rosales, *Towards pharmacogenomics knowledge discovery with the semantic web*. Briefings in Bioinformatics, 2009.
11. *Drugs Approved for Breast Cancer*. [cited 2013 July 3rd]; Available from: <http://www.cancer.gov/cancertopics/druginfo/breastcancer>.
12. *Resource Description Framework (RDF)*. [cited 2013 July 17th]; Available from: <http://www.w3.org/RDF/>.
13. *OWL Web Ontology Language*. [cited 2013 July 17th]; Available from: <http://www.w3.org/TR/owl-features/>.
14. *An Executive Intro to Ontologies*. 2009; Available from: <http://www.mkbergman.com/900/an-executive-intro-to-ontologies/>.
15. *Simplified molecular-input line-entry system (SMILES)*. [cited 2013 July 3rd]; Available from: [http://en.wikipedia.org/wiki/Simplified\\_molecular-input\\_line-entry\\_system](http://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system).
16. *The IUPAC International Chemical Identifier (InChI)*. [cited 2013 July 3rd]; Available from: <http://www.iupac.org/home/publications/e-resources/inchi.html>.
17. *PubChem Entrez Programming Utilities*. [cited 2013 July 3rd]; Available from: <http://www.ncbi.nlm.nih.gov/books/NBK25500/>.
18. *NCI Chemical Identifier Resolver*. [cited 2013 July 3rd]; Available from: <http://cactus.nci.nih.gov/chemical/structure>.
19. Butina, D., *Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets*. Journal of Chemical Information and Computer Sciences, 1999. **39**(4): p. 747-750.
20. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics*. Journal of Chemical Information and Computer Sciences, 2003. **43**(2): p. 493-500.
21. *Protégé*. [cited 2013 July 3rd]; Available from: <http://protege.stanford.edu/>.
22. Baader, F. 2003: Cambridge university press.
23. *Tamoxifen*. [cited 2013 July 3rd]; Available from: <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0012290/?report=details>.
24. Lee, J.-Y., et al., *Effect of combined treatment with progesterone and tamoxifen on the growth and apoptosis of human ovarian cancer cells*. Oncology reports, 2012. **27**(1): p. 87-93.
25. *Clinical Trail for Tamoxifen Compared With Thalidomide in Treating Women With Ovarian Epithelial Cancer, Fallopian Tube Cancer, or Primary Peritoneal Cancer*. [cited 2013 July 3rd]; Available from: <http://clinicaltrials.gov/ct2/show/record/NCT00041080>.
26. *Clomiphene* [cited 2013 July 3rd]; Available from: <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0009673/?report=details>.
27. Dhaliwal, L.K., et al., *Tamoxifen: An alternative to clomiphene in women with polycystic ovary syndrome*. Journal of human reproductive sciences, 2011. **4**(2): p. 76.
28. *Tamoxifen citrate and Rheumatoid arthritis*. [cited 2013 July 3rd]; Available from: <http://www.ehealthme.com/ds/tamoxifen+citrate/rheumatoid+arthritis>.

## DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES

Anna L. Tyler

*The Jackson Laboratory  
Bar Harbor, ME, USA  
E-mail: anna.tyler {at} jax.org*

Dana C. Crawford

*Department of Molecular Physiology & Biophysics,  
Vanderbilt University  
Nashville, TN 37240, USA  
E-mail: dana.c.crawford {at} vanderbilt.edu*

Sarah A. Pendergrass

*Department of Biochemistry and Molecular Biology, Pennsylvania State University,  
Pennsylvania State University  
University Park, PA 16802, USA  
E-mail: sap29 {at} psu.edu*

### Introduction

Pleiotropy, the phenomenon in which one gene influences more than one phenotype, was first defined over 100 years ago by Ludwig Plate.<sup>1</sup> Since that time our understanding of pleiotropy has changed and expanded to incorporate knowledge of molecular genetics. With this increase in knowledge has come an increase in an appreciation for the importance of pleiotropy in human health and evolutionary dynamics, but also a corresponding increase in confusion about how pleiotropy should be measured, and how the context in which pleiotropy is measured affects its interpretation. The purpose of this session is to offer a general view of pleiotropy and of the different approaches used to study this phenomenon. During the session, we will be examining a series of questions that are currently being discussed in the literature:

- Which genetic elements can be defined as pleiotropic?
- How are phenotypes defined and counted?
- How prevalent is pleiotropy? Does every gene affect every phenotype, or is pleiotropy more limited?
- How does pleiotropy influence, and how is it influenced by, evolution?
- How does an understanding of pleiotropy improve our understanding of human health?

### *Defining Genetic Elements*

Pleiotropy requires defining a genetic element that affects multiple phenotypes. But which genetic element is appropriate? A gene, a chromosomal segment with high linkage disequilibrium, a mutation? Plate's original definition of pleiotropy came long before the discovery of

DNA, and referred to a “unit of inheritance”.<sup>1</sup> A unit of inheritance may refer to a single nucleotide polymorphism (SNP), or a gene, or a large segment of the genome containing multiple genes. Individual mutations may affect a single gene, or multiple genes.<sup>2,3</sup> In this session, the problem is addressed in a variety of ways. **Darabos et al.** use SNPs and genes, while **Philip et al.** use expression quantitative trait loci (eQTL) as genetic elements. The choices made in different experiments clearly have implications for the interpretation of pleiotropy, although the implications of these choices is still being debated.

### *Defining Phenotypes*

The concept of counting phenotypes is perhaps an even more difficult issue than identifying genetic elements. As Wagner and Zhang point out, a biologist may see two traits, femur length and tail length, where a mathematician familiar with rotation of coordinate systems may see only one: FeTail.<sup>3</sup> Further disagreements may arise as to whether two correlated traits such as femur length and femur width are one trait or two. Finally, there is discussion about whether the relationship between traits represents yet another phenotype that can be affected by genetic manipulation. Relationship QTL, or rQTL, which change the relationships between phenotypes have been identified in mice,<sup>4,5</sup> and are likely present in other organisms as well.

To reduce the number of subjective choices, some studies, such as **Philip et al.** in this session, use mRNA expression levels as phenotypes. mRNA expression is relatively easy to measure comprehensively and at scale. However, problems such as high-dimensional data with relatively few samples and correlation between phenotypes arise in these studies. One class of methods used to address these problems is dimensionality reduction, and in this session **Philip et al.** discuss one such dimensionality reduction approach.

Measurement of physiological traits in humans is addressed by **Hall et al.**. This paper examines the concept of standardized, high-throughput phenotyping in a range of medically relevant areas from physiological measurements available in electronic medical records to environmental exposures.

### *The Prevalence of Pleiotropy*

In addition to specific relationships between individual genes and phenotypes, many studies of pleiotropy are concerned with quantifying pleiotropy itself. Ronald Fisher promoted the idea of “universal pleiotropy” in which every gene affects every phenotype to some extent either directly or indirectly.<sup>6</sup> This idea was implicit in his geometric model of adaptation.<sup>6,7</sup> However, since the 1930’s molecular genetics experiments have revealed a more modular model of pleiotropy.<sup>8</sup> In modular pleiotropy, gene actions are limited to a specific set of processes or phenotypes and are relatively independent from other phenotypes<sup>3,9</sup>

Modular pleiotropy is supported in the literature. Wagner and Zhang<sup>3</sup> review the results from experiments in yeast, nematode and mouse, using a variety of methods of counting genetic elements and phenotypes. In each of these experiments, the vast majority of genetic elements affect very few phenotypes, while only a few elements affect a large number of phenotypes. These distributions are surprisingly consistent across the different experiments. This limited scope of the majority of genes supports the hypothesis that gene action is relatively limited to



phenotypes modules. The paper by **Darabos et al.** presented in this session shows additional evidence in support of a modular view of pleiotropy.

### ***Pleiotropy in an Evolutionary Context***

Whether pleiotropy is universal or modular has an impact on how pleiotropic genes are influenced by selection. Complex organisms have vastly more cell types than prokaryotes, but only about four-fold more genes.<sup>10</sup> The necessary increase in pleiotropy per gene that this statistic suggests could limit the evolvability in complex organisms due to potentially wide-spread effects of single mutations. Des Marais and Rausher<sup>11</sup> have proposed that gene duplication may provide an escape from these evolvability limitations, as each gene copy can take over a subset of the original gene's functions. Other studies have addressed molecular mechanisms by which genes evolve to be more pleiotropic. This process may preferentially recruit genes to new biological processes rather than adding new biological functions.<sup>3</sup> For example, new processes might include changes in tissue expression, subcellular localization, interacting partners and context-sensitive transcription.<sup>3</sup>

In this session, the relationship between pleiotropy and evolution will be addressed by the keynote speaker, **James Cheverud** in his talk titled "Genetic Variation and Evolution of Pleiotropy."

### ***Pleiotropy and Human Health***

The importance of pleiotropy in human health is undeniable. Pleiotropy coupled with dynamic networks that exist between the genetic architecture, signaling pathways, intermediate phenotypes, and outcome traits can be an important part of health and disease and may become important for network-based medicine.<sup>12</sup> Phenomics, phenome scans, and phenome-wide association studies may provide a high-throughput way for exploring both pleiotropy and the diseasome.<sup>13–19</sup> Identifying genetic variation that confers both protection for some traits/outcomes but risk for others may both highlight important genetic regions, and also show important features of larger biological networks. Knowing which genes influence which phenotypes may aid in drug repurposing for genetically related diseases, as well as predicting off-target effects of targeted therapies. All papers in this session address human health either directly or indirectly. **Philip et al.** investigate QTL that interact to affect kidney health in a mouse model of kidney disease. **Darabos et al.** explore the relationships between SNPs, genes and pathways, and phenotypes to show novel molecular relationships between human diseases. And finally **Hall et al.** discuss methods of standardized, high-throughput phenotype measurement in patients with type 2 diabetes (T2D).

### **Session Contributions**

The keynote lecture for this session will be given by **James Cheverud** who has worked extensively on pleiotropy and the evolution of pleiotropy in mammals. He will speak on the "Genetic Variation and Evolution of Pleiotropy."

**Philip et al.** investigate epistasis and pleiotropy at the transcript level in an F2 mouse cross designed to examine kidney function. This paper presents a method called the Com-

bined Analysis of Pleiotropy and Epistasis (CAPE) which combines information across multiple phenotypes to infer directional interactions between genetic variants. This method has previously been used to examine pleiotropy related to physiological traits and now focuses on pleiotropy at the level of transcription. The authors found loci on eight chromosomes that interact to influence three expression modules. This method was further able to distinguish between which markers are truly pleiotropic and affect more than one module, and which are indirectly pleiotropic, affecting multiple modules through interactions with other genetic loci. This paper directly addresses several open issues in pleiotropy research including dimension reduction for high-dimensional phenotype spaces and distinguishing direct pleiotropy from indirect pleiotropy.

**Darabos et al.** also combine analysis of epistasis and pleiotropy. This paper constructs a bipartite network of genetic elements and phenotypes reported in GWAS data and other public repositories in a method similar to that used to construct the human diseasome.<sup>20</sup> However, unlike the diseasome, the network constructed by **Darabos et al.** includes non-disease phenotypes, such as hair color, as well as risk-associated SNPs that fall outside of coding regions. The authors constructed networks at three different levels of resolution: SNPs, genes, and pathways. These networks show that most genes have limited pleiotropic effects, supporting a model of modular pleiotropy. The pathway-base network also proves to be particularly informative and shows well established links between glaucoma and blood pressure, as well as glaucoma and type 2 diabetes. The network also shows a novel relationship between glaucoma and Alzheimer's disease, a connection that has only recently begun to be investigated. This paper shows the powerful predictions in human health that can be made by taking into account both epistasis and pleiotropy.

**Hall et al.** conduct an environment-wide association study (EWAS) to investigate contributions of environmental exposures and lifestyle choices to type 2 diabetes (T2D) in a high-throughput manner. The study employs a combination of resources, including electronic medical health records, the PhenX toolkit for standardized exposure measurement, and the Diet History Questionnaire. The authors find that moderate alcohol use is associated with decreased risk of T2D, and that low amounts of activity during leisure time, as well as smoking are positively associated with T2D. These relationships replicated in two independent populations and are supported by previous literature. This paper demonstrates the importance and practicality of standardized, high-throughput measurements of human phenotypes and environmental exposures, a field that is critical to further study of pleiotropy in humans.

## Acknowledgments

We thank all of the authors who contributed papers to this session and all of the anonymous reviewers who gave their time and expertise to review those papers.

## References

1. F. W. Stearns, *Genetics* **186**, 767 (2010).
2. J. Hodgkin, *The International Journal of Developmental Biology* **42**, 501 (1998).
3. G. P. Wagner and J. Zhang, *Nature Reviews Genetics* **12**, 204 (2011).

4. J. M. Cheverud, T. H. Ehrich, T. T. Vaughn, S. F. Koreishi, R. B. Linsey and L. S. Pletscher, *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* **302**, 424 (2004).
5. M. Pavlicev, J. P. Kenney-Hunt, E. A. Norgard, C. C. Roseman, J. B. Wolf and J. M. Cheverud, *Evolution* **62**, 199 (2008).
6. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, 1930).
7. H. A. Orr, *Evolution* **54**, 13 (2000).
8. J. J. Welch and D. Waxman, *Evolution* **57**, 1723 (2003).
9. P. Mitteroecker, *Evolutionary Biology* **36**, 377 (2009).
10. N. Lane and W. Martin, *Nature* **467**, 929 (2010).
11. D. L. Des Marais and M. D. Rausher, *Nature* **454**, 762 (2008).
12. A.-L. Barabási, N. Gulbahce and J. Loscalzo, *Nature Reviews Genetics* **12**, 56 (2011).
13. R. M. Bilder, F. Sabb, T. Cannon, E. London, J. Jentsch, D. S. Parker, R. Poldrack, C. Evans and N. Freimer, *Neuroscience* **164**, 30 (2009).
14. M. B. Lanktree, R. G. Hassell, P. Lahiry and R. A. Hegele, *Journal of Investigative Medicine* **58**, 700 (2010).
15. D. Houle, D. R. Govindaraju and S. Omholt, *Nature Publishing Group* **11**, 855 (2010).
16. A. Rzhetsky, D. Wajngurt, N. Park and T. Zheng, *Proceedings of the National Academy of Sciences* **104**, 11694 (2007).
17. N. Ghebranious, C. A. McCarty and R. A. Wilke, *Personalized Medicine* **4**, 175 (2007).
18. S. A. Pendergrass, K. Brown-Gentry, S. Dudek, A. Frase, E. S. Torstenson, R. Goodloe, J. L. Ambite, C. L. Avery, S. Buyske, P. Bůžková *et al.*, *PLoS Genetics* **9**, p. e1003087 (2013).
19. J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden and D. C. Crawford, *Bioinformatics* **26**, 1205 (2010).
20. K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).

# USING THE BIPARTITE HUMAN PHENOTYPE NETWORK TO REVEAL PLEIOTROPY AND EPISTASIS BEYOND THE GENE

CHRISTIAN DARABOS, SAMANTHA H. HARMON, JASON H. MOORE

*Institute for the Quantitative Biomedical Sciences, The Geisel Medical School at Dartmouth College,  
Lebanon, NH 03756, U.S.A.*

*E-mail: Christian.Darabos@dartmouth.edu*

With the rapid increase in the quality and quantity of data generated by modern high-throughput sequencing techniques, there has been a need for innovative methods able to convert this tremendous amount of data into more accessible forms. Networks have been a corner stone of this movement, as they are an intuitive way of representing interaction data, yet they offer a full set of sophisticated statistical tools to analyze the phenomena they model. We propose a novel approach to reveal and analyze pleiotropic and epistatic effects at the genome-wide scale using a bipartite network composed of human diseases, phenotypic traits, and several types of predictive elements (i.e. SNPs, genes, or pathways). We take advantage of publicly available GWAS data, gene and pathway databases, and more to construct networks different levels of granularity, from common genetic variants to entire biological pathways. We use the connections between the layers of the network to approximate the pleiotropy and epistasis effects taking place between the traits and the predictive elements. The global graph-theory based quantitative methods reveal that the levels of pleiotropy and epistasis are comparable for all types of predictive element. The results of the magnified “glaucoma” region of the network demonstrate the existence of well documented interactions, supported by overlapping genes and biological pathway, and more obscure associations. As the amount and complexity of genetic data increases, bipartite, and more generally multipartite networks that combine human diseases and other physical attributes with layers of genetic information, have the potential to become ubiquitous tools in the study of complex genetic and phenotypic interactions.

*Keywords:* Pleiotropy; Epistasis; Eye Diseases; Glaucoma; Network; GWAS; Human Phenotype Network; SNPs; Pathways;

## 1. Introduction

Genetic diseases and propensities have been at the center of the biomedical world for decades. From simple Mendelian diseases that obey the one-gene-one-phenotype paradigm, to complex genetic disorders, geneticists are working on developing novel methods to diagnose, treat, cure, and even prevent these diseases. At the center of prevention lie the information and education of patients on their personal genetic risk landscape. Because of the sheer number and complexity of genetic interactions within any given organism, and with its environment, genetic disorders and traits cannot be studied in isolation of one another or of external factors. The cascading effects of genomic mutations can extend to entire organisms, and having a global understanding of the ramifications of these mutations, including all the affected phenotypes and diseases, is becoming crucial. Two phenomena flawlessly illustrate the underlying complexity of genetic variations: pleiotropy, when a single mutation affects several traits, and epistasis, when multiple mutations in distant parts of the genome have synergetic, usually non-linear, effects on a single phenotype. From a system’s biology perspective, the preferred visualization methods for these interactions are networks of human diseases and traits. Networks offer an intuitive representation of phenotypic and genotypic interactions, while at the same time

allowing sophisticated quantitative statistical analysis of their intrinsic properties.

Although the concepts of epistasis and pleiotropy are over a 100 years old, they are widely under-appreciated due to their perceived rarity. State-of-the-art genome-wide association studies (GWAS) most often look for individual genes with large impacts on a single phenotype. The impact of genetic mutation cannot be studied in isolation, even if the attempt is to bridge the gap between a single gene and a single phenotype. Predictive elements, such as single nucleotides (SNPs), loci, genes, or entire biological pathways interact at all levels of granularity. The pervasiveness and strength of biomolecular interactions require a step back from reductionist biology and an acknowledgement of the importance of biological networks and pathways.

In this work, we propose to go beyond the gene as a unit of mutation, and use SNPs as a smaller unit, and biological pathways as a larger unit. We take a bird's eye view of the effect of genetic mutations on human phenotypes. It is often arduous to distinguish between certain types of pleiotropy and epistasis. The effect of a single mutation rippling through a pathway can be confused with the combined effect of distinct mutations. We therefore decide to study these phenomena in unison. We propose to use bipartite networks made of both phenotypes and predictive elements, constructed with GWAS data and other publicly available genetic databases. These networks allow us to identify the pleiotropic and epistatic interactions at the system's level. By studying several types of human phenotype networks (HPNs) based on predictive elements of different scales, we quantify the fundamental structural differences of these networks, as well as the amount of pleiotropic and epistatic information they contain. Finally, we magnify a specific phenotypic region of the HPN: the "glaucoma" region, which groups the disease and all its first and second neighbors. We offer a close up view of pleiotropic and epistatic interactions within a specific sub-network.

## 2. Background

In this section, we offer a cursory overview of the concepts of pleiotropy and epistasis. Furthermore, we define the fundamental concepts of HPNs, how they are constructed, and how they differ from one another (Section 2.2);

### 2.1. *Concepts of Pleiotropy and Epistasis*

Ludwig Platt and William Bateson first introduced the concepts of pleiotropy and epistasis, respectively, to explain observed inconsistencies in Mendelian inheritance and in the *one-gene-one-phenotype* paradigms.<sup>1,2</sup> To adapt with progress with genetics, the definition of pleiotropy has changed since it was first coined in 1910, and remains somewhat loose. A thorough history of pleiotropy in the past 100 year can be found in Stearns' 2010 review.<sup>3</sup> It refers to the general phenomenon in which a single gene dictates two or more seemingly unrelated phenotypic traits. In some cases, the definition is limited to a single mutation in a locus that affects multiple traits. It is however widely accepted that there is more than one type of pleiotropy. Grüneberg<sup>4</sup> in 1938 correctly distinguished between two major types he called "genuine" and "spurious" pleiotropy. Genuine pleiotropy refers to a single locus responsible for the production of two distinct gene products, whereas spurious involves a single gene product utilized in two different

ways. Furthermore, he distinguished a second form of spurious pleiotropy in which the single primary product initiates a cascade of events with different phenotypic consequences. Spurious pleiotropy can be said to perturb the biological pathways. Since then, more refined subdivisions have emerged. To help us navigate the various types of pleiotropy, Hodgking's survey offers classifications, descriptions, and examples of seven types of pleiotropy<sup>5</sup> (Table 1).

Table 1. A classification of different types of pleiotropy. Adapted from Hodgking's study<sup>5</sup>

Type	Situation
Artefactual	Adjacent but functionally unrelated genes affected by the same mutation
Secondary	Simple primary biochemical disorder leading to complex final phenotype
Adoptive	One gene product used for quite different chemical purposes in different tissues
Parsimonious	One gene product used for identical chemical purposes in multiple pathways
Opportunistic	One gene product playing a secondary role in addition to its main function
Combinatorial	One gene product employed in various ways, and with distinct properties, depending on its different protein partners
Unifying	One gene, or cluster of adjacent genes, encoding multiple chemical activities that support a common biological function

Actual genetic mechanisms of pleiotropy are extremely diverse. Genuine pleiotropy encapsulates pleiotropy at the mRNA-processing level, multiple or overlapping loci reading frames, alternative splicing, and multifunctional proteins, to mention only a few. Spurious pleiotropy covers single loci mutations that produce deviation in the gene product affecting other genes or regulatory elements located further down the biological pathways. Indeed, new gene products may promote or repress the expression of other genes. They may initiate alternate gene-gene and protein-protein interactions and alternate mRNA and microRNA productions, which may in turn affect seemingly unrelated phenotypes. Pleiotropic genes offer a unique insight into the complexities of biomolecular interaction networks.

In epistasis, on the other hand, the phenotypic contribution of a gene and its gene products depends on the specific genotype of a locus at a different genomic position. From the origin of the word, “standing upon”, we can derive the modern definition of epistasis, or epistatic gene effects, in which the expression of an allele at one locus *masks* the expression of an allele at another locus.<sup>6</sup> Epistasis is therefore usually the result of multiple genetic mutations at different loci. In this age of Genome-Wide Association Studies (GWAS), epistatic studies can be conducted at the genome level, quantitatively studying the masking and combined effect of single point mutations (SNPs).

Both epistasis and pleiotropy are exceptions to the one-gene-one-phenotype Mendelian rules of genetics. They are, however, far from being rare deviations.<sup>7</sup> Epistasis and pleiotropy are ubiquitous inherent properties of biological systems, and they are necessary byproducts of biomolecular networks.<sup>8</sup> Most phenotypes are the result of interactions between thousands of genes, as well as between genes and their environment. Because of the widespread connectivity within networks, the effects of a single mutation or variation can spread through thousands of gene-gene interactions, resulting in multiple phenotypes, or pleiotropy. The connections through which a variant's effects propagate define the molecular basis for epistatic interactions and how they translate into an observed phenotype. Because of their close relatedness, it is

not unreasonable to conclude that a similar set of quantitative tools can be applied to study both phenomena, sometimes simultaneously. In the present study, these tools are Bipartite Human Phenotype Networks.

## 2.2. *Human Phenotype Network (HPN)*

In recent years there has been a trend toward studying disease through network based analysis of various systems of connections between diseases. The result is the Human Disease Network (HDN). The nodes in the HDN represent human genetic disorders and the edges represent various connections between disorders, such as gene-gene or protein-protein interactions, to name only a few. The underlying connections of the HDN contribute to the understanding of the basis of disorders, which in turn leads to a better understanding of human disease.

One study by Goh, *et al.*,<sup>9</sup> explored the Human Disease Network (HDN), limiting its analysis to the genes shared by different diseases. Another study by Li *et al.*<sup>10</sup> traced the SNPs connecting disease traits. In 2009, Silpa Suthram *et al.*<sup>11</sup> found that when diseases were analyzed by disease-related mRNA expression data in combination with the human protein interaction network, there were significant genetic similarities between certain diseases, and some of the correlated diseases shared drug treatments, as well. This could help us target certain genes for treatment. In 2009, Barrenas *et al.*<sup>12</sup> further studied genetic architecture of complex diseases by doing a GWAS, and found that complex disease genes are less central than the essential and monogenic disease genes in the human interactome.

GWAS identify common genetic variants, such as SNPs, found in the genotype of different individuals in association with phenotypic traits. Using GWAS data, we extend the HDN to include not only diseases, but also general *phenotypes*, encompassing behavioral traits and physical attributes, such as hair color, and explore large portions of non-coding variations in the human genome. We call this more complete representation the Human Phenotype Network.<sup>13</sup> We rely on the catalog of published GWAS maintained by the National Human Genome Research Institute (NHGRI) at the National Institute of Health (<http://www.genome.gov/gwastudies/>) as a primary source of phenotypic data. It aggregates studies that report SNP(s)-to-phenotype(s) and gene(s)-to-phenotype(s) associations. The NHGR catalog used in this study, downloaded in June 2013, reports 646 phenotypes associated with 2,000+ genes and 6,000+ SNPs.

Over 90% of risk-associated SNPs (raSNPs) identified by the GWAS fall outside of coding regions,<sup>14</sup> stressing the requirement for a more global assessment of phenotypic associations. In this work we explore methods of building the HPN that go beyond previously mentioned gene-centric HDN approaches. An interesting side-effect of all the methods presented below is that before obtaining a HPN, the algorithm produces a bipartite network (see Section 3), which is the property that allows us to study the pleiotropic and epistatic information in the genetic association data. The HPN is obtained by projecting the bipartite network onto the phenotype space.

The following sections present our methods for building the HPNs based on different predictive elements. We start at the smallest predictive element, the SNP, then move on to SNP clusters, to genes, and finally to complete biological pathways. These offer varying density

of the information contained with both the bipartite network and the projected HPN.

### 2.2.1. Genetic Variations based HPN

For each phenotype in the catalog (Fig. 1, Step 1), we define its risk-associated variome (RAV) as the complete set of its associated raSNPs (Step 2). To address the low genomic coverage provided by GWAS, we associate each raSNP with all SNPs found in *linkage disequilibrium* (ldSNPs) using the HapMap project data<sup>15</sup> (Step 3). SNPs in linkage disequilibrium form clusters of variants that statistically tend to appear in the same patient.<sup>16</sup> The HapMap project aims at building a repository of describing the common patterns found in human genetic variations (<http://hapmap.ncbi.nlm.nih.gov/>). The resulting imputed variome (iRAV) will allow us to establish connections between diseases/traits that share blocks, i.e. that have overlapping iRAVs (Step 4).

**iRAV-based HPN.** In a previous study, we presented a model of iRAV-based HPN which included the phenotype-to-raSNPs association from GWAS, and added the HapMap project data to build clusters of variants for each phenotype (iRAVs).<sup>13</sup> Phenotypes in the iRAV-HPN are linked when they share overlapping iRAVs. The algorithm (in Figure 1) produces a bipartite network of phenotypes and iRAVs.

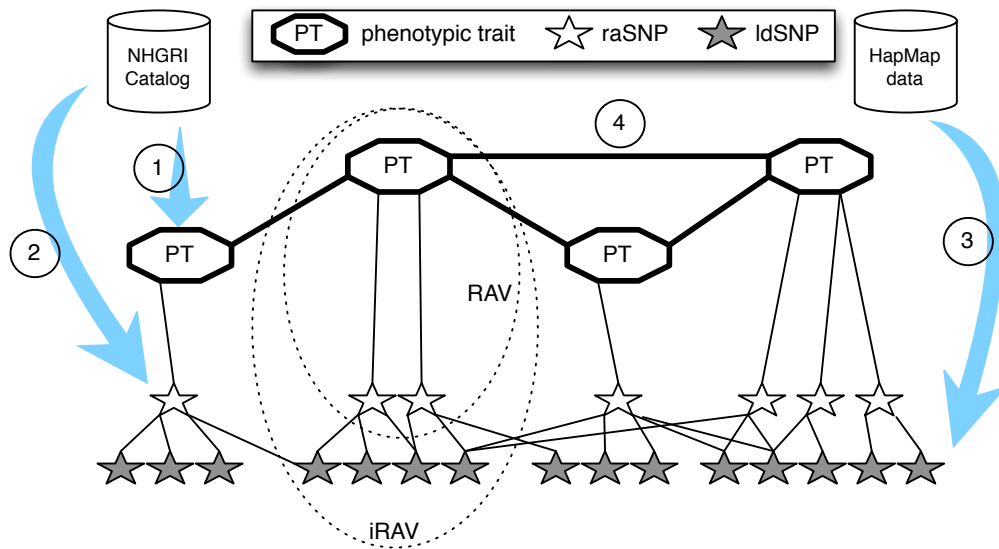


Fig. 1. Step-by-step description of the method to obtain the HPN. The circled numbers correspond to the steps of the method description above.

**RAV-based HPN.** Linking phenotypes that share at least one raSNP, we build the RAV-HPN. This approach is similar to that of Li *et al.*<sup>10</sup> The phenotypes are linked based only on shared risk variants, i.e. overlapping RAVs, not including ldSNPs/iRAVs. This approach produces a HPN that is less dense than the iRAV-HPN, identifying fewer phenotype associations. The algorithm is similar to that described in Figure 1, omitting Step 3.



### 2.2.2. Gene-based HPN

Leveraging the gene(s)-to-phenotype(s) associations contained in the GWAS catalogue, we construct the gene-only based HPN (gHPN). Indeed, the GWAS catalog reports for each phenotype both the associated and the mapped genes in which the SNPs fall. This approach is analogous to that of Goh *et al.*<sup>9</sup> To increase the genetic coverage of each phenotype, we use the Broad Institute's GeneCruiser ([genecruiser.broadinstitute.org](http://genecruiser.broadinstitute.org)) to identify the gene closest to a SNP, or for which SNPs fall in a known regulatory region. If this gene is not already associated with the SNPs phenotype, we include it in the study. This method increases the number of genes by 138, from 2,339 genes to 2,477. The algorithm is similar to that shown in Figure 1, omitting Step 3, and white star symbols are now genes, not raSNPs.

### 2.2.3. Biological Pathways based HPN

Expanding on the gHPN, we build a pathway-based HPN (pHPN).<sup>17</sup> Biological pathways represent elaborate series of cascading biochemical reactions occurring within the cell, and possibly receiving external signals.<sup>18</sup> Pathways govern all major cellular functions, such as cell cycle, cell respiration, and apoptosis (programmed cell death). Biochemical compounds, (e.g. nucleic acids, proteins, complexes and small molecules) participating in reactions form a network of biological processes and are grouped into pathways. KEGG Pathway ([kegg.jp](http://kegg.jp)) is an open-access collection of manually curated and peer-reviewed pathway database, containing the structured information about the elements, enzymes, and genes (via their gene products) within many known pathways.

Relying on the gene(s)-to-phenotype(s) data used to construct the gHPN, genes were further linked to enriched pathways using KEGG. By building these associations, we were able to link phenotypes associated with genes involved in the same pathways in the pHPN. The algorithm is analogous to that in Figure 1, except that the white star symbols represent genes, and the grey stars are pathways.

## 3. Pleiotropy and Epistasis in the Bipartite HPNs

The HPN resulting from either method described in Section 2.2 can be represented as a mathematical object: a graph.<sup>19</sup> In this work, the terms “graph” and “network” are used interchangeably. Formally, a network is a collection of nodes and edges connecting them. The degree,  $k$ , of a node is the number of edges incident upon the node. The average degree of the network is the average of all  $k$ . The degree distribution function,  $P(k)$ , of the network describes the fraction of nodes within the network with degree  $k$ . The clustering coefficient (CC) of a network measures the degree to which nodes tend to form closely knit communities with a higher than average connectivity.<sup>20</sup> The CC of networks found in nature, in particular social and biological networks, show a higher degree of clustering than that observed in randomized networks of identical size. The average path length of a network (APL) represents the average of the minimum number of edges separating any two vertices. Finally, the network's diameter is defined as the greatest distance between any pair of vertices.

In our study, we start by building a bipartite network,<sup>21</sup> consisting of two disjoint sets of nodes. The nodes are connected in such a way that the nodes of one set will have no

connections between them, but can only be connected to nodes of the other set. The use of a bipartite network is natural when dealing with two different types of data sets (Figure 2b), in our case phenotypes (e.g. the rectangles) and RAVs, iRAVs, genes, or pathways (e.g. the circles). This type of network gives us three distinct degree distributions, one for each projection, and one for the bipartite network. Each degree distribution shows how many links each node has. Nodes in a projection of a bipartite network are connected if they share at least one node in the other group. This gives us the ability to see the interactions within each set.

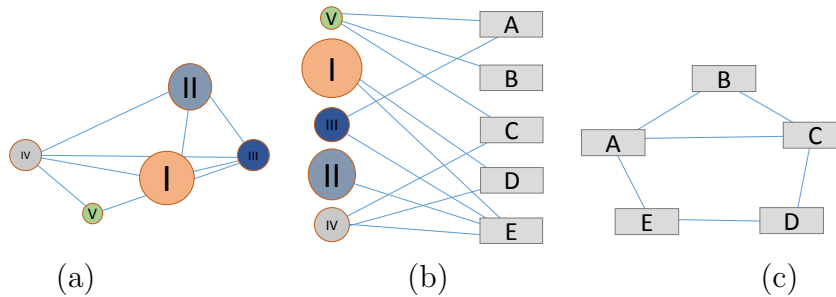


Fig. 2. Bipartite Network schematic. A bipartite network (b) made of 2 data sets the “circles”, and the “rectangles”. Projections in the “circle” space (a) and in “rectangle” space (c).

The data from the bipartite network can be projected onto either data space (Figure 2a,c). In both cases, the nodes are connected to one another through a vertex of the other space. By ignoring the different types of data, all network properties described above remain valid on the bipartite network (as a single data set network) and on either projection. We illustrate the iRAV-HPN resulting from the projection onto the phenotype space in Figure 3. In the context of this study, the qualitative nature of the projected HPN does not contain much information about the phenotypic pleiotropy and epistasis. Therefore, we only show one example of projected HPN to give a sense of the complexity of the data and the necessity for quantitative methods.

#### 4. Characterizing and Quantifying Pleiotropy and Epistasis in the HPNs

Early studies have made use of network theory in studying both pleiotropy and epistasis. Global statistical properties of networks, such as the “shape” of the degree distribution and an above average CC place gene expression networks in the small-world<sup>20</sup> or scale-free<sup>19</sup> family of networks.<sup>22</sup> This indicates that most of the nodes (genes) in the network are of a low degree  $k$ . However, a small minority of the vertices are highly connected (hubs). Put in the context of the present work, a few genes have extensive pleiotropic/epistatic effects, but most genes only affect/are affected by a small number of phenotypes. The quantitative structural analyses of the protein interaction networks of model organisms have highlighted the importance of properties such as the diameter and the APL. Li *et al.*<sup>23</sup> determined that the diameter was  $\sim 4 - 5$  edges, meaning that each gene in the genomes studied affected on average four or five proteins. This finding also corroborates the conjecture that pleiotropy and epistasis are confined to genomic modules, and cannot generally affect any pairs/set of loci in the genome.<sup>24</sup>



level, we will study the pHPN and use biological pathways to quantify pleiotropy and epistasis. Admittedly, these interpretations of pleiotropy and epistasis may somewhat stray from the commonly accepted definitions, but they are in line with the loose nature of the phenomena, where both have sub-types that relate to all degrees of granularity.

Relying on the data in the bipartite HPNs, we calculate the number of phenotypes connected to each predictive element. We use the average connectivity of the predictive element as a proxy for measuring the global pleiotropy (Table 2). We also present the degree distribution of the predictive element subset, showing the effect of pleiotropy at each predictive element level (Figure 4). Inversely, the average epistatic effect of predictive elements on phenotypes can be calculated as the average degree of the phenotype subset in the bipartite HPN (Table 2). The degree distribution of the phenotype subset conveys the distribution of epistatic effects that different predictive elements have on the phenotypes (Figure 4).

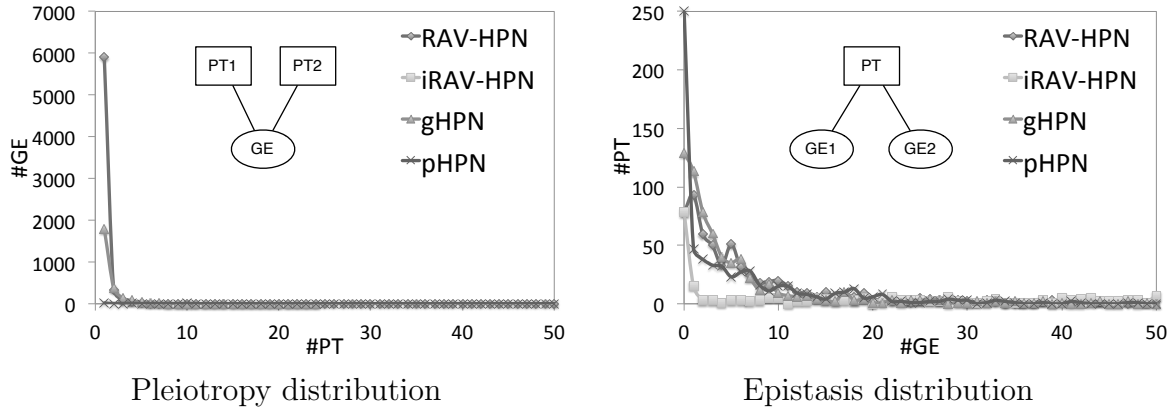


Fig. 4. Pleiotropy and Epistasis Distributions. The pleiotropy distributions shows the distribution of the number of phenotypes (PT) for each predictive element, i.e. the distribution of predictive elements influencing multiple phenotypes (see inset). The epistasis distribution shows the number of predictive elements for each phenotype, i.e. the distribution of phenotypes ruled by multiple predictive element (see inset).

Both pleiotropy and epistasis distributions are right-skewed with a heavy tail, which denotes the presence of hubs. The pleiotropy distributions show that most predictive elements only influence a few phenotypes, however, a small minority of predictive elements influence a large number (50+) phenotypes. Similarly, the epistasis distribution depicts that most phenotypes can be associated with only a few predictive elements. However, a small number of phenotypes rely on the signaling of a large number of predictive elements. Although somewhat simplistic, these results are, for all models, in line with the findings of Featherstone *et al.*<sup>22</sup> We acknowledge that the manner in which the average effects are computed may capture more than just the pleiotropic and epistatic effects. However, these results, due to their ubiquity, reflect a biologically plausible property of the system. We run a full array of quantitative statistical analysis of the different HPNs, including the average pleiotropic and epistatic effects (Table 2).

As the models increase in granularity, the networks become denser with more edges, decreasing APL and diameter. This is to be expected: as the network has more connections,

Table 2. Quantitative Properties of the different HPNs. (PT : phenotype, GE: predictive element.)

	RAV-HPN	iRAV-HPN	gHPN	pHPN
LCC size (%nodes)	295 (45%)	401 (62%)	430 (67%)	396 (61%)
#edges	932	2845	2556	40K
avg. degree / weighted	6.31 / 10.03	14.19 / 37.54	11.88 / 16.85	204.1 / 497.6
APL / diameter	3.7 / 10	2.96 / 8	2.96 / 6	1.48 / 3
avg. CC	0.58	0.59	0.57	0.79
modularity / communities	0.62 / 26	0.55 / 24	0.49 / 10	0.10 / 4
isolate vertices	351	245	216	250
<b>avg. pleiotropy (#PT/#GE)</b>	1.12	1.12	1.58	27.06
<b>avg. epistasis (#GE/#PT)</b>	11.11	285.64	6.07	5.69

the distance between nodes decreases. The values agree with Li *et al.*<sup>23</sup> findings. The above average CC and the shape of the degree distributions (not shown here for space reasons) put the HPNs in the scale-free region of the network topology spectrum. Additionally, we note that the modularity and number of communities drops with increasing granularity and network density. The number of isolated nodes provides an insight into how many phenotypes have no detected genetic connection to any other phenotype. Finally, we see that the average pleiotropy remains relatively constant until we look at the pHPN, which biological pathways tend to affect  $\sim 27$  phenotypes in average. Otherwise, predictive elements do not in general impact more than 1-2 phenotypes. This proves the necessity to apply a biologically relevant filter to the pHPN in order to extract the “backbone” of the network, containing the most relevant genetic influences.<sup>17</sup> The average epistatic effect is also reasonably steady, except when ldSNPs are included in the iRAV-HPN. This is due to the fact that now both raSNPs and ldSNPs are directly associated to the phenotypes.

### 5. Clinical Implications: the Example of Glaucoma

As previously stated, each HPN differs in terms of the number of edges branching from each phenotype node. Moving from the gHPN to the pHPN provides a great deal more information, but the network itself becomes extremely complex and difficult to analyze visually. The pHPN can help to explain the shared etiology of glaucoma and other diseases by revealing a substantial number of interactions unseen in the gHPN. Ultimately, studying predictive elements from a global perspective, using networks, could contribute to novel discoveries in pleiotropic drug therapies.

The HPNs confirm well-known interactions, such as between glaucoma and blood pressure (BP). Studies have linked the two for years and drugs used to treat glaucoma, such as beta-blockers and alpha-adrenergic agonists,<sup>25</sup> are known to affect BP. In fact, patients with cardiovascular problems are advised against taking beta-blockers, a treatment for the high intraocular pressure (IOP) associated with glaucoma, because of its effect on heart rate and BP.<sup>26</sup> Moreover, many studies have shown that BP and ocular perfusion are important factors in the pathogenesis of glaucoma. For example, studies have linked increases in blood pressure to slight increases in IOP. Going further, the “Blue Mountains Eye Study” found that systemic hypertension was significantly associated with an increased risk of primary open-angle glaucoma (POAG), independent of the effect of BP on IOP. Systemic hypertension was the

greatest risk factor for POAG.<sup>26</sup> Blood pressure is a first neighbor to Glaucoma in the pHPN, suggesting the validity of the model. They are linked by the umbrella *pathways in cancer*. Diabetes mellitus is another well-documented disease known to interact with glaucoma.<sup>27</sup> Type 1 diabetes is a direct neighbor and Type 2 diabetes is a second (indirect) neighbor. Type 1 diabetes and glaucoma are linked by the *cell cycle* and *HTLV-I infection* pathways. Type 2 diabetes and glaucoma share common gene: CDKN2B-AS. In the gHPN, on the other hand, Type 2 diabetes is a first neighbor, but Type 1 diabetes and blood pressure are only second neighbors to glaucoma. Additionally, the pHPN allows us to see connections that are not included in the gHPN, which could lead to new advances in treatments for the linked diseases. For instance, Alzheimers disease is a second neighbor of glaucoma. Both are neurodegenerative diseases and their similarities have recently begun to receive significant attention. Inoue *et al.* maintain that elevated levels of biomarkers for Alzheimers are more often found in patients with open-angle glaucoma (OAG) than in patients with cataracts.<sup>28</sup> In addition, Alzheimer's and OAG share pathways such as *cell death mechanisms (apoptosis)*, *reactive oxygen species (ROS) production*, *mitochondrial dysfunction* and *vascular abnormalities*.<sup>29</sup> Apoptosis of the neural ganglia cells is a major issue in glaucoma. In the gHPN, the link between Glaucoma and Alzheimers disease is not readily apparent by looking at the graph – it becomes a third neighbor. Another interesting link is to the “smoking behavior” phenotype, although this is only readily apparent in the pHPN where it is a first neighbor to glaucoma. The two share the umbrella *pathways in cancer*. Association studies have shown that smoking behavior is correlated with central corneal thickness in OAG and might also be a risk factor for POAG.<sup>30</sup>

## 6. Conclusions & Future Work

The study of genetic diseases is progressing at an unprecedented pace, thanks to modern high-throughput sequencing technology and to the development of modeling techniques at the crossroads of bioinformatics and mathematics. Bipartite HPN models are capable of leveraging the massive amount of GWAS and other readily-accessible genetic data, and collapsing the information into a single, manageable source. The projection of the HPN helps analyze phenotypic interactions.<sup>13,17</sup> The overall structure of the connections between the layers of the bipartite HPN, on the other hand, allows us to estimate in a quantitative manner the pleiotropic and epistatic effect at a global level, for multiple types of predictive elements. Finally, by magnifying regions of the HPN, we are able to highlight previously documented phenotypic interactions, supported by genes and biological pathways evidence as a proof of concept. The bipartite HPNs are flexible, scalable, and intuitive models. HPNs are potentially useful to study phenotypic links, as well as uncover novel pleiotropy and epistasis effects at the single variation level, at the gene level, and all the way to the biological pathway. Future work will involve collapsing the multiple HPNs into an aggregated model. This step will however require the information to be filtered in a biologically sensible manner. Further refinements of the model will include the detection of different types of pleiotropy and epistasis. Finally, we are working on statistical and cross-validation approaches to validate the E&P significance.

## Acknowledgments

Financial supported by NIH grants R01 EY022300, LM009012, LM010098, AI59694.

## References

1. W. Bateson, *Science* **26**, 649 (Nov 1907).
2. L. Plate, *Festschrift für R. Hertwig* **II**, 537 (1910).
3. F. W. Stearns, *Genetics* **186**, 767 (Nov 2010).
4. H. Gruneberg, *Proceedings of the Royal Society of London. Series B, Biological Sciences* **125**, pp. 123 (1938).
5. J. Hodgkin, *Int J Dev Biol* **42**, 501 (1998).
6. A. J. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin and W. M. Gelbart, *Introduction to Genetic Analysis. 7th edition.* (W. H. Freeman, 2000).
7. J. H. Moore, *Hum Hered* **56**, 73 (2003).
8. A. L. Tyler, F. W. Asselbergs, S. M. Williams and J. H. Moore, *Bioessays* **31**, 220 (Feb 2009).
9. K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabasi, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).
10. H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li and Y. A. Lussier, *Journal of the American Medical Informatics Association : JAMIA* **19**, 295 (January 2012).
11. S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie and A. J. Butte, *PLoS Comput Biol* **6**, p. e1000662 (02 2010).
12. F. Barrenas, S. Chavali, P. Holme, R. Mobini and M. Benson, *PLoS ONE* **4**, p. e8090 (11 2009).
13. C. Darabos, K. Desai, R. Cowper-Sal-lari, M. Giacobini, M. Lupien and J. H. Moore, Inferring human phenotype networks from genome-wide genetic associations, in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics - 11th European Conference, EvoBIO 2013, Vienna, Austria, April 3-5, 2013. Proceedings*, eds. M. Giacobini, L. Vanneschi and W. S. BushLecture Notes in Computer Science (Springer, to appear, 2013).
14. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proc Natl Acad Sci U S A* **106**, 9362 (Jun 2009).
15. C. International HapMap, *Nature* **437**, 1299 (Oct 2005).
16. D. S. Falconer and T. F. C. Mackay, *Introduction to Quantitative Genetics (4th Edition)* (Prentice Hall, February 1996).
17. C. Darabos, M. J. White, B. E. Graham, D. Leung, S. Williams and J. H. Moore, *TBC2013 - The 3rd Annual Translational Bioinformatics Conference* (in press).
18. C. H. Schilling, S. Schuster, B. O. Palsson and R. Heinrich, *Biotechnol Prog* **15**, 296 (May-Jun 1999).
19. M. Newman, *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).
20. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
21. T. Zhou, J. Ren, M. c. v. Medo and Y.-C. Zhang, *Phys. Rev. E* **76**, p. 046115 (Oct 2007).
22. D. E. Featherstone and K. Broadie, *Bioessays* **24**, 267 (Mar 2002).
23. R. Li, S.-W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen and G. A. Churchill, *PLoS Genet* **2**, p. e114 (Jul 2006).
24. J. J. Welch and D. Waxman, *Evolution* **57**, 1723 (Aug 2003).
25. B. Chae, T. Cakiner-Egilmez and M. Desai, *Insight* **38**, 5 (Winter 2013).
26. V. P. Costa, E. S. Arcieri and A. Harris, *Br J Ophthalmol* **93**, 1276 (Oct 2009).
27. K. S. Oswal, R. R. Sivaraj, P. I. Murray and P. Stavrou, *BMC Res Notes* **6**, p. 167 (Apr 2013).
28. T. Inoue, T. Kawaji and H. Tanihara, *Invest Ophthalmol Vis Sci* (Jul 2013).
29. D. Wang, Y. Huang, C. Huang, P. Wu, J. Lin, Y. Zheng, Y. Peng, Y. Liang, J.-H. Chen and M. Zhang, *BMC Ophthalmol* **12**, p. 59 (2012).
30. J. A. Ghiso, *J Glaucoma* **22 Suppl 5**, S36 (Jun-Jul 2013).

**ENVIRONMENT-WIDE ASSOCIATION STUDY (EWAS) FOR TYPE 2 DIABETES IN THE  
MARSHFIELD PERSONALIZED MEDICINE RESEARCH PROJECT BIOBANK**

**MOLLY A. HALL**

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State  
University, 512 Wartik Lab, University Park, PA 16802, USA  
Email: mah546@psu.edu*

**SCOTT M. DUDEK**

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State  
University, 512 Wartik Lab, University Park, PA 16802, USA  
Email: sud23@psu.edu*

**ROBERT GOODLOE**

*Center for Human Genetics Research, Vanderbilt University, Nashville TN, 37232, USA  
Email: robert.goodloe@chgr.mc.vanderbilt.edu*

**DANA C. CRAWFORD**

*Center for Human Genetics Research, Vanderbilt University, Nashville TN, 37232, USA  
Email: crawford@chgr.mc.vanderbilt.edu*

**SARAH A. PENDERGRASS**

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State  
University, 503 Wartik Lab, University Park, PA 16802, USA  
Email: sap29@psu.edu*

**PEGGY PEISSIG**

*The Marshfield Clinic  
Marshfield, WI, USA  
Email: Peissig.Peggy@securityhealth.org*

**MURRAY BRILLIANT**

*The Marshfield Clinic  
Marshfield, WI, USA  
Email: BRILLIANT.MURRAY@mcrf.mfldclin.edu*

**CATHERINE A. MCCARTY**

*Essentia Institute of Rural Health, Duluth, MN, USA  
Email: CMcCarty@eirh.org*

**MARYLYN D. RITCHIE**

*Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State  
University, 512 Wartik Lab, University Park, PA 16802, USA  
Email: Marylyn.ritchie@psu.edu*

Environment-wide association studies (EWAS) provide a way to uncover the environmental mechanisms involved in complex traits in a high-throughput manner. Genome-wide association studies have led to the discovery of genetic variants associated with many common diseases but do not take into account the environmental component of complex phenotypes. This EWAS assesses the comprehensive association between environmental variables and the outcome of type 2 diabetes (T2D) in the Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP). We sought replication in two National Health and Nutrition Examination Surveys (NHANES). The Marshfield PMRP currently uses four tools for measuring environmental exposures and outcome traits: 1) the PhenX Toolkit includes standardized exposure and



phenotypic measures across several domains, 2) the Diet History Questionnaire (DHQ) is a food frequency questionnaire, 3) the Measurement of a Person's Habitual Physical Activity scores the level of an individual's physical activity, and 4) electronic health records (EHR) employs validated algorithms to establish T2D case-control status. Using PLATO software, 314 environmental variables were tested for association with T2D using logistic regression, adjusting for sex, age, and BMI in over 2,200 European Americans. When available, similar variables were tested with the same methods and adjustment in samples from NHANES III and NHANES 1999-2002. Twelve and 31 associations were identified in the Marshfield samples at  $p < 0.01$  and  $p < 0.05$ , respectively. Seven and 13 measures replicated in at least one of the NHANES at  $p < 0.01$  and  $p < 0.05$ , respectively, with the same direction of effect. The most significant environmental exposures associated with T2D status included decreased alcohol use as well as increased smoking exposure in childhood and adulthood. The results demonstrate the utility of the EWAS method and survey tools for identifying environmental components of complex diseases like type 2 diabetes. These high-throughput and comprehensive investigation methods can easily be applied to investigate the relation between environmental exposures and multiple phenotypes in future analyses.

## 1. Introduction

Computational methods to assess environmental exposures are essential to elucidate the complex nature of common human phenotypes. Genome-wide association studies (GWAS) have allowed for greater understanding of the genetic component of complex traits and identification of numerous loci associated with these traits [1]. They have provided a high-throughput approach for comprehensive testing of variants across the genome. However, this approach fails to consider the richly diverse and complex environment with which humans interact throughout the life course.

While GWAS have uncovered thousands of single nucleotide polymorphisms (SNPs) associated with disease, much remains unclear about the heritability and mechanisms that lead to common, complex human diseases [1,2]. It is likely that environmental exposure greatly impacts the genetic and cellular systems at play for many complex traits [2]. Environment-wide association studies (EWAS) [3] provide a method to test a variety of exposures across the human environment in a high-throughput, unbiased manner, much like GWAS tests for genetic effects. The utility of the EWAS approach was demonstrated for type 2 diabetes (T2D) using an array of laboratory measurements to identify a diverse number of exposures associated with T2D [3]. Such comprehensive laboratory measurements are rare and only assess exposures at a fixed time point without consideration of the various exposures throughout an individual's lifetime. Thus, there is a need to evaluate comprehensive and standardized survey tools that enable assessment of exposures and lifestyle choices over time and comparison of results across multiple studies.

The PhenX (consensus measures for Phenotypes and eXposures) toolkit (<https://www.phenxtoolkit.org/>) was developed as a resource for collecting standardized measures of phenotypes and environmental exposures [4]. Measures are available across 27 domains covering alcohol, tobacco, and other substance use; demographics; mental health; environmental exposures; diet; and disease, among others. In addition to providing information on traits, many of these measures can be used to ascertain information on environment, lifestyle, and environmental exposures. Other valuable resources for environmental measures include 1) the Measurement of a Person's Habitual Physical Activity, a questionnaire measuring a person's work, leisure, and sport activity level [5] (Baecke), and 2) the Dietary History Questionnaire (<http://riskfactor.cancer.gov/DHQ/>), a food frequency questionnaire [6,7] (DHQ).

Electronic health records (EHR) are a growing resource for measuring health outcomes in individuals, as they contain vast amounts of medical data including records of diagnoses, procedures, and clinical laboratory measurements [8]. These data can be used, with electronic

algorithms, to systematically define cases and controls for numerous phenotypes of interest, such as type 2 diabetes. The Electronic Medical Records and Genomics (eMERGE) Network combines EHR data from sites across the United States and currently utilizes electronic phenotyping algorithms for over a dozen phenotypes [9]. The Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP) [10], part of the eMERGE Network, is one site currently employing EHR phenotyping as well as the PhenX Toolkit, the Measurement of a Person's Habitual Physical Activity (Beacke), and the Dietary History Questionnaire (DHQ). Taken together, the PMRP is a rich phenotypic resource for genomic and environmental association analyses to dissect the architecture of complex traits.

Here, we present the results of an EWAS for type 2 diabetes using survey questions from the PhenX Toolkit, DHQ, and Beacke surveys from the Marshfield PMRP. To seek replication of these results with similar survey questions when available, we used data from the National Health and Nutrition Examination Surveys (NHANES) [11]. To the authors' knowledge, this is the first EWAS performed using EHR data. Environment-wide association studies provide a methodology to test environmental measures in a comprehensive, high-throughput manner. Integration of EWAS with phenome-wide association studies (PheWAS) [12-14] and genome-wide association studies (GWAS) [1] will further elucidate the complex interplay of gene and environment in common traits as well as the ways in which exposures modulate pleiotropy. Using multiple exposure and outcome variables to assess environment and lifestyle factors using EWAS will provide a richer understanding of the architecture of complex traits.

## 2. Methods

### 2.1. Marshfield PMRP and Type 2 Diabetes Case Identification

The Marshfield PMRP is a population based biobank with ~20,000 subjects, aged 18 years and older, enrolled in the Marshfield Clinic healthcare system in central Wisconsin [10]. DNA, plasma, and serum samples are collected at the time the enrollee completes a written informed consent document, with allowance for ongoing access to the linked electronic health records (EHR). PMRP participants also complete questionnaires, including responses regarding smoking history, occupation, physical activity, diet, and a variety of other PhenX measures. A subset of the Marshfield PMRP subjects completed the PhenX survey, the DHQ, and/or the Measurement of a Person's Habitual Physical Activity (Table 1).

The NHGRI funded eMERGE network (Electronic Medical Records and Genomics) has implemented robust electronic phenotyping algorithms to select cases and controls for a number of different phenotypes/outcomes [9]. Using an algorithm developed by eMERGE [15], T2D patients were diagnosed by their records from the Marshfield EHR. The Marshfield samples were originally selected for eMERGE based on their cataract case-control status; however, this is an example of the reusability of biobank samples for additional traits. T2D cases were defined as having the following in their EMR: a T2D ICD-9 medical billing code, information about insulin medication, abnormal glucose or HbA1c levels, or more than two diagnoses of T2D by a clinician. All T2D cases with an ICD-9 code for T1D were removed from further analyses. All control subjects had to have at least 2 clinical visits, at least one blood glucose measurement, normal blood glucose or HbA1c levels, no ICD-9 codes for T2D or any related condition, no history of being on insulin or any diabetes related medication, and no family history of T1D or T2D.

**Table1. Marshfield Type 2 Diabetes Case/Control Sample Size for Each Questionnaire**

	Questionnaire	Total Sample Size	# Cases T2D	# Controls
<b>Total</b>	PhenX	2,243	433	1,810
	DHQ	2,606	559	2,047
	Activity	2,571	552	2,018
<b>Male</b>	PhenX	898	204	694
	DHQ	1,051	260	791
	Activity	1,035	257	778
<b>Female</b>	PhenX	1,345	229	1,116
	DHQ	1,555	299	1,256
	Activity	1,535	295	1,240
<b>Age</b>	All	> 50		
<b>Ancestry</b>	All	European		

## **2.2. Environmental Variable Measurements**

### **2.2.1 Phenx Toolkit**

The PhenX Toolkit ([www.phenxtoolkit.org](http://www.phenxtoolkit.org)) was accessed to develop a self-administered questionnaire to assess environmental and lifestyle factors. Some of the PhenX measures were chosen because of the potential for gene/environment associations with age related cataract - which is a primary disease of interest for PMRP (smoking, alcohol, ultraviolet light exposure), some were chosen because of the potential for validation against prior PMRP questionnaire data and medical history information (demographics, physical activity, family history of heart attack, history of stroke) and the rest were chosen because of the potential for future research and cross-site collaborations (hypomania/mania symptoms, hand dominance) within the network funded through administrative supplements to collect PhenX measures. The time to complete the questionnaire ranged from 20 to 40 minutes in pre-testing, depending on how many questions were logical skips. The 32-page self-administered questionnaire was mailed to all eligible subjects with a cover letter and return address envelope. A second mailing was employed to increase the response rate. Subjects were offered \$10 for their time to complete the questionnaire.

PhenX survey data were entered and merged with prior PMRP questionnaire information from the Marshfield Clinic electronic medical record. For validation purposes, the electronic medical record was considered to be the gold standard where possible. Two hundred fifty-five measures from the PhenX Toolkit were included for our analysis. Questions included a range of topics from the following classes: alcohol use, smoking, demographics, depression, mania, activity, residential environment, and UV exposure.

### 2.2.2. *Diet History Questionnaire*

Food frequency questionnaires (FFQs) are widely used to assess dietary intake in epidemiologic studies because they are more representative of usual intake and less expensive to implement than other methodologies including weighed food records and 24-hour dietary recalls because they are usually self-administered. Inclusion of aids to estimate portion sizes is essential to improve the accuracy and validity of FFQs [7]. Self-administered food frequency questionnaires (FFQ) are available on approximately 2/3 of the PMRP cohort to quantify usual dietary intake of all major nutrients. The selected FFQ, the Diet History Questionnaire (DHQ) (<http://riskfactor.cancer.gov/DHQ/>), was developed by researchers at the National Cancer Institute (NCI) and has been shown to be superior to the commonly used Willett FFQ and similar to the Block FFQ for estimating absolute nutrient intakes [7]. All three FFQs produce similar results after statistical adjustment for total energy intake. The list of foods and portion sizes on the DHQ was developed from nationally representative data, the USDA's 1994-1996 Continuing Survey of Food Intakes by Individuals, and is therefore most appropriate for use with this study population. The DHQ comprises 124 separate food items and asks about portion sizes for most foods. In addition, there are 10 questions about nutrient supplement intake. The DHQ was printed and scanned by National Computer Systems as has been done for all recent studies conducted at the NCI using the DHQ. The completed DHQ was mailed to National Computer Systems for scanning. After scanning, the data from the questionnaires are stored in ASCII format and then uploaded into the nutrient analysis software package. Diet\*Calc software, available from the National Institutes of Health, is used for the nutrient analyses of the DHQ data (<http://riskfactor.cancer.gov/DHQ/dietcalc/>). This is the software package that was used for analysis of the DHQ for the Eating at America's Table Study. The DHQ is mailed to participants with their appointment reminders so that they can complete it prior to their appointment to save them time. The Research Project Assistants reviews all DHQs to ensure that they have been completed. Fifty-six measures of dietary intake were assessed for this EWAS that covered the following domains: vitamin, fat, protein, carbohydrate, fiber, cholesterol, caloric, grain, vegetable, caffeine, and alcohol intake.

### 2.2.3 *Measurement of a Person's Habitual Physical Activity*

As with measurement of dietary intake for epidemiologic studies, there are a number of different validated tools that have been used in the past. The agreement between physical activity questionnaire and gold standard tends to be somewhat lower than for dietary intake, but is reasonable for ranking relative activity levels in groups. The researchers preferred to use a previously developed physical activity assessment tool to allow comparison with results from other study populations. Requirements of the selected tool included: 1) self-administered, 2) previously validated, and 3) validated for use in a similar study population across a range of ages. The selected physical activity questionnaire, the ARIC/Baecke questionnaire, is self-administered, validated for use in both men and women, and currently being used in a large, prospective study in the US [16]. The questionnaire has been shown to have high reliability and accurate assessment of both high intensity activity and light intensity activity such as walking. It comprises 16 questions and generates three indices of activity: 1) a work index, 2) a sport index, and 3) a leisure-time index. This one-page self-administered physical activity questionnaire is mailed along with appointment reminders and the Diet History Questionnaire (DHQ). Information from the completed physical activity questionnaires are entered twice into a

Microsoft Access database. The two entries are compared to ensure accuracy of the data entry. The three physical activity indices (work, sport, and leisure-time) are calculated and the data merged with anthropometric, dietary, and demographic data for subsequent analyses.

#### *2.2.4. National Health and Nutrition Examination Surveys (NHANES)*

NHANES III Phase 2, conducted between 1991-1994, and NHANES 1999-2002 measures the health and nutritional habits of participants by collecting medical, dietary, demographic, laboratory, lifestyle, and environmental exposure data using questionnaire and laboratory measures. The data of NHANES were collected by the National Center on Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). All participants were consented by the CDC at the time of the survey and sample collection.

To seek replication of the Marshfield results, we identified measures similar to the most significant Marshfield PMRP EWAS results in NHANES III and NHANES 1999-2002. Because different survey methods were utilized between Marshfield PMRP and the NHANES, measures were chosen when they matched a significant broad environmental “class”. For example, many smoking measures were included in the most significant EWAS results and any smoking measure found in either NHANES was included for replication. T2D case/control status was defined using an algorithm previously described [17].

#### *2.3. Statistical Analysis*

A total of 314 environmental variables were included in our analysis of the Marshfield data. Logistic regression was used, adjusting for age, sex, and body mass index (BMI), with PLATO [18]. Control was coded as 1 and case as 2. All significant results were investigated to ensure that all top ranking associations had greater than 10 responses for both cases and controls. Results in figures 1 and 2 were plotted using PheWAS View [19].

For the NHANES data, logistic regression was used for all association testing, adjusting for age, sex, and BMI, in 46 to 3,964 samples (sample sizes varied for each measure) of European ancestry (self-identified non-Hispanic whites) for a total of 116 environmental variables from NHANES III (84) and NHANES 1999-2002 (32). All significant EWAS results were assessed to ensure sample size was greater than 10 for cases and controls for each variable.

### **3. Results**

In this environment-wide association study of 314 variables for type 2 diabetes, we found 12 results with a p-value less than 0.01 in the Marshfield Clinic samples. Due to the exploratory and hypothesis generating nature of this method, we are presenting all the results with a p-value less than 0.05 (31 results). Figure 1 displays the most significant EWAS associations in the Marshfield sample.

All variables could be placed into seven broad environmental “classes”: smoking, alcohol use, mania, depression, activity, diet, UV exposure, and residence. Table 2 includes all results with a p-value less than 0.05 by environment class and displays the survey question for each measure from the PhenX Toolkit.

Top Marshfield EWAS Results for Type 2 Diabetes



**Figure 1. The most significant association results in the Marshfield sample using PhenX Toolkit, DHQ, and Measurement of a Person's Habitual Physical Activity surveys.** The PhenX variables are listed along the Y-Axis. The first track shows the results of our EWAS, with  $-\log_{10}$  of the p-value plotted from most significant result at the top and descending in order. The next track shows the magnitude and direction of the effect. Case/control status was coded as 1=Control, 2=Case.

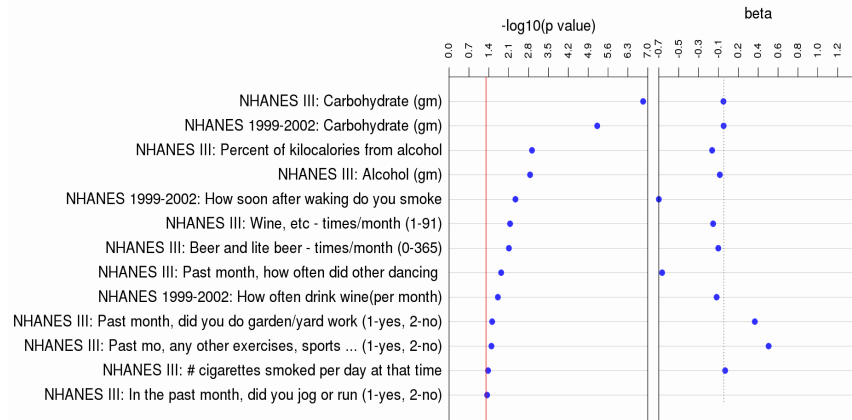
**Table 2. EWAS Variable Classes, Specific PhenX Toolkit Questions, and the EWAS Marshfield PMRP results**

Class	Survey: Variable	PhenX Toolkit Question	P-value	Beta
Alcohol	PhenX: Alcohol 30Day Frequency	Think specifically about the past 30 days, from [DATEFILL], up to and including today. During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?	6E-04	-0.03
	PhenX: Alcohol Withdrawal Hallucination	When you stopped, cut down or went without drinking, did you ever experience any of the following problems for most of the day for 2 days or longer? Did you see or hear things that weren't there? (Yes=1, No=2)	0.022	-3.041
	PhenX: Lifetime Use	In your entire life, have you had at least 1 drink of any kind of alcohol, not counting small tastes or sips? (Yes=1, No=2)	0.035	0.4655
	PhenX: Alcohol Use Liver Disease	There are several health problems that can result from long stretches of drinking. Did drinking ever cause you to have liver disease or yellow jaundice? (Yes=1, No=2)	0.037	-1.894
	PhenX: Smoking At Home	Does anyone who lives here smoke cigarettes, cigars, or pipes anywhere inside this home? (Yes=1, No=2)	6E-04	-0.889
	PhenX: Exposure Smoke Childhood	How many hours were you exposed to smoke from other people's cigarettes or tobacco products during childhood per day?	0.003	0.0064
	PhenX: Former Smoker Quantity 1DayB	Former smokers who did not ever smoke every day for the at least 6 months: when you last smoked every day, on average how many cigarettes did you smoke each day?	0.006	0.2683
	PhenX: Exposure Smoke Work	Were you exposed to smoke from other people's cigarettes or tobacco products during adulthood at work? (Yes=1, No=2)	0.007	-0.375

Smoking Exposure	PhenX: Former Smoker More 1stHour	Did you smoke more frequently during the first hours after waking than during the rest of the day? (Yes=1, No=2)	0.019	-0.689
	PhenX: Former Smoker 1stSmoke Time	How soon after you wake up do/did you smoke your first cigarette?	0.02	-0.202
	PhenX: Exposure Smoke Home	Were you exposed to smoke from other people's cigarettes or tobacco products during adulthood at home? (Yes=1, No=2)	0.021	-0.309
	PhenX: Former Smoker Quantity 1DayA	Former smokers who smoked cigarettes every day for at least 6 months: when you last smoked every day, on average how many cigarettes did you smoke each day?	0.039	0.0171
	PhenX: Exposure Smoke Present Time Hours	At the present time, how many hours per day are you exposed to the smoke of others?	0.041	0.0943
	PhenX: Exposure Smoke Adulthood Home Hours	How many hours per day were you exposed to smoke from other people's cigarettes or tobacco products during adulthood at home?	0.048	0.0046
Diet	DHQ: Caffeine(mg)	NA	0.001	-0.0005
Activity	Activity: Leisure Index	NA	0.002	-0.27
	Activity: Sports Index	NA	0.003	-0.31
	PhenX:Leisure Activity	Please check the box next to the one statement which best describes the way you spent your leisure-time during most of the last year.	0.014	-0.132
Residence	PhenX: House Gas Powered Device	Are any gas powered devices stored in any room, basement, or attached garage in this (house/apartment)? (Yes=1, No=2)	0.003	0.4187
	PhenX: House Farm	Is this property actively used as a farm or ranch? (Yes=1, No=2)	0.01	0.5382
	PhenX: Dwelling Type	What is the type of dwelling? (1=Detached house, 2=Duplex/Triplex, 3=Row house, 4=Low rise apartment (1-3 floors), 5=High rise apartment (>3 floors), 6=Mobile home / Trailer7=Other)	0.01	0.0684
	PhenX: Building Residence Age	When did you start living there?	0.024	0.0072
	PhenX: Air Conditioning Stop Month	During which month (do you usually/would you) stop using air conditioning?	0.028	0.1891
Depression	PhenX: Energy Level	Please indicate the one response that best describes your energy level for the past seven days. (0 = There is no change in my usual level of energy. 1 = I get tired more easily than usual. 2 = I have to make a big effort to start or finish my usual daily activities (for example,shopping, homework, cooking or going to work). 3 = I really cannot carry out most of my usual daily activities because I just don't have the energy.)	0.005	0.2365
	PhenX: Depression Number Weeks	About how many weeks altogether did you feel this way? Count the weeks before, during and after the worst two weeks. The total period of depression/loss of interest was:	0.044	-0.022
Mania	PhenX: Mania Increased Sex	Please try to remember a period when you were in a "high" state. In such a state: I am more interested in sex, and/or have increased sexual desire (Yes=1, No=2)	0.023	0.3615
	PhenX: Mania Impatient	Please try to remember a period when you were in a "high" state. In such a state: I am more impatient and/or get irritable more easily (Yes=1, No=2)	0.044	-0.321
UV Exposure	PhenX: Weekend Sun Hours Last Decade	On a typical weekend day in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?	0.027	-0.158
	PhenX: Weekday Sun Hours Last Decade	On a typical weekday in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?	0.031	-0.151
	PhenX: Tanning Booth	Have you ever used a tanning booth? (Yes=1, No=2)	0.042	0.4621
	PhenX: Sunlamp Times	About how many times have you used a sunlamp in your life?	0.048	0.8917

When available, similar questions from NHANES that fell into one of the above phenotype classes were included to seek replication. Measures were available in alcohol use, smoking exposure, diet, activity, depression, and mania but not in residence and UV exposure. Seven of the results were significant at  $p < 0.01$  and thirteen at  $p < 0.05$  with the same direction of effect as the related Marshfield associations (Figure 3).

## Replicating EWAS Results in NHANES



**Figure 2. Replicating results of the most significant Marshfield EWAS associations from NHANES III and NHANES 1999-2002.** Results were considered a replication if the p-value was  $< 0.05$  p-value and showed the same direction of effect as the Marshfield analyses. Controls were coded as 1 and Cases as 2. This figure is in the same format as Figure 1, with NHANES measurements on the y-axis ordered by descending association significance. The tracks show the p-value significance of the association in  $-\log_{10}(\text{p-value})$  and the magnitude and direction of the effect.

The most significant survey questionnaire result in the Marshfield EWAS was *alcohol frequency in the last 30 days*, which was inversely associated with type 2 diabetes status. This relationship was also observed for two related measures in NHANES III: alcohol consumption questions *beer and lite beer -times/month* and *wine, etc - times/month* and one in NHANES 1999-2002: alcohol consumption question: *How often drink wine (per month)*. Never having alcohol was associated with T2D status in Marshfield and did not replicate in either NHANES, though a similar, but not exact, measure was available and tested. Experiencing excessive alcohol use symptoms like hallucination due to alcohol withdrawal and liver disease from excess alcohol use was associated with having T2D in the Marshfield sample. Neither of these measures were available in either NHANES for comparison.

A number of significant results in Marshfield included measurements of first and second hand smoking exposure. Cigarette or other tobacco smoke exposure at home or at work, and for a greater number of hours during childhood, adulthood, and present time were all associated with T2D status. Additionally, for former smokers, greater number of cigarettes per day, smoking more frequently during the first hours of the day, and smoking earlier in the day were also associated with having T2D. Two of the smoking measures replicated in NHANES III: *number of cigarettes smoked/day when smoked* and NHANES 1999-2002: *how soon after waking do you smoke?* with the same direction of effect.

The two most significant results from the DHQ for the EWAS in Marshfield were a metric of caffeine consumption: caffeine (mg), which was inversely associated with T2D status and a metric of the consumption of carbohydrates (g). The caffeine measurement did not replicate in either NHANES, though increased coffee intake has been previously reported as having an association with lowered risk of T2D [20]. Carbohydrate intake did not meet the significance threshold of p-value less than 0.05 in Marshfield, but was included in the replication analysis because it was the second most significant DHQ result. When this association was investigated in NHANES III and NHANES 1999-2002 it was the most significant result for both studies.



#### 4. Discussion

Using a systematic, high-throughput EWAS method, we identified and replicated novel as well as established associations between environmental exposures and T2D. The replicating results of the association between less alcohol use per month and T2D status is consistent with prior research that demonstrates that moderate alcohol use is associated with decreased risk of T2D [21,22]. The association between T2D status and the specific symptoms of hallucination and liver disease has not been observed in the literature, to the best of the authors' knowledge. However, prior research has indicated that binge drinking and high levels of alcohol use are associated with increased risk of T2D [21,22]. It is possible that these results are spurious, or that there may be some mechanism at play by which these extreme alcohol-related measures are related to T2D. Comparison with other studies for this measure is necessary before conclusions can be drawn.

The relationship between increased smoking exposure and having T2D is also well established [23-25]. Activity level also has a well-documented link with T2D [26-28]. Here we observed a number of results from both Marshfield and NHANES III that demonstrate this association. Work activity was not significantly associated with T2D in the PhenX or Baecke measures. However, lower amounts of leisure and sports activity was associated with T2D status in Marshfield. This relationship was validated with similar measures in NHANES III: *dancing, gardening/yard work, sports, and running or jogging in the past month*.

A number of associations from the residence, depression, mania, and UV exposure classes in Marshfield did not replicate in either NHANES. This could indicate that these were false positive findings, or it could also be due to differences in measures that were used, deviation in survey question wording, or low sample sizes for a given question. Additionally, many of these results could not be evaluated for replication in either NHANES because they were not available. This demonstrates the need for standardized measures of environmental exposures, as the utilization of these measures will allow the validation of significant results across multiple studies.

Another limitation to this EWAS design is the difficulty in determining whether associations occurred simply due to T2D diagnosis. For instance, the activity questions measured activity for the past month and did not include information on activity level during childhood or if activity level changed when T2D symptoms were experienced. It is possible that the individuals with T2D participated in less leisure and sport activity due to symptoms but had greater activity levels earlier in life. Similarly, the inverse association observed between T2D and carbohydrate intake may be reflective of individuals who are restricting carbohydrate intake due to T2D diagnosis, a common dietary treatment for the disease [29]. This issue indicates the importance of gathering environmental variables that measure multiple points of an individual's lifetime. Additionally, this approach does not currently consider the full spectrum of environmental exposures. Limitations in the types of exposures assessed, and when they are collected, restricts thorough understanding of all the environmental components involved in the development of complex diseases such as T2D. Future incorporation of biological exposure data such as toxins [30] and nutrients [31] will provide additional data on the exposures associated with complex traits.

Environment-wide association studies allow the testing of multiple environmental exposures for association with common disease. Here, we demonstrate the utility of this approach for research using health record data, a novel use for this type of resource. Using this systematic EWAS approach, exposures will be identified as potential causative agents for complex traits. Significant associations can be investigated for gene-environment interactions [32,33].

Incorporating genetic data will lead to a more complete understanding of the mechanisms that lead to complex phenotypes, such as T2D. Similar to the PheWAS [12-14] method, the EWAS approach can be used to test for association between a diverse array of exposures and numerous phenotypes to discover the types of exposure that are associated with multiple traits. The search for interactions between environmental variables and genetic loci, as well as the independent exposures involved in multiple traits, will further elucidate the genetic and environmental architecture of complex human phenotypes.

## 5. Acknowledgements

This work was supported in part by NIH U01 HG004798 and its ARRA supplements and by NIH grants HG006389 and HG006385 in addition to an administrative supplement from PhenX RISING (NOT-HG-11-009). We would like to thank Dr. Geraldine McQuillan and Jody McLean for their help in accessing the Genetic NHANES data. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Center for Disease Control and Prevention.

## 6. References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106: 9362-9367.
2. Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456: 18-21.
3. Patel CJ, Bhattacharya J, Butte AJ (2010) An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 5: e10746.
4. Hamilton CM SL, Pratt JG, Maiese D, Hendershot T, et al. (2011) The PhenX Toolkit: get the most from your measures. *Am J Epidemiol* 253-260.
5. Baecke JA BJ, Frijters JE (1982) A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am J Clin Nutr* 36: 936-942.
6. Thompson FE, Subar AF, Brown CC, Smith AF, Sharbaugh CO, et al. (2002) Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. *J Am Diet Assoc* 102: 212-225.
7. Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, et al. (2001) Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires : the Eating at America's Table Study. *Am J Epidemiol* 154: 1089-1099.
8. Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 12: 417-428.
9. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, et al. (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 4: 13.
10. McCarty CA, WR GP, Westbrook SD, Caldwell MD (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Medicine* 49-79.
11. (CDC) CfDCAp (2013) National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol). Hyattsville, MD: U.S: Department of Health and Human Services, Centers for Disease Control and Prevention.
12. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26: 1205-1210.

13. Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, et al. (2011) The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet Epidemiol* 35: 410-422.
14. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, et al. (2013) Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 9: e1003087.
15. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, et al. (2012) 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: 212-218.
16. Baecke JA, Burema J, Frijters JE (1982) A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am J Clin Nutr* 36: 936-942.
17. Haiman CA, Fesinmeyer MD, Spencer KL, Buzkova P, Voruganti VS, et al. (2012) Consistent directions of effect for established type 2 diabetes risk variants across populations: the population architecture using Genomics and Epidemiology (PAGE) Consortium. *Diabetes* 61: 1642-1647.
18. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, et al. (2010) Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. *Pac Symp Biocomput*: 315-326.
19. Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD (2012) Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Min* 5: 5.
20. van Dam RM, Willett WC, Manson JE, Hu FB (2006) Coffee, caffeine, and risk of type 2 diabetes: a prospective cohort study in younger and middle-aged U.S. women. *Diabetes Care* 29: 398-403.
21. Carlsson S, Hammar N, Grill V, Kaprio J (2003) Alcohol consumption and the incidence of type 2 diabetes: a 20-year follow-up of the Finnish twin cohort study. *Diabetes Care* 26: 2785-2790.
22. Pietraszek A, Gregersen S, Hermansen K (2010) Alcohol and type 2 diabetes. A review. *Nutr Metab Cardiovasc Dis*.
23. Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J (2007) Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *JAMA* 298: 2654-2664.
24. Yeh HC, Duncan BB, Schmidt MI, Wang NY, Brancati FL (2010) Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study. *Ann Intern Med* 152: 10-17.
25. Xie XT, Liu Q, Wu J, Wakui M (2009) Impact of cigarette smoking in type 2 diabetes development. *Acta Pharmacol Sin* 30: 784-787.
26. Hu G, Qiao Q, Silventoinen K, Eriksson JG, Jousilahti P, et al. (2003) Occupational, commuting, and leisure-time physical activity in relation to risk for Type 2 diabetes in middle-aged Finnish men and women. *Diabetologia* 46: 322-329.
27. Helmrigh SP, Ragland DR, Leung RW, Paffenbarger RS, Jr. (1991) Physical activity and reduced occurrence of non-insulin-dependent diabetes mellitus. *N Engl J Med* 325: 147-152.
28. Laaksonen DE, Lindstrom J, Lakka TA, Eriksson JG, Niskanen L, et al. (2005) Physical activity in the prevention of type 2 diabetes: the Finnish diabetes prevention study. *Diabetes* 54: 158-165.
29. Nielsen JV, Joensson EA (2008) Low-carbohydrate diet in type 2 diabetes: stable improvement of bodyweight and glycemic control during 44 months follow-up. *Nutr Metab (Lond)* 5: 14.
30. Rappaport SM, Smith MT (2010) Epidemiology. Environment and disease risks. *Science* 330: 460-461.
31. Tzoulaki I, Patel CJ, Okamura T, Chan Q, Brown IJ, et al. (2012) A nutrient-wide association study on blood pressure. *Circulation* 126: 2456-2464.
32. Patel CJ, Chen R, Butte AJ (2012) Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease. *Bioinformatics* 28: i121-126.
33. Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ (2013) Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum Genet* 132: 495-508.

# DISSECTION OF COMPLEX GENE EXPRESSION USING THE COMBINED ANALYSIS OF PLEIOTROPY AND EPISTASIS

VIVEK M. PHILIP, ANNA L. TYLER, and GREGORY W. CARTER\*

*The Jackson Laboratory,  
Bar Harbor, ME, 04609, USA*

*\*E-mail: greg.carter@jax.org*

Global transcript expression experiments are commonly used to investigate the biological processes that underlie complex traits. These studies can exhibit complex patterns of pleiotropy when *trans*-acting genetic factors influence overlapping sets of multiple transcripts. Dissecting these patterns into biological modules with distinct genetic etiology can provide models of how genetic variants affect specific processes that contribute to a trait. Here we identify transcript modules associated with pleiotropic genetic factors and apply genetic interaction analysis to disentangle the regulatory architecture in a mouse intercross study of kidney function. The method, called the combined analysis of pleiotropy and epistasis (CAPE), has been previously used to model genetic networks for multiple physiological traits. It simultaneously models multiple phenotypes to identify direct genetic influences as well as influences mediated through genetic interactions. We first identified candidate *trans* expression quantitative trait loci (eQTL) and the transcripts potentially affected. We then clustered the transcripts into modules of co-expressed genes, from which we compute summary module phenotypes. Finally, we applied CAPE to map the network of interacting module QTL (modQTL) affecting the gene modules. The resulting network mapped how multiple modQTL both directly and indirectly affect modules associated with metabolic functions and biosynthetic processes. This work demonstrates how the integration of pleiotropic signals in gene expression data can be used to infer a complex hypothesis of how multiple loci interact to co-regulate transcription programs, thereby providing additional constraints to prioritize validation experiments.

*Keywords:* pleiotropy, genetic interaction, genetic network.

## 1. Introduction

The widespread adoption of genomic technologies has greatly increased the power and scope of genetic studies. One especially fruitful approach to understanding how genetic variation affects biological processes is the study of the genetics of gene expression.<sup>1–5</sup> In these studies, transcript levels are treated as panels of thousands of phenotypes that quantify the cellular composition and gene expression of a tissue sample that is related to a physiological phenotype such as disease. These data are commonly analyzed to identify expression quantitative trait loci (eQTL), which are specific chromosomal regions that associate with the expression level of a given transcript.

Associated eQTL are generally classified as local, *cis*-acting variants that affect the expression of a gene located near the associated variant, or remote, *trans*-acting variants that affect the expression of a gene located at a distance (*i.e.* outside of linkage disequilibrium (LD) or on another chromosome). The more common *cis* associations have the straightforward biological interpretation of a sequence variant directly affecting the self transcript production, stability, or splicing. However, *trans* associations are often more difficult to interpret. The structure of gene regulatory networks suggests that these *trans* associations are caused by transcription factors or other proteins that bind and regulate DNA or RNA. The co-regulatory structures

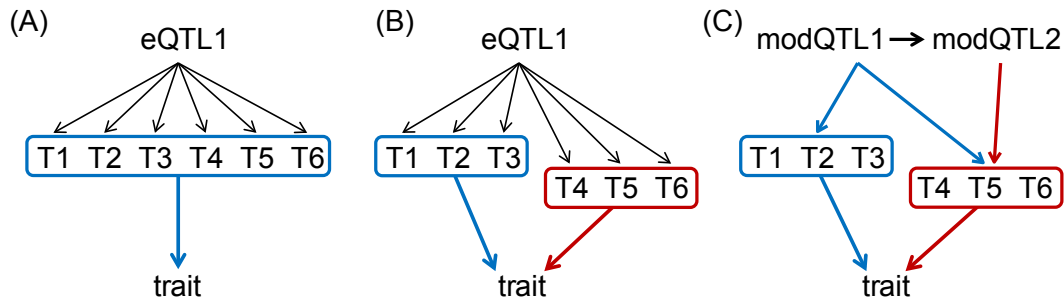


Fig. 1. Hypothetical regulatory architecture of transcripts ( $T1, \dots, T6$ ) that serve as an endophenotype for an organism-level trait. (A) Simple model in which all transcripts are associated with *trans*-acting *eQTL1* and part of a single underlying biological process affecting the trait. (B) Model with transcripts grouped into two modules that combine to affect the trait. Models (A) and (B) are indistinguishable using single-locus association. (C) Model obtained with co-expression clustering and CAPE analysis, in which the eQTL has been replaced by two multiple module QTL (modQTL). The genetic effects now map to the two modules distinctly, and the modQTL are linked by a directional influence mapping feed-forward regulation from *modQTL1* to the red module via *modQTL2*.

of these networks, in which proteins regulate multiple transcripts in complex hierarchies,<sup>6</sup> suggest that a genetic variation in one regulatory gene could have significant effects on the expression of multiple target transcripts. This would generate extensive pleiotropy as many redundantly regulated transcripts would associate with the variant. While this is pleiotropy in the sense that one genetic variant is influencing multiple traits, it is somewhat trivial in that the multiple traits are redundant outputs of the same regulatory module. This effect can be efficiently modeled by first finding modules of co-expressed transcripts that map to the common *trans*-acting module QTL (modQTL). Pleiotropy between modQTL, in which a single variant is associated with multiple distinct gene modules, is more informative in the sense of a single variant affecting multiple regulatory programs in a more complex genetic architecture (Figure 1). Distinguishing between trivial and informative pleiotropy can be difficult for complex regulatory networks in which multiple regulatory variants combine to affect hundreds of transcript outputs.

In this paper, we address this problem by modeling interacting *trans* associations for modules of co-expressed genes. We use kidney transcript data from a panel of F2 mouse intercross progeny to dissect the genetic regulation of multiple biological processes that affect overall kidney function in these genetically diverse mouse models. We use co-expression analysis to identify gene modules with correlated expression and common function and derive summary endophenotypes that describe transcriptional states. We next use a combined analysis of pleiotropy and epistasis (CAPE<sup>7</sup>) to simultaneously assess patterns of pleiotropy and statistical interactions between *trans* modQTL, in order to infer the variant-to-variant ordering of regulatory influences on the multiple processes. This approach improves the interpretation of genetic interactions in terms of directed QTL-to-QTL influences that map how a given locus suppresses or enhances the effects of a second locus. By integrating evidence of epistasis across multiple phenotypes, the CAPE method can improve power to detect modQTL interactions and assign directionality to the relationship. Furthermore, CAPE inherently parses

QTL-to-phenotype associations into direct effects and effects modified through genetic interactions, thereby separating the target transcripts into subsets that are influenced by distinct combinations of modQTL. In the case of transcript data, the result is a model of how multiple modQTL affect one another and, in turn, the regulation of multiple modules of co-expressed genes (Figure 1C). The resulting network model provides a clearer dissection of the nature of the observed pleiotropy and generates more specific hypotheses of variant activity and action.

## 2. Methods

We followed a multi-step strategy to systematically identify and model multiple gene modules that underlie kidney health and disease. The procedure is outlined in Figure 2, and consisted of three main steps: a preliminary eQTL analysis to identify transcripts affected by one or more genetic factors; clustering of the affected transcripts into co-expressed gene modules; and a network analysis to map how the gene modules are regulated by multiple interacting genetic loci. We began with a study of gene expression related to kidney function in a mouse intercross.<sup>8</sup> An F2 intercross population was derived from the kidney damage-susceptible SM/J inbred strain and the nonsusceptible MRL/MpJ inbred strain. Male SM/J mice exhibit kidney dysfunction, as measured by an increase in urinary albumin-to-creatinine ratio (ACR). To identify causal genetic loci, ACR was measured in 173 male F2 progeny. Significant QTL were mapped on chromosomes (Chrs) 1, 4, and 15, with an additional suggestive QTL on Chr 17.<sup>8</sup> This established ACR as a trait affected by multiple QTL that vary between the SM/J and MRL/MpJ lines.

### 2.1. Data

To identify the biological pathways and processes underlying the ACR results, mRNA was collected from whole kidneys of the 173 F2 animals. Data generation and processing is de-

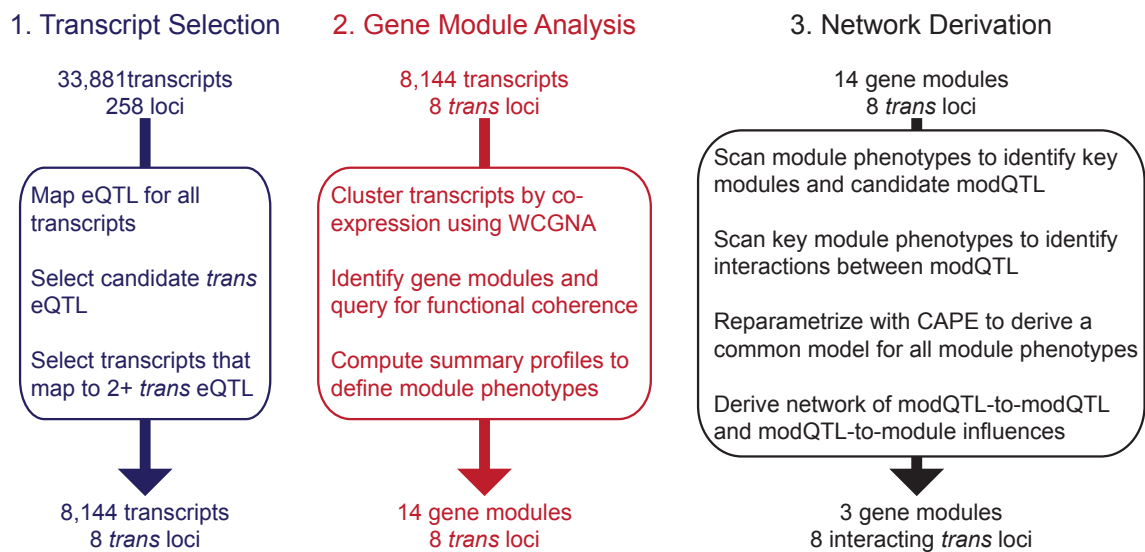


Fig. 2. Overview of analytical strategy.

scribed in depth in the initial publication<sup>8</sup> and will be summarized here. All mice were genotyped using an array that contained 258 polymorphisms that were informative between the MRL/MpJ and SM/J strains. RNA samples were labeled and hybridized to the mouse gene 1.0 ST microarray (Affymetrix, Santa Clara, California). Microarray data were imported in R (<http://www.r-project.org>) and processed using the *affy* package from Bioconductor (<http://bioconductor.org>). Normalization of the data was performed using robust multi-array average without any background subtraction. In total, 33,881 probe sets were considered.<sup>8</sup> Data were downloaded from the QTL Archive (<http://www.qtlarchive.org>).

## 2.2. Transcript Selection

Following the initial study, we performed eQTL scans using R/qtl<sup>9</sup> to test the association of every transcript with every marker. Transcript expression data were subjected to a Van der Waerden transformation<sup>10</sup> prior to eQTL mapping. Pseudo-markers were generated at 2 cM spacing for each chromosome and Haley-Knott regression was performed genome-wide for each transcript. To identify suggestive eQTL ( $P < 0.63$ ), we followed the originally-reported LOD thresholds of 2.23 and 1.44 for autosomes and the X chromosome, respectively.<sup>8</sup> This comprised a set of candidate transcripts with at least one suggestive association, each potentially regulated by one or more genetic loci. Because we were interested in analyzing overlapping patterns of pleiotropy, we further reduced this list to a set of transcripts that were associated with at least two distinct suggestive eQTL.

## 2.3. Co-Expression Modules

Since the co-regulation of multiple genes is expected to be manifest as co-expression in array data, we next performed weighted gene correlation network analysis (WGCNA)<sup>11</sup> to identify gene modules. WGCNA has been widely and successfully used to parse sets of transcripts into co-expressed modules, particularly in genetic mapping populations.<sup>12</sup> A comprehensive list of tutorials on WGCNA can be found at <http://www.genetics.ucla.edu/horvath/CoexpressionNetwork>. WGCNA generates an adjacency matrix based on the underlying absolute values of Pearson correlations among all pairs of transcripts raised to a user-defined power  $\beta$ . Here, the  $\beta$  parameter was set to 6 in order to generate the scale-free topology criterion as defined by Zhang and Horvath.<sup>13</sup> For each module, we separately obtained the first principal component (termed “eigengenes” in WGCNA) to represent the summary expression pattern for that module. We hereafter refer to these quantitative expression vectors as *module phenotypes* since they represent composite phenotypes (and the term eigengene may be confused with our distinct concept of an eigentrait in Section 2.4). Modules were queried for coherent functions using the R package GOstats.<sup>14</sup> Both Gene Ontology annotations<sup>15</sup> and KEGG pathways<sup>16</sup> were queried for functional overrepresentation. GO enrichment significance scores were corrected for multiple tests using the decorrelation of GO graph structure.<sup>17</sup>

## 2.4. CAPE Network Derivation

The combined analysis of pleiotropy and epistasis (CAPE) is an approach to modeling two or more phenotypes across a population harboring genetic variation. Detailed explanations

of the method have been published elsewhere<sup>7</sup> and will be briefly summarized here. CAPE is designed to translate data from genetic studies with multiple traits into an integrated model that accounts for variance across all phenotypes. As input, the method requires two or more quantitative phenotypes and a matrix of genotype values at markers across the genome. Variants can be engineered mutations such as gene knockouts or amplifications, or natural variants that are commonly used to map QTL. In this work, the variants will be the modQTL associated with module phenotypes. The model of variants affecting phenotypes is obtained by multivariate linear regression followed by a novel reparametrization of the results.<sup>7</sup> For a given pair of genetic variants, this reparametrization recasts the set of interaction coefficients (one for each trait) in terms of two coefficients that describe how each variant suppresses or enhances the effects of the other. This procedure translates trait-specific interaction terms into trait-independent, directed edges between the two variants, providing a common model of gene action that consistently fits all traits. These quantitative, variant-to-variant influences can be readily interpreted as genetic suppression or enhancement. When combined with the variant-to-phenotype edges, the final output is a directed network of both direct and indirect effect of variants on multiple traits. CAPE is available as an R package (<http://cran.r-project.org/web/packages/cape>), which was used in our analysis.<sup>18</sup>

We first identified a subset of modules suitable for CAPE. Each module phenotype was first scanned for modQTL associations,<sup>12</sup> with candidate loci identified using a suggestive threshold ( $P < 0.63$ ) based on a null distribution generated from 100 permutations. Genetic markers were used as loci for regression, with homozygous MRL/MpJ markers coded as 0, heterozygous markers as 0.5, and homozygous SM/J markers as 1. CAPE modules were then selected by identifying module phenotypes with a combination of candidate modQTL that included both shared and unique associations, and exhibited some degree of correlation (Figure 3). These criteria are essential to the CAPE method, given that it requires biologically related phenotypes (e.g. all modules related to kidney function) that also exhibit unique signals from which to draw functional distinctions.

The selected module phenotypes and sample genotypes were then used as input for the R implementation of CAPE.<sup>18</sup> As a first step in the analysis, CAPE decomposes all phenotypes into *eigentraits* using singular value decomposition (SVD). This procedure reorganizes the phenotypes into common and distinct signals that are expected to map to common and distinct genetic loci. Each eigentrait is scanned for its own QTL, and a user-defined number of eigentraits are selected for further analysis. This allows one to filter non-genetic signals in the data and maximizes efficiency in the analysis. In this case, the eigentraits were linear combinations of the module phenotypes. A suggestive threshold was used ( $P < 0.63$ , determined via 200 permutation tests) and the union of all suggestive markers comprised the set of markers to undergo pair-wise association tests.

Pair-wise regression models were derived and reparametrized following the CAPE method.<sup>7,18</sup> In all except specified instances, default CAPE parameters were used. To avoid effects due to LD, we omitted marker pairs with genotypes showing Pearson correlation above 0.6. Effects from QTL to eigentraits are then recomposed to map modQTL-to-phenotype influences. We performed 100,000 permutations to generate empirical  $P$  values for each parameter



in the model, and then performed a false discovery rate (FDR) correction<sup>19</sup> to compute  $q$  values. For the final network model, we used a significance cutoff of  $q < 0.05$  on both variant-to-variant and variant-to-phenotype influences.

### 3. Results

#### 3.1. *Selected Transcripts*

We performed eQTL scans on 33,881 probe transcripts across 254 independent genetic markers. This procedure yielded 53,134 suggestive associations for 26,097 transcripts, including both *cis*- and *trans*-acting loci (Table S1). In order to restrict our analysis to pleiotropic loci, we identified the number of *trans* eQTL per chromosome. This varied from 5977 transcripts associated with Chr 1 to 1101 transcripts associated with Chr 10. Since we were particularly interested in the loci associated with the ACR phenotype we concentrated our analysis on the top eight chromosomes, which comprised 60% of the associations. As in the previous study,<sup>8</sup> Chrs 1, 4, 15, and 17 were among the top *trans* chromosomes. With our weak significance cutoff, we also found four additional candidate chromosomes (Chrs 2, 6, 7, and 11). These patterns suggested widespread co-regulation of hundreds of genes by a few genetic loci. To explore potential pleiotropic effects, we selected the 8,144 transcripts associated with two or more of these chromosomes in order to analyze how these loci affect transcripts both jointly and distinctly. This provided us a large number of overlapping endophenotypes while maintaining focus on a tractable number of biological processes.

#### 3.2. *Gene Modules Analysis*

WCGNA was performed on the 8,144 transcripts identified in the previous step. We obtained 14 distinct modules, which were automatically assigned color identifiers by the software. The number of genes per module ranged from 25 to 1299 (Table S2). We queried each module for functional overrepresentation and found GO and KEGG associations for nearly all modules at a significance of  $P < 10^{-4}$  (Table S3). We observed a diversity of processes across modules, which included small organic molecule metabolism, macromolecule metabolism, immune processes, and structural development. However, the largest modules were concentrated in metabolic and transcriptional processes. These module results generally matched the KEGG pathways identified in the original analysis of the data,<sup>8</sup> which were obtained through a different analytical procedure.

We next assessed correlations between module phenotypes. Since the CAPE method relies on moderately correlated data, we sought pairs of modules with similar, but not redundant, profiles. The module phenotypes exhibited absolute Pearson correlations ranging from 0.001 to 0.8 (Figure S1).

#### 3.3. *Single-Locus Genome Scans*

We performed single-locus scans on the 14 module phenotypes to assess common associations and pleiotropic loci (Figure S2). As expected, most (82%) of the suggestive ( $P < 0.63$ ) modQTL were located on the eight chromosomes that were pre-selected for associations with individual

Table 1. Summary of gene modules used in CAPE analysis.

Module	Genes	Suggestive modQTL	Representative GO Function	Representative KEGG Pathway
blue	969	2,4,7,9,11,15	oxoacid metabolic process ( $6 \times 10^{-13}$ )	fatty acid metabolism ( $8 \times 10^{-8}$ )
grey	1299	1,4,9,11,17	oxidation-reduction process ( $1 \times 10^{-4}$ )	oxidative phosphorylation ( $3 \times 10^{-3}$ )
turquoise	1228	1,17	translational initiation ( $8 \times 10^{-5}$ )	cell cycle ( $5 \times 10^{-5}$ )

transcripts. Chrs 1, 4, 11, and 17 had the greatest number of associations, suggesting a strong biological overlap with the ACR phenotype. The number of suggestive modQTL ranged from one locus (magenta module) to eight loci (brown module).

### 3.4. *Pair-Wise Scans and Interaction Network*

We next performed two-locus interaction scans and CAPE reparametrization to derive a network of pleiotropic effects on gene modules. We selected modules with partial pleiotropy and correlation for further analysis, since modules with simpler genetic associations would not require genetic dissection with CAPE. We selected the three largest modules for CAPE analysis, summarized in Table 1. These modules met the criteria of exhibiting moderate correlations (Figure 3A) and had suggestive associations with one or more pleiotropic modQTL (Table 1). They comprised 78% of the annotated genes in all modules together, thereby accounting for the vast majority of expression variance in the data set. All modules had multiple significantly enriched annotations (Table S3). The blue module contained specific acid metabolic processes and transport genes. The grey module was concentrated in metabolic processes, programmed cell death, and catabolism. Although WCGNA assigns the grey color to transcripts that do not belong to any other module based on correlated expression, and therefore might not be co-expressed in some cases, our pre-selection of transcripts based on eQTL associations generated a grey module phenotype with sufficient common signal to generate modQTLs and a gene set with common functional annotations. Genes in the turquoise module were associated with gene expression and RNA metabolism, and other cell cycle processes. While it would have been feasible to include additional modules in the analysis, many of the modules had relative weak associations and poor correlation with other modules (Figures S2 and S3), suggesting CAPE analysis would provide little additional information. Furthermore, the addition of phenotypes associated with non-pleiotropic modQTL will likely have distinct genetic etiology, and thus can weaken significance of CAPE results by adding genetically independent variance.<sup>7</sup>

We performed SVD on the three selected module phenotypes to obtain three eigentraits (Section 2.4), which represent linear combinations of the three module phenotypes (Figure 3B). We scanned each eigentrait for QTL associations and found that most of our candidate modQTL were associated with the first and/or second eigentrait, suggesting that the genetically-driven variance is captured by these two composite phenotypes. Additionally, the first two eigentraits are of comparable weight and together account for 87% of the global variance. We therefore used these two eigentraits in our analysis, which is the default for CAPE.<sup>18</sup> A total of 54 candidate markers were identified by pooling those markers with suggestive ef-

fects, leading to 1303 marker pairs tested after removing pairs in LD. After performing the interaction analysis (Section 2.4) we transformed the eigentraits back to the original module phenotypes. This transformation does not change modQTL-to-modQTL influences.<sup>7</sup> An adjacency matrix of significant results for all marker pairs is shown in Figure 4. This non-symmetric matrix maps directed edges from each source marker to each target marker or target phenotype (rightmost columns).

A summary interaction network is shown in Figure 5. To avoid redundant interactions and nodes due to adjacent markers within a given modQTL, each modQTL-containing chromosome is represented by a single node. Although the pleiotropic modQTL and genetic interactions consistently map to the same regions on the indicated chromosomes (Figures 4 and S4), the relatively large intervals preclude reliable identification of candidate genes and therefore we simply represent the modQTL with chromosome names. Network nodes represent the effect of the SM/J allele at each modQTL. Thus the modQTL-to-phenotype edges represent the effects of a SM/J allele at the modQTL, and negative modQTL-to-modQTL interaction represents the presence of a SM/J variant at one locus suppressing another SM/J variant at a second locus. All interactions between modQTL were negative, consistent with the vast majority of findings in intercross experiments.<sup>20</sup> This may be due to functional redundancy between modQTL, suggesting that variants within pathways underlie the interactions.<sup>21–23</sup> In sum, we detected six significant modQTL-to-modQTL interactions between chromosome pairs and 15 significant modQTL-to-phenotype interactions.

Our interaction network most prominently detected interactions between Chr 1, 4, and 15. These correspond to QTL previously associated with ACR and kidney health,<sup>8</sup> and also comprised the most significant influences in our analysis. The co-suppression observed between Chr 1 and Chr 15 and between Chr 4 and Chr 15 suggest candidate genes of similar function underlie these modQTL. This genetic co-suppression was frequently observed for knockdowns of genes in the same pathways in a previous study of fly cell proliferation,<sup>23</sup> and is a consequence of highly redundant effects when SM/J alleles are present at both loci. We also note

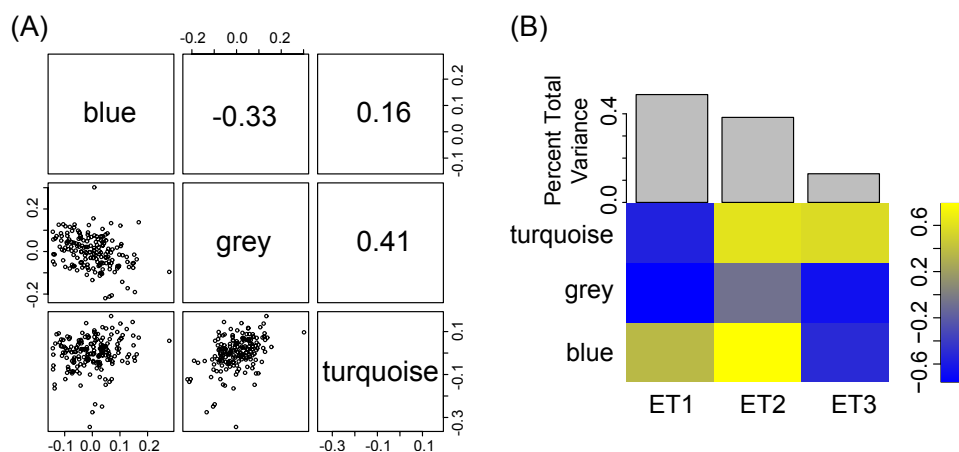


Fig. 3. Correlation structure of the three module phenotypes selected for CAPE analysis. (A) Pearson correlations and scatter plots of each pair of module phenotypes. (B) The three module phenotypes decomposed into orthogonal eigentraits, showing phenotype composition and global variance fraction for each eigentrait.

that upon conditioning on interaction effects, these modQTL are pleiotropic, with each significantly influencing both the blue and grey modules. Interestingly, the turquoise module is primarily influenced by a network of interactions between modQTL on Chrs 7, 9, and 17. The Chr 9 modQTL suppression of the Chr 17 modQTL is an example of how the CAPE method can identify indirect effects between loci, in that the Chr 9 SM/J-derived effects on the turquoise and grey modules are mediated via the presence of an SM/J allele at the Chr 17 locus. The hypothesis is that Chr 9 allele indirectly acts to suppress the Chr 17 allele, and this conditional dependence on the Chr 17 modQTL renders the Chr 9 modQTL only marginally significant when considered in isolation (Figure S2).

#### 4. Discussion and Conclusions

The CAPE method has been developed to map networks of how multiple genetic variants interact to affect multiple phenotypes, thereby identifying shared and distinct genetic etiology of complex traits. Here, we have applied this approach to address the regulation of kidney gene expression in an inbred mouse intercross. This required a focused approach to identifying patterns of co-expressed genes, followed by an application of the CAPE algorithm that separated the co-regulation of those genes in a network of causal genetic loci.

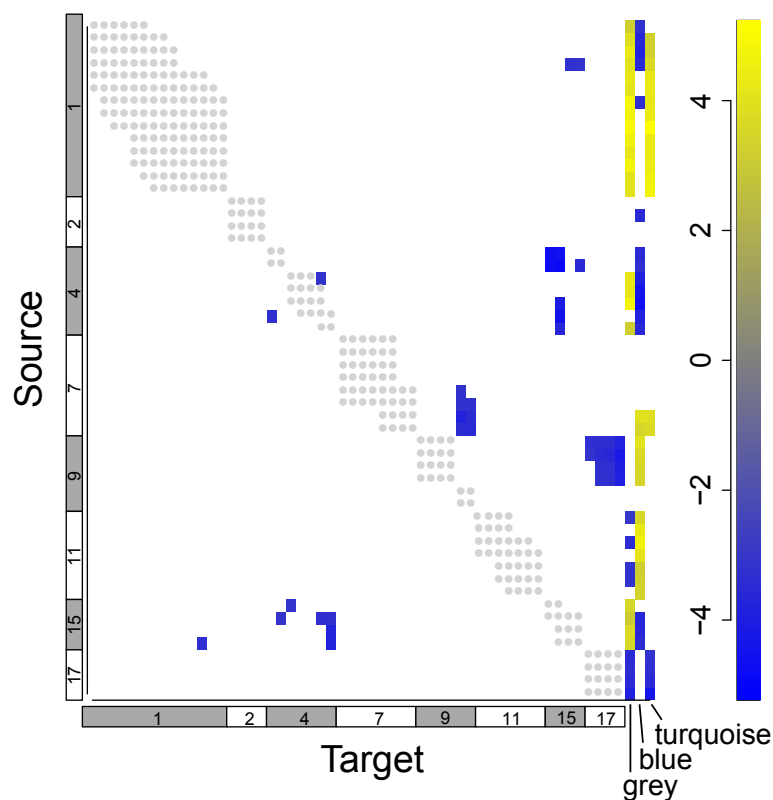


Fig. 4. Adjacency matrix of interactions derived with CAPE (FDR  $q < 0.05$ ). Markers are designated as sources or targets of directed interactions, and marker-to-phenotype influences are in the rightmost columns. Only candidate markers are shown with chromosome locations labeled, and grey dots marking pairs that were not tested due to LD. Standardized effects (effect divided by standard error) are shown to reflect significance.

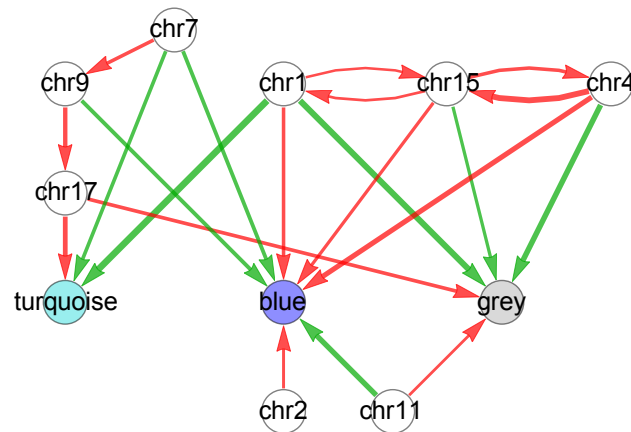


Fig. 5. Summary interaction network derived with R/cape, with interacting modQTL labeled by chromosome location on white nodes and gene modules on nodes colored by WCGNA assignments. Width of positive (green) and negative (red) edges represent significance in terms of standardized effect size.

#### 4.1. Co-Expressed Gene Modules as Complex Pleiotropy

By clustering transcripts into modules, we efficiently identified common *trans*-acting modQTL that regulate multiple co-expressed genes.<sup>12</sup> Although this strategy will not detect the majority of *cis* eQTL, which can be readily detected through direct associations of each individual transcript, it quickly identifies *trans* modQTL that exhibit pleiotropy by affecting multiple gene modules. Furthermore, the coherent expression patterns within each gene module were used as summary traits representing the activity levels of multiple biological processes. This allowed the use of the CAPE approach to map an interacting network of causal gene variants, providing an enhanced view of how multiple genetic variants commonly and differentially affected multiple gene expression patterns in the kidneys of genetically diverse mice.

#### 4.2. How Genetic Interactions Modify Pleiotropic Effects

By simultaneously analyzing genetic interactions across multiple module phenotypes, we were able to identify cases in which pleiotropic modQTL are directly associated with a module and cases in which the modQTL was indirectly affecting a module via interaction with a second modQTL. This separation provides an improved genetic model of how the modQTL might affect overall kidney health through two or more processes. The interaction cascade observed for Chrs 7, 9, and 17 suggests a series of co-dependent effects from the SM/J variant at these loci (Figure 5). When all three modQTL are inherited from SM/J, the model implies that the Chr 9 and Chr 17 modQTL are suppressed and ineffective, leading to an overall Chr 7 positive effect on the expression of the turquoise and blue modules. However, changing this scenario with an MRL/MpJ allele at the Chr 9 modQTL implies the Chr 17 modQTL counteracts the effect of the Chr 7 modQTL on the turquoise module, leaving the primary effect of the Chr 7 modQTL on the blue module only and therefore diminishing its pleiotropic effect. Examples of epistasis-dependent pleiotropy are a key element of hypotheses generated from CAPE, and their inference requires a systematic integration of both epistasis and pleiotropy in a single model of genetic effects.

### 4.3. *Overlapping Patterns of Pleiotropy to Model Complex Traits*

At the core of the CAPE method is the use of multiple QTL with partially overlapping patterns of pleiotropy over a panel of complex traits. The information coded in these patterns is used to constrain models of genetic interactions and, at the same time, map pleiotropic effects as either independent or dependent on other QTL. Thus the appropriate choice of phenotypes in analysis is essential. The most direct method is to perform single-locus scans for all phenotypes to identify shared QTL regions, with the assumption that the causal variant is common to all phenotypes. However, the sensitivity of QTL significance on limited sample numbers can rarely preclude that a QTL that falls slightly below a significance threshold is in fact causal.

In this work, we have surmounted this problem by allowing highly permissive significance thresholds for pre-selection of potentially interacting loci. Nevertheless, some of our modules exhibited few suggestive modQTL or unique loci, such as the distal Chr 6 locus that dominates the magenta module scan (Figure S2). An alternative, related approach is to select phenotypes with moderately correlated values across all samples, such as Pearson correlations of 0.3-0.8. Excessive correlation among phenotypes generates redundant genetic associations, which are ineffective for the CAPE approach, while a lack of sufficient correlation between phenotypes introduces too many conflicting signals to arrive at a common genetic model. Finally, we note that an excess of complex phenotypes can reduce the ability of CAPE to find a common genetic model. While the number of phenotypes that can be co-analyzed is theoretically unlimited, the core of the analysis is based on a dimensional reduction of a series of epistasis coefficients (one for each phenotype) to two influence parameters describing how a pair of QTL influence each other in either direction.<sup>7</sup> While the method maximizes the amount of phenotype information in two degrees of freedom independently for each locus pair, conflicting data can weaken the interaction signal. Indeed, in an earlier study of global transcript data that directly modeled principal components instead of more focused co-expression modules, it was found that simultaneously modeling more than three components diluted the power to detect interactions.<sup>7</sup> This finding applies whether the additional components are interpreted as experimental noise or additional biological signal.

### 4.4. *Potential Extensions and Validation*

The genetic models obtained by CAPE are formulated in terms of inferred influences that quantify the associated effects of variants on (1) all phenotypes; and (2) the effective weight of other variants on the phenotypes. The resulting networks structure provides a hypothesis of regulatory architecture, but does not provide any direct evidence of molecular binding. When available, the network can be used as a template for the integration of complementary molecular interaction data, with candidate regulatory interactions limited by the sign and direction of each variant-to-variant influence.<sup>24</sup> In systems lacking existing molecular interaction data, the inferred networks can serve to direct experimental validation to specific combinations of loci. For example, the binding sites of a candidate transcription factor may be predicted to be modified by the presence of a second *trans*-acting variant. This could be directly assayed with chromatin immunoprecipitation experiments performed with and without the second variant. This framework can guide follow-up investigations by providing additional constraints to

prioritize candidate regulators.

## Supplementary Material

Tables S1-S3 and Figures S1 and S2 are located at <http://carterdev.jax.org/psb2014>.

## Acknowledgments

We thank Ron Korstanje for assistance with the data. This work was supported by NIGMS grants P50 GM076468 and K25 GM079404, and NCI grant CA034196. The content does not necessarily represent the official views of NIGMS, NCI, or NIH.

## References

1. R. Jansen and J.-P. Nap, *Trends Genet* **17**, 388 (2001).
2. R. Brem, G. Yvert, R. Clinton and L. Kruglyak, *Science* **296**, 752 (2002).
3. Schadt, E.E., S. Monks, T. Drake, A. Lusi, N. Che, V. Colinayo, T. Ruff, S. Milligan, J. Lamb, G. Cavet, P. Linsley, M. Mao, R. Stoughton and S. Friend, *Nature* **422**, 297 (2003).
4. E. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. Hsu, J. Mountz, N. Baldwin, M. Langston, D. Threadgill, K. Manly and R. Williams, *Nat Genet* **37**, 233 (2005).
5. J. Keurentjes, J. Fu, I. Terpstra, J. Garcia, G. van den Ackerveken, L. Snoek, A. Peeters, D. Vreugdenhil, M. Koornneef and R. Jansen, *Proc Natl Acad Sci U S A*. **104**, 1708 (2007).
6. H. Yu and M. Gerstein, *Proc Natl Acad Sci U S A*. **103**, 14724 (2006).
7. G. W. Carter, M. Hays, A. Sherman and T. Galitski, *PLoS Genet* **8**, p. e1003010 (2012).
8. R. S. Hageman, M. S. Leduc, C. R. Caputo, S.-W. Tsaih, G. A. Churchill and R. Korstanje, *J Am Soc Nephrol* **22**, 73 (2011).
9. K. W. Broman, H. Wu, S. Sen and G. A. Churchill, *Bioinformatics* **19**, 289 (2003).
10. B. L. van der Warden, *Koninklijke Nederlandse Akademie van Wetenschappen* **55**, 453 (1952).
11. P. Langfelder and S. Horvath, *BMC Bioinformatics* **9**, 559 (2008).
12. A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis and S. Horvath, *PLoS Genet* **2**, p. e130 (2006).
13. B. Zhang and S. Horvath, *Stat Appl Genet Mol Biol* **4**, p. 17 (2005).
14. S. Falcon and R. Gentleman, *Bioinformatics* **23**, 257 (2006).
15. M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin and G. Sherlock, *Nat Genet* **25**, 25 (2000).
16. M. Kanehisa and S. Goto, *Nucleic Acids Res* **28**, 27 (2000).
17. A. Alexa, J. Rahnenführer and T. Lengauer, *Bioinformatics* **22**, 1600 (2005).
18. A. L. Tyler, W. Lu, J. J. Hendrick, V. M. Philip and G. W. Carter, *PLoS Comp Bio* (2013), in press.
19. Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B* **57**, 289 (1995).
20. W. Huang, S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt, L. D. Julien F. Ayroles, K. W. Jordan, F. Lawrence, M. M. Magwire, K. B. Crystal B. Warner, Y. Han, M. Javadi, J. Jayaseelan, S. N. Jhangiani, D. Muzny, L. P. Fiona Onger, Y.-Q. Wu, Y. Zhang, X. Zou, E. A. Stone, R. A. Gibbs and T. F. C. Mackay, *Proc Natl Acad Sci U S A*. **109**, 15553 (2012).
21. L. Avery and S. Wasserman, *Trends Genet* **8**, 312 (1992).
22. D. Segré, A. Deluna, G. Church and R. Kishony, *Nat Genet* **37**, 77 (2005).
23. G. W. Carter, *G3* **3**, 807 (2013).
24. G. W. Carter, S. Prinz, C. Neou, J. P. Shelby, B. Marzolf, V. Thorsson and T. Galitski, *Mol Syst Biol* **3**, p. 96 (2007).

## **PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY**

JENNIFER LISTGARTEN

*Microsoft Research, 110 Glendon Avenue, Suite PH1, Los Angeles, CA*

Email: jennl@microsoft.com

OLIVER STEGLE

*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom*

Email: oliver.stegle@ebi.ac.uk

QUAID MORRIS

*University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, ON M5S 3E1, Canada*

Email: quaid.morris@utoronto.ca

STEVEN E BRENNER

*Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720*

Email: brenner@compbio.berkeley.edu

LEOPOLD PARTS

*University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, ON M5S 3E1, Canada*

Email: leopold.parts@utoronto.ca

Genotyping and large-scale molecular phenotyping are already available for large patient cohorts and will soon become routinely available for all patients. Exome or complete genome sequences are being increasingly collected and are explored as newborn screening technologies. These data are setting the stage for rapid advances in personalized medicine, enabling better disease classification, more precise treatment, and improved disease prevention. Robust statistical and computational methods for analyzing these data are critical to realizing the promise of genome-based medicine. The challenges span from accurate low level analyses of high throughput datasets to identification of causal links between different layers of molecular information, and incorporating them into diagnostics. Important analysis problems include accurate phenotypic characterization, identifying and correcting for latent structure, dealing with missing data, deciding at what level to test (e.g. single base pair values, sets of polymorphisms, exonic regions, etc.), data heterogeneity, the problem of multiple testing, integrating various modalities, deducing functional consequences *in silico*, addressing data quality, and making sense of new data types as they become available.



For example, in genome-wide association studies, population structure and family relatedness can reduce power and cause spurious associations. In gene expression and epigenetic studies, experimental artifacts and environmental influences have been shown to corrupt results of naive analyses. All of these problems can be tackled by various classes of latent variables, such as those related to Principal Components Analysis and probabilistic variations thereof, linear and non-linear mixed models. These models learn latent factors from the large scale of the data---that is, patterns which permeate many of the features, and therefore speak to wide-spread “contamination”. By removing these broad patterns, we hope to be left with the true associations; however, being certain of this is difficult [1-14].

Using patient genotype to inform treatment in the clinic is limited by our ability to accurately predict the impact of genetic variation, and the lack of models for its mechanistic effect. While whole genome sequencing has been successfully used to identify causal mutations for severe developmental disorders and other Mendelian diseases, use of genotype information has not yet permeated clinical practice, save for a handful of single locus tests [15]. Personalized approaches, however, are becoming increasingly common in applications to cancer treatment, albeit these are at present mostly limited to a research setting. Questions that remain are whether to treat the sequence data as clinical test, and only report known causal locus results for any phenotype under heavy regulation, or whether to broadly disclose any incidental findings. Many found variants are of unknown effect, and precise statistical models, as well as convenient software are needed to help practitioners make decisions [16]. To this end, efforts such as Critical Assessment of Genome Interpretation [17,18] are performing controlled experiments to probe the limits of our ability to predict phenotype from genotype.

The path from genotype to disease state goes through intermediate phenotypes [19]. To modulate the disease risk or trait, one of the molecular intermediates must be changed in a controlled way using small molecules or changes in environment, but one current limitation is finding out the right targets for these interventions. A first level of understanding should come from genetic mapping studies - to which extent do the loci responsible for heritable disease risk affect intermediate traits? Some progress has been made on this front over the last years, especially for RNA levels [20,21], but also protein and metabolite abundances [22,23], with much remaining to be done. The next task is distinguishing the actual drivers of ailment from traits that do respond to genotype, but do not cause disease. Causal models, such as Mendelian randomization methods, will play a crucial role in separating out the molecular causes of disease from the high-dimensional state of the organism [24,25].

Clinical grade confidence in data and methods to use genome information for providing better treatment has been difficult to establish, with heterogeneous and imperfect medical records also remaining a real bottleneck. However, each year brings more rigor and agreement in applications of genome-based personalized medicine in the field. Still, much work is required in all areas, from basic discovery of molecular mechanisms of disease pathology, to statistical methods of causality and publicly available computational infrastructure to deliver on the promise of genetic information in the clinic. The payoffs will be large.

## Session contributions

The session keynote is given by **Robert Gentleman**, who has spearheaded the use of computational methods in biology and medicine [26], and is currently employing them to design cancer therapeutics.

The availability of inexpensive partial genotype data, and increasingly cheaper full genome sequencing to complement traditional diagnostic markers has fuelled the promise of personalised genomic medicine. However, genetic tests inform the diagnosis and treatment for only a minority of heritable disease cases in clinic today. This is partly due to low explanatory power of common small effect variants that underlie the common disease risk, but also due to larger effect alleles not being well captured by standard genotyping arrays as they have low frequency. In our session, **Martin et al.** analyse the performance of different genotyping platforms for imputing rare coding variation. Perhaps somewhat surprisingly, they find that genotyping arrays dedicated to measuring rare exome variants can be less useful in imputing unobserved rare variants than dense common variant arrays. This occurs because the latter are actually able to tag unmeasured variants (including rare ones) better than the specialized rare variant chips.

Interpreting incidental findings in whole-genome sequencing is difficult, and can take up considerable time of clinicians and genetic counselors. **Daneshjou et al.** will present PATH-SCAN, a publicly available tool that automatically annotates the variants that have been designated as pathogenic by ClinVar. The tool is expected to accelerate the analysis of genes that have been recommended by the American College of Medical Genetics and Genomics to be followed up and reported to the patient.

Also in our session, **Zhe et al.** tackle the problem of employing genotype and endophenotypes (intermediate phenotype) in disease diagnosis. Focussing on dissecting the genetic basis of Alzheimer's disease, a neurodegenerative disorder, they apply a latent variable model to the genotypes, magnetic resonance imaging, and diagnosis label where all the three types of observed features are sparse manifestations of a single continuous underlying disease state. After learning the model parameters, they then use them to predict disease state in a patient cohort, achieving better performance compared to current alternatives, and also uncovering several potentially causal links between genotype and the measured endophenotypes.

An important role for personalized medicine is in predicting frequency of drug side effects from genotype. **Oetjens et al.** genotyped 34 genes for 127 heart transplant recipients, 35 of whom had an adverse reaction to an immune suppressor. Incorporating data from electronic medical records, known predisposition to chronic kidney disease, and broad variance components in the genotype, the authors identified a single non-synonymous variant that significantly increased the risk of renal failure. Their study serves as a nice proof-of-principle that even with limited sample size and number of genotyped loci, genotype-dependent side effects can be identified using statistical analyses of longitudinal data.

**Parikh et al.** consider the problem of simultaneously inferring gene expression networks from a series of evolving conditions (*e.g.*, healthy tissue *versus* cancer stages) to identify functional roles of individual genes and pinpoint the causal changes. They propose a model that finds a sparse representation of the gene co-expression patterns sharing information across the different stages in a principled manner, and one which accounts for potential differences in the network structures.

Finding individual-specific contributors to immune response can help inform therapy of viral infections. In our session, **Perina et al.** propose a bag of words model to describe the distribution of epitopes presented by cells that are targeted by immune surveillance mechanisms. Their approach is able to better explain the correlations between individual epitopes compared to alternatives. For a clinical application, they test the models on a cohort of HIV patients to find links between distribution of epitopes and the viral load.

## References

1. Lippert C., et al. *The benefits of selecting phenotype-specific variants for applications of mixed models in genomics*. Sci Rep. 2013. **3**:1815.
2. Listgarten, J., et al. *FaST-LMM-Select for addressing confounding from spatial structure and rare variants*. Nat Genet, 2013. **45**(5):470-1.
3. Listgarten J., et al. *A powerful and efficient set test for genetic markers that handles confounders*. Bioinformatics, 2013. **29**(12):1526-33.
4. Listgarten J., et al. *Improved linear mixed models for genome-wide association studies*. Nature Methods, 2012, doi:10.1038/nmeth.2037.
5. Listgarten, J. *Correction for Hidden Confounders in the Genetic Analysis of Gene Expression*. PNAS, 2010.
6. Stegle, O. et al. *A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies*. PLoS Comp Biol, 2010.
7. Parts, L. et al. *Joint genetic analysis of gene expression data with inferred cellular phenotypes*. PLoS Genet. 201. **7**(1):e1001276.
8. Stegle O., et al. *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses*. Nat. Protoc. 2012. **7**(3):500-7.
9. Fusi, N., et al. *Detecting regulatory gene-environment interactions with unmeasured environmental factors*. Bioinformatics. 2013. **29**(11):1382-9.
10. Fusi, N. et al. *Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies*. PLoS Comp Biol, 2012.
11. Balding, D. *A tutorial on statistical methods for population association studies*. Nat Rev Genet, 2006. **7**(9): 781-91.
12. Quon, G., et al. *ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing*. Bioinformatics, 2009. **25**(21):2882-9.
13. Quon, G., et al. *Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction*. Genome Medicine, 2013. **5**:29.
14. Qiao, W., et al. *PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions*. PLoS Computational Biology, 2012. **8**(12): e1002838.
15. McCarthy J.J., et al. *Genomic medicine: a decade of successes, challenges, and opportunities*. Sci Transl Med. 2013. **5**(189):189sr4.
16. Jacob, J.J. et al. *Genomics in clinical practice: lessons from the front lines*. Sci Transl Med., 2013. **5**(194):194cm5.

17. Radivojac, P., et al. *A large-scale evaluation of computational protein function prediction*. Nat Methods. 2013. **10**(3):221-7.
18. Callaway, E. *Mutation prediction software rewarded*. Nature, 2010. doi:10.1038/news.2010.679
19. Chakravarti, A. et al. *Distilling pathophysiology from complex disease genetics*. Cell, 2013. **55**(1):21-6.
20. Lappalainen T., et al. *Transcriptome and genome sequencing uncovers functional variation in humans*. Nature. 2013. **501**(7468):506-11.
21. Parts L., et al. *Extent, causes and consequences of small RNA expression variation in human adipose tissue*. PLoS Genet, 2012. **8**(5):e1002704.
22. Kettunen, J., et al. *Genome-wide association study identifies multiple loci influencing human serum metabolite levels*. Nat Genet. 2012.**44**(3):269-76.
23. Johansson, Å., et al. *Identification of genetic variants influencing the human plasma proteome*. PNAS, 2013. **110**(12):4673-8.
24. Gagneur J., et al. *Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype*. PLoS Genet. 2013. **9**(9):e1003803.
25. Fall, T., et al. *The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis*. PLoS Med, 2013. **10**(6):e1001474.
26. Gentleman, R.C., et al. *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biol. 2004. **5**(10):R80.

## **PATH-SCAN: A REPORTING TOOL FOR IDENTIFYING CLINICALLY ACTIONABLE VARIANTS**

ROXANA DANESHJOU<sup>†1</sup>, ZACHARY ZAPPALA<sup>†1</sup>, KIM KUKURBA<sup>1</sup>, SEAN M BOYLE<sup>1</sup>, KELLY E ORMOND<sup>1</sup>, TERI E KLEIN<sup>1</sup>, MICHAEL SNYDER<sup>1</sup>, CARLOS D BUSTAMANTE<sup>1</sup>, RUSS B ALTMAN<sup>\*1,2</sup>, STEPHEN B MONTGOMERY<sup>\*1,3</sup>

*1. Department of Genetics, Stanford University  
Stanford, CA 94061, United States*

*2. Department of Bioengineering, Stanford University  
Stanford, CA 94061, United States*

*3. Department of Pathology, Stanford University  
Stanford, CA 94061, United States*

The American College of Medical Genetics and Genomics (ACMG) recently released guidelines regarding the reporting of incidental findings in sequencing data. Given the availability of Direct to Consumer (DTC) genetic testing and the falling cost of whole exome and genome sequencing, individuals will increasingly have the opportunity to analyze their own genomic data. We have developed a web-based tool, PATH-SCAN, which annotates individual genomes and exomes for ClinVar designated pathogenic variants found within the genes from the ACMG guidelines. Because mutations in these genes predispose individuals to conditions with actionable outcomes, our tool will allow individuals or researchers to identify potential risk variants in order to consult physicians or genetic counselors for further evaluation. Moreover, our tool allows individuals to anonymously submit their pathogenic burden, so that we can crowd source the collection of quantitative information regarding the frequency of these variants. We tested our tool on 1092 publicly available genomes from the 1000 Genomes project, 163 genomes from the Personal Genome Project, and 15 genomes from a clinical genome sequencing research project. Excluding the most commonly seen variant in 1000 Genomes, about 20% of all genomes analyzed had a ClinVar designated pathogenic variant that required further evaluation.

---

<sup>†</sup> Co-first author

<sup>\*</sup> Co-last author

## 1. Background and Significance

The era of personalized genomics received a jumpstart in 2007, when 23andMe, deCODEme, and Navigenics began to offer Direct to Consumer (DTC) personal genetic testing.<sup>1</sup> Reports from these companies include genotyping of up to hundreds of thousands of loci with phenotypic interpretation for dozens to hundreds of traits and conditions based mainly upon genome wide association studies (GWAS).<sup>2,3</sup> The use of such genetic information in a clinical setting has been slower to develop, although several academic medical centers have established genomic medicine programs.<sup>4</sup> Moreover, with the falling price of next generation sequencing, the number of whole genomes and exomes being sequenced is steadily increasing.<sup>4,5</sup> Whole genome or exome sequencing provides much more data than genotyping, especially with regards to rare and private mutations. As a consequence, incidental findings in an individual's genome beyond the scope of the research or clinical question are likely to exist. There is some debate surrounding the handling of the so-called "incidentalome", particularly since novel, rare, or private mutations may be difficult to interpret and a full interpretation is cost prohibitive in most settings.<sup>6</sup> Recently, the American College of Medical Genetic and Genomics (ACMG) put out a report with recommendations on which incidental findings should be specifically analyzed and reported.<sup>7</sup> In this case, "incidental findings" refer to pathogenic or potentially pathogenic variants discovered in a subset of genes during whole genome or exome sequencing, regardless of the reason sequencing was ordered.<sup>7,8</sup> The list of 57 genes covering 24 conditions put forward by the ACMG are those that have medically actionable outcomes. For example, the list includes *BRCA1* and *TNNI3*, mutations in which can lead to breast cancer and hypertrophic cardiomyopathy, respectively.<sup>7</sup> Currently, it is not known exactly what percentage of individual genomes will carry such variants, and an understanding of the pathogenic burden will allow researchers to better understand the resources required to evaluate such variants. Here, we present a publicly available tool, PATH-SCAN, which annotates genomes for ClinVar designated pathogenic variants in the list of genes recommended by the ACMG.<sup>7</sup>

## 2. Methods

PATH-SCAN allows a researcher or individual to analyze and annotate individual exomes or genomes for a set of pathogenic variants identified in the ClinVar database in the genes put forward by the ACMG. These annotations are presented in a report with genomic information and links to additional information. Due to the consequences of many of these variants, security and privacy are mainstays of the PATH-SCAN program. PATH-SCAN maintains complete privacy by performing all analyses on an individual's local machine, similar to a previously described genotype analysis tool, INTERPRETOME.<sup>9</sup> PATH-SCAN offers an option to anonymously submit data to our research group allowing us to use crowd sourcing to determine the prevalence of pathogenic variants found in the ACMG gene list.

### 2.1. Pathogenic Variant Selection

Pathogenic variants were selected from the National Center for Biotechnology Information's (NCBI) ClinVar variant call file (VCF). From this database of variants, those variants with at least

one submission as “pathogenic” were extracted and annotated with links to other clinically relevant databases. Since the ClinVar database is a collaborative database with potentially variable quality in individual variant results, we filtered out any variant that was tagged with a “variant suspect” code. A variant might be labeled as such for several reasons, including being called from an old genomic alignment or a suspected paralog. From this list, we then extracted only the variants that mapped to the 57 genes listed in the ACMG report. Gene boundaries were determined using GRCh37.p10.<sup>10</sup> In total, ClinVar had records for 994 variants designated as pathogenic across 57 genes. These variants are included in the PATH-SCAN package. The original ClinVar VCF can be found here [http://www.ncbi.nlm.nih.gov/variation/docs/human\\_variation\\_vcf/](http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/). The use of other databases is allowed in case an individual wishes to use an alternative database for annotation (see Appendix).

## 2.2. Analysis Tool

Our cross-platform program, PATH-SCAN, utilizes a database of 994 variants to scan personal genomes and annotate them. The annotations produced by PATH-SCAN are made available to the end user or researcher as a local html page with a simplified user interface for increased accessibility and transparency. To assist interpretation of this information and provide a model for future genome interpretation tools, each recognized variant and annotation is presented alongside links to relevant educational resources, including ClinVar, OMIM, and consolidated Gene Reviews from the National Center for Biotechnology Information (NCBI).

Crowd sourcing data collection was accomplished by making a submission link available that transfers de-identified anonymous information back to our data collection server. In order to prevent any privacy concerns regarding this data collection, PATH-SCAN only transmits the total number of pathogenic variants annotated for each gene (e.g. the total pathogenic burden per gene of an individual genome) as well as a unique key to prevent duplicate submissions from unwary users. Additional information such as ancestry is optional to transmit. The unique key is calculated by PATH-SCAN automatically by hashing the personal genome file using the SHA-2 family of cryptographic functions. In addition to these security measures, a privacy message is presented before the user can submit their data. For the personal genomes we had direct access to, the full annotations made by PATH-SCAN were used to collect data on individual diseases and variants as well as aggregate distributions of pathogenic variants across individuals.

PATH-SCAN is a command line utility that was developed in Python 2.7.5 and has no external dependencies. The PATH-SCAN program comes pre-loaded with the existing database of pathogenic variants. We also have the ability to load updated databases pending re-releases of the ACMG recommendation or for custom made variant databases. PATH-SCAN will automatically detect and process variant call files (VCFs), tab-separated variant (TSV) files from Complete Genomics, and SNP chip results from 23andMe. Because 23andMe only genotypes SNPs, PATH-SCAN will not scan data in this form for indels. For a whole genome VCF file that is 336 MB, PATH-SCAN runs in 24 seconds on a machine with 16GB of RAM and a 2.3 ghz processor. The script and database are bundled and available for download online at: <http://montgomerylab.stanford.edu/pathscan.zip>.

### 2.3. Applying PATH-SCAN to existing datasets: 1000 Genomes, Personal Genomes Project, and a clinical sequencing project

We pilot tested PATH-SCAN on the 1092 individuals from the 1000 Genomes project publicly available low coverage ( $\sim 4\times$ ) genomes.<sup>11</sup> We also investigated how ancestry affected the number of variants found in each population. Additionally, we tested PATH-SCAN on exome chip data for 2123 individuals from the 1000 Genomes project. These individuals overlap with the 1092 whole genome data.<sup>11</sup>

We also tested our tool on 163 Genomes downloaded from the Personal Genomes Project, which were in the Complete Genomics format ([www.personalgenomes.org/community.html](http://www.personalgenomes.org/community.html)).<sup>12</sup> We only considered variants called with high quality. High quality variants are called on homozygous calls with a quality score greater than or equal to 20 and heterozygous calls with a quality score greater than or equal to 40 under the maximum likelihood variable allele fraction.

In addition to the larger scale, low-coverage studies previously discussed, we tested our tool on a clinical sequencing project consisting of 15 individuals (3 trios and 4 unrelated individuals).

## 3. Results

### 3.1. Pathogenic variants studied

By filtering ClinVar for variants with evidence of pathogenicity in the subset of ACMG guideline genes, we selected 994 variants that our tool evaluates. These variants include 651 single nucleotide polymorphisms (SNPs) and 343 small insertions/deletions (indels). 65.5% of the pathogenic variants evaluated were SNPs, evenly distributed across all 12 non-synonymous nucleotide-to-nucleotide transversions. Variants were not evenly distributed across the 57 genes, with *BRCA1* and *BRCA2* having the largest number of variants (Figure 1). An example of the output of PATH-SCAN can be seen in Figure 2.

ACTC1	8	KCNQ1	26	PKP2	2	STK11	12
APC	16	LMNA	47	PMS2	5	TGFBR1	7
APOB	12	MEN1	11	PRKAG2	10	TGFBR2	15
BRCA1	121	MLH1	18	PTEN	20	TMEM43	1
BRCA2	159	MSH2	12	RB1	12	TNNI3	13
CACNA1S	8	MSH6	2	RET	56	TNNT2	9
COL3A1	17	MYBPC3	6	RYR1	34	TP53	23
DSG2	5	MYH7	40	RYR2	10	TPM1	6
DSP	10	MYL3	3	SCN5A	38	TSC1	9
FBN1	37	NF2	13	SDHAF2	1	TSC2	14
GLA	39	NTRK1	12	SDHB	11	VHL	20
KCNH2	20	PCSK9	3	SDHD	18	WT1	15

Figure 1: Total number of pathogenic variants found per gene in ClinVar. In total there were 994 variants distributed across the 57 genes specified by the ACMG recommendations.



## PATH-SCAN

Note that all end-user services are undertaken with your privacy in mind; no data is transferred to our server and the entire annotation process is carried out on your machine.

#	Gene	Condition	RSID	Chromosome	Position	Sample	OMIM Reports	Gene Reviews
1	<a href="#">BRCA1</a>	<a href="#">Hereditary Breast and Ovarian Cancer</a>	<a href="#">rs80358145</a>	17	41199659	0	<a href="#">604370 - 612555</a>	<a href="#">Gene Review on PubMed</a>
2	<a href="#">BRCA1</a>	<a href="#">Hereditary Breast and Ovarian Cancer</a>	<a href="#">rs80358145</a>	17	41199659	1	<a href="#">604370 - 612555</a>	<a href="#">Gene Review on PubMed</a>

**OPTIONAL** Please consider submitting these completely anonymized results for research purposes (no identifying information will be sent).



PATH-SCAN is licensed under a [Creative Commons Attribution 3.0 Unported License](#).

Figure 2: Sample output of PATH-SCAN. Information regarding the affected variant (including chromosome, position, rsID, and gene) are displayed alongside relevant information including what condition this variant is expected to have pathology in and links to clinical reviews and publications regarding the condition. A crowd-sourcing form is available at the bottom of the page if users wish to submit de-identified information to our servers.

### 3.2. PATH-SCAN identifies variants in 1000 Genomes Data

Out of 1092 individuals with low coverage genome data, 633 have at least one ClinVar designated pathogenic variant reported in one of the ACMG genes. Out of the 2123 exome-chipped individuals (which overlaps with the 1092 individuals with whole genomes), 997 individuals had at least one variant reported. The most common variant seen was rs1805124 (*SCN5A*), which was seen in 41.2% of individuals (Table 1). This variant has an allele frequency of about 20% in the 1000 Genomes population. Excluding this very common variant, out of 1092 low coverage genomes, 225 individuals had at least one pathogenic variant in one of the ACMG genes, and 237 individuals had at least one pathogenic variant in the exome chip data.

Table 1: Variants and individual frequencies seen in the 1000 Genomes Project Data. Absent data from the exome chip columns due to incomplete sequencing coverage in those individuals. Frequencies represent frequency of individuals with at least one copy of the variant and not allele frequencies.

Gene	Disease	rsID	4x Genome (1,092 indiv.)	Freq.	Exome Chip (2,123 indiv.)	Freq.
<i>APC</i>	Familial adenomatous polyposis	rs137854567	2	0.002	-	-
		rs1801166	8	0.007	-	-
<i>DSP</i>	Arrhythmogenic right-ventricular cardiomyopathy	rs121912998	4	0.004	-	-
<i>LMNA</i>	Hypertrophic cardiomyopathy, dilated	rs57830985	1	0.001	-	-
<i>MSH6</i>	Lynch syndrome	rs2020912	11	0.010	13	0.006

SCN5A	Romano–Ward long QT syndrome types 1, 2, and 3, Brugada syndrome	rs1805124	450	0.412	852	0.401
		rs41261344	26	0.024	72	0.034
		rs45620037	1	0.001	-	-
		rs7626962	26	0.024	65	0.031
SDHB	Hereditary paraganglioma–pheochromocytoma syndrome	rs11203289	19	0.017	-	-
		rs33927012	17	0.016	30	0.014
SDHD	Hereditary paraganglioma–pheochromocytoma syndrome	rs11214077	20	0.018	-	-
		rs34677591	13	0.012	-	-
STK11	Peutz–Jeghers syndrome	rs59912467	28	0.026	61	0.029
TP53	Li–Fraumeni syndrome	rs28934576	1	0.001	-	-
TSC1	Tuberous sclerosis complex	rs118203576	48	0.044	-	-
		rs118203657	5	0.005	-	-

### 3.3. *PATH-SCAN identifies variants in the Personal Genomes Project*

We applied PATH-SCAN to 163 genomes in Complete Genomics format. 77 of these individuals were found to have at least one variant. The most common variant, once again, was rs1805124 (Table 2). Excluding this variant, 27 individuals had at least one variant in one of the ACMG guidelines genes.

Table 2: Variants and counts seen in 163 Personal Genomes

Gene	Disease	rsID	PGP Genomes (163 individuals)
APC	Familial adenomatous polyposis	rs1801166	5
DSG2	Arrhythmogenic right-ventricular cardiomyopathy	rs193922639	2
FBN1	Marfan syndrome, Loeys–Dietz syndromes, and familial thoracic aortic aneurysms and dissections	rs137854475	1
KCNQ1	Romano–Ward long QT syndrome types 1, 2, and 3, Brugada syndrome	rs267607197	1
RET	Multiple endocrine neoplasia type 2; Familial medullary thyroid cancer	rs77724903	1
SCN5A	Romano–Ward long QT syndrome types 1, 2, and 3, Brugada syndrome	rs1805124	62
		rs41261344	1
		rs137854610	1
SDHB	Hereditary paraganglioma–pheochromocytoma syndrome	rs33927012	7
SDHD	Hereditary paraganglioma–	rs11214077	5

	pheochromocytoma syndrome	rs34677591	1
STK11	Peutz-Jeghers syndrome	rs59912467	1
TNNT2	Hypertrophic cardiomyopathy, dilated cardiomyopathy	rs121964857	1
TSC1	Tuberous sclerosis complex	rs118203657	1

### 3.4. Analyzing variant burden across populations

We looked at the variant detection in the different 1000 Genomes populations (Table 3). Because of the high allele frequency of rs180524, we looked at the frequencies with and without this SNP.

Table 3: Number of variants seen in the different 1000 Genomes populations. ACB- African Caribbean in Barbados; ASW - HapMap African ancestry individuals from Southwest US; CDX- Chinese Dai in Xishuangbanna, China; CEU – Utah residents with Northern and Western European ancestry; CHB - Han Chinese in Beijing; CHD - Chinese in metropolitan Denver, CO; CHS – Southern Han Chinese; CLM - Colombian in Medellin, Colombia; FIN -HapMap Finnish individuals from Finland; GBR - British individuals from England and Scotland; GIH - HapMap Gujarati India individuals from Texas; IBS - Iberian populations in Spain; JPT – Japanese in Tokyo, Japan; KHV - Kinh in Ho Chi Minh City, Vietnam; LWK - Luhya individuals in Webuye, Kenya; MKK- HapMap Maasai individuals from Kenya; MXL - HapMap Mexican individuals from LA California; PEL - Peruvian in Lima, Peru; PUR- Puerto Rican in Puerto Rico; TSI – Tuscans from Italy; YRI- Yoruba from Ibadan, Nigeria

Population	4x Genome Samples (1092 total)	Avg. variant count/ person 4x Genome	Avg. variant count/person 4x Genome w/o rs180524	Exome Chip Samples (2123 total)	Avg. variant count/person Exome Chip	Avg. variant count/person Exome Chip w/o rs180524
ACB	0	-	-	98	71/0.72	20/0.20
ASW	61	44/0.72	12/0.20	97	63/0.65	10/0.10
CDX	0	-	-	100	36/0.36	24/0.24
CEU	85	45/0.53	13/0.15	104	41/0.39	5/0.05
CHB	97	44/0.45	20/0.21	100	44/0.44	20/0.2
CHD	0	-	-	1	0/0	0/0
CHS	100	35/0.35	21/0.21	150	44/0.37	35/0.23
CLM	60	47/0.78	21/0.35	107	52/0.46	2/0.19
FIN	93	45/0.48	13/0.14	100	40/0.4	4/0.04
GBR	89	53/0.60	13/0.15	101	54/0.53	9/0.09
GIH	0	-	-	93	42/0.45	4/0.04
IBS	14	18/1.29	4/0.29	147	87/0.59	12/0.08
JPT	89	35/0.39	10/0.11	100	37/0.37	10/0.1
KHV	0	-	-	118	56/0.47	38/0.32
LWK	97	64/0.66	9/0.09	100	64/0.64	6/0.06
MKK	0	-	-	31	22/0.71	1/0.03
MXL	66	47/0.71	26/0.39	100	35/0.35	5/0.05
PEL	0	-	-	104	46/0.44	0/0
PUR	55	49/0.89	21/0.38	111	62/0.56	3/0.02

TSI	98	69/0.70	22/0.22	100	57/0.57	9/0.09
YRI	88	85/0.97	25/0.28	161	129/0.80	24/0.15

In 1092 Genomes, the average number of variants per genome ranged from 0.35 (CHS) to 1.29 (IBS). Without rs180524, the average number of variants per person ranged from 0.09 (LWK) to 0.39 (MXL). Populations that were closely related had similar average variants per person (Figure 3). Particular populations, such as LWK, had a much lower variant count than other populations when rs180524 was not taken into consideration.

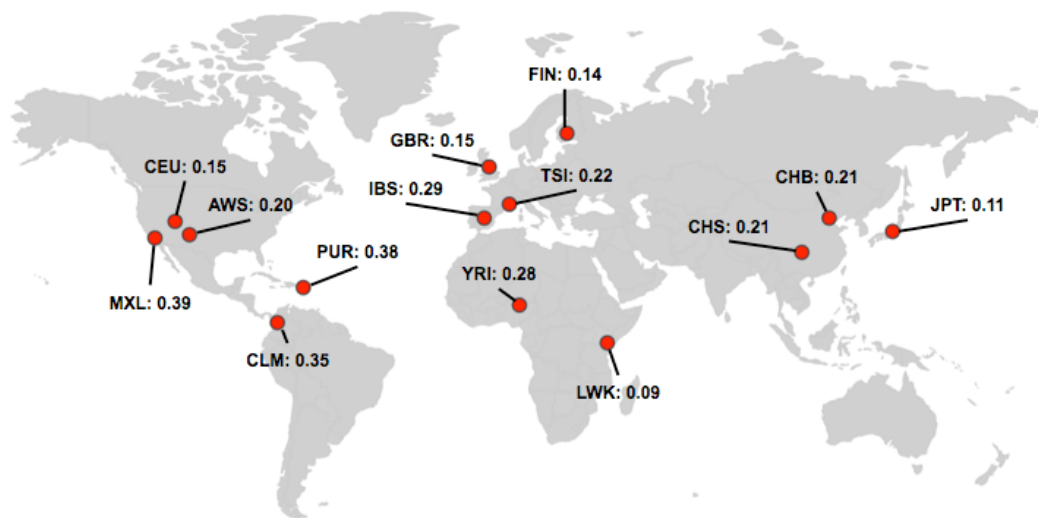


Figure 3: Average variants per individual in 1092 Genomes (with rs1805124 removed due to high allele frequency in all populations).

### 3.5. Applying *PATH-SCAN* to a clinical genome sequencing project

In a clinical genome sequencing project consisting of 15 individuals, 2 subjects had 2 ClinVar pathogenic variants, 5 subjects had 2 ClinVar pathogenic variants, and 8 subjects had 0 ClinVar pathogenic variants. The variant list was not directly reported to us due to IRB constraints.

## 4. Discussion

Since the ACMG report on incidental findings was published, there has been much debate around explicitly searching for and reporting variants in the ACMG's gene list.<sup>8,13</sup> Issues have included the difficulty of substantiating which variants are pathogenic, the cost of additional screening, and the lack of information about how often variants are seen and how many each individual could possibly carry. Here, we present a tool, which serves as an example of how technicians, researchers, clinicians, and individuals may screen for potentially pathogenic and actionable variants. Furthermore, we have applied this tool on existing datasets and have made it available for public use in order to gauge the frequency that potentially pathogenic variants in the ACMG genes are observed.

#### 4.1. Variant Selection

One of the major issues was outlined in the original ACMG report: “The Working Group recognized that there is no single database currently available that represents an accurately curated compendium of known pathogenic variants, nor is there an automated algorithm to identify all novel variants meeting criteria for pathogenicity.”<sup>7</sup> For the purposes of this project, we selected the ClinVar database, because the variants submitted come directly from patient data. We selected only those variants that had at least one submission indicating that the variant was pathogenic in nature. A limitation of this approach is the inclusion of variants that may have conflicting submissions listing the variant as pathogenic and benign, and issues such as sample size and study population can contribute to this confusion about variant interpretation. However, the ClinVar curators are making an effort to review submissions. We recognize that variants labeled as pathogenic by ClinVar may not be viewed as so when analyzed by a clinical laboratory, genetic counselor, or clinician. However, their presence in a genome or exome will warrant evaluation in order to determine if they should be acted upon. Thus, understanding the frequency of such variants will allow us to draw conclusions about the amount of resources required to properly vet variants in the ACMG guidelines genes.

Another limitation of our database choice is that we do not pick up novel, rare, or private mutations that are not currently annotated in ClinVar. However, since we could not reliably make any inference about the pathogenicity of such variants, we selected not to include them in our publicly available tool. Finally, because most research studies are done in individuals of European descent, there is likely an overrepresentation of variants that are pathogenic in populations of European descent.<sup>14</sup>

We do note that the pathogenic variants in the ClinVar database are not evenly distributed between genes. The number of pathogenic variants reported in a gene can be influenced by several factors – including the length of the gene, the amount of selective pressure, and the number of studies focusing on the gene. Interestingly, *BRCA1* and *BRCA2* had the largest number of pathogenic variants. This could be due to the extensive studies on these genes and their role in hereditary breast and ovarian cancer.

#### 4.2. Findings in the 1000 Genomes Data and Personal Genomes Project

Our successful application of PATH-SCAN to the 1000 Genomes data sets confirmed the ability of our tool to process whole genomes. In 1092 low pass genomes, 566 individuals had a pathogenic variant in one of the ACMG genes.

The most observed variant was rs1805124 (H558R), seen in 41.2% of individuals. The population allele frequency of this variant is about 20% in 1000 Genomes. This is a prime example of the challenge with implementing an automatic system to follow up on potentially pathogenic variants in ACMG genes. *SCN5A* H558R has been associated with atrial fibrillation and changes in cardiac conduction.<sup>15,16</sup> Multiple studies have also demonstrated that the presence of this variant combined with other rare *SCN5A* variants perturbs heart electrophysiology.<sup>17–19</sup> However, there are also studies in which this variant may mitigate the effects of a particular mutation that causes Brugada syndrome.<sup>20</sup> Finally, it should be noted that this variant is quite

common in the general population. As Klitzman et al. noted in response to the ACMG Guidelines, ‘pathogenic’ variants with a high frequency in the population but a low corresponding disease prevalence may cause unnecessary alarm.<sup>13</sup> Because this variant can affect disease risk when other mutations are also present, its presence would require evaluation of the entire gene and family history by an experienced genetic counselor or clinician. This example supports the need for comprehensive follow up of variants that are thought to be pathogenic.

Excluding rs1805124 (H558R), 233 individuals out of 1092 carried an incidental finding. These other variants were less common, with less than 5% of individuals carrying any single variant. These variants included risks for such conditions as colon cancer (rs1801166) and cardiomyopathy (rs121912998), which can profoundly impact health and lifestyle.<sup>21,22</sup>

When we looked across the populations, we saw that there were differences in the average number of variants per person. Because many of these variants were derived from studies done in individuals of European ancestry, differences could be attributed to this selection bias.<sup>14</sup> Furthermore, different populations likely have different variants driving their total variant counts due to differences in population allele frequency. In the case of LWK, which had a very low average variant per person count when the most common variant was removed, we are likely missing population specific pathogenic variants. Another complex issue brought up by ancestry is pathogenicity – variants that may be causative and pathogenic in one population may not have the same penetrance or impact in another.<sup>14</sup> With our crowdsourcing tool, ancestry will be an option that individuals can submit; we hope that this will allow us to get a more accurate picture of the distribution of these variants across individuals of different and mixed ancestries.

We also note that since these are low coverage genomes (~4x), some variants reported could be false. Genomes sequenced to clinical standards would have much higher coverage and have more confident calls. Thus, this data may be skewed by false positives.

To evaluate our tool on Complete Genomics data and higher coverage genomes, we applied PATH-SCAN to 163 genomes made publicly available from the Personal Genomes Project. Once again, rs1805124 (H558R) was the most common variant. However, excluding this variant, 17% of genomes had variants of interest. Overrepresentation of certain variants may occur if individuals in the Personal Genome Project are related. Several of these variants were low frequency at a population level, as they did not appear in the 1000 Genomes data. Our tool assists in the evaluation of such variants by pinpointing them within minutes of scanning a genome.

#### **4.3. Using *PATH-SCAN* on Clinical Genomes**

Finally, we ran PATH-SCAN on a clinical genome sequencing cohort of fifteen individuals. The output provided a starting point for the evaluation of variants in the project. Previously, people used a gene-based approach to look at all variants in a gene of interest and then used manual curation to select variants for further evaluation.

#### **4.4. *PATH-SCAN* as a quantitative evaluation tool**

PATH-SCAN is a publicly available tool; individuals using it can choose to anonymously submit their pathogenic burden (i.e. the number of variants seen in their genome) and ancestry to our

server. Over time, we aim to use crowdsourcing to get a more accurate number of how often potentially pathogenic variants are seen and how ancestry affects these numbers.

The current iteration of our tool serves as the foundation for additional functionalities in development. Because ClinVar designated pathogenic variants may not truly be pathogenic, we are currently working on adding variant effect prediction scores, such as PolyPhen and SIFT to our tool.<sup>23,24</sup>

We have found that even with the most common pathogenic variant removed, a substantial percentage of individuals still carry variants in ACMG guidelines genes that require additional investigation. Of course, due to the limitations of the ClinVar database, many of these variants may be benign. However, we feel that each variant needs to be evaluated in the context of other mutations, clinical history, and family history by a clinician or genetic counselor. While not all of these variants may be ultimately reported back, evaluating these variants will require additional resources. Thus, understanding how often such variants occur is key to assessing the resource utilization of following the ACMG Guidelines. In the past few months, there has much debate surrounding the ACMG Guidelines and their implementation. Our tool PATH-SCAN aims to streamline the identification of variants in ACMG recommended genes that warrant further investigation and to provide data on how often each variant is seen.

## 5. Acknowledgments

RBA and TEK are funded by NIH/NIGMS NIGMS R24 GM61374. SBM is funded by the Edward Mallinckrodt Jr. Foundation. RD is funded by Stanford MSTP and T32 HG000044. ZZ is funded by NSF GRFP and T32 HG000044. KEO is funded by 5 P50 HG003389-05

## 6. Appendix A

### PATH-SCAN Manual

**Download** <http://montgomerylab.stanford.edu/pathscan.zip>

**Requirements:** Python 2.7.5; a web browser

### Command Line Interface

A full description of the CLI for PATH-SCAN follows:

```
$ python pathscan.py <genome file> [--suppress | --db <database>]
```

**<genome file>** is either a VCF file, a Complete Genomics TSV file, or a 23andMe SNP file.

**--suppress** If this flag is specified PATH-SCAN will only report data on the command line.

**--db <database file>** Can be used to specify a different database file. The database format is a TAB-delimited file with 9 columns, all required. First column is chromosome, second is position, third is RSID, fourth is the reference allele, fifth is the alternate allele, sixth is the gene name, seven is the gene review ID numbers (can be replaced with a '.'), eight is the OMIM ID number (can be replaced with a '.'), and the ninth is the clinical significance code from ClinVar (can be replaced with a '.').

## References

1. Gurwitz, D. & Bregman-Eschet, Y. Personal genomics services: whose genomes? European journal of human genetics : EJHG **17**, 883–9 (2009).

2. Kalf, R. R. J. et al. Variations in predicted risks in personal genome testing for common complex diseases. *Genetics in medicine : official journal of the American College of Medical Genetics* 1–7 (2013). doi:10.1038/gim.2013.80
3. Vernez, S. L., Salari, K., Ormond, K. E. & Lee, S. S.-J. Personal genome testing in medical education: student experiences with genotyping in the classroom. *Genome medicine* **5**, 24 (2013).
4. Manolio, T. a et al. Implementing genomic medicine in the clinic: the future is here. *Genetics in medicine : official journal of the American College of Medical Genetics* **15**, 258–67 (2013).
5. Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31–46 (2010).
6. Kohane, I., Masys, D. & Altman, R. The incidentalome: a threat to genomic medicine. *JAMA: the journal of the ...* **296**, 212–215 (2006).
7. Green, R. C. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics* (2013). doi:10.1038/gim.2013.73
8. Allyse, M. & Michie, M. Not-so-incidental findings: the ACMG recommendations on the reporting of incidental findings in clinical whole genome and whole exome sequencing. *Trends in biotechnology* **31**, 439–441 (2013).
9. Karczewski, K. & Tirrell, R. Interpretome: A freely available, modular, and secure personal genome interpretation engine. *Pac. Symp. ...* (2012).
10. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–45 (2004).
11. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
12. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nature reviews. Genetics* **9**, 406–11 (2008).
13. Klitzman, R., Appelbaum, P. S. & Chung, W. Return of Secondary Genomic Findings vs Patient Autonomy Implications for Medical Care. *JAMA* **310**, 369–370 (2013).
14. Rosenberg, N. a et al. Genome-wide association studies in diverse populations. *Nature reviews. Genetics* **11**, 356–66 (2010).
15. Chen, L. et al. Polymorphism H558R in the Human Cardiac Sodium Channel SCN5A Gene is Associated with Atrial Fibrillation. *Journal of International Medical Research* **39**, 1908–1916 (2011).
16. Gouas, L. et al. Association of KCNQ1, KCNE1, KCNH2 and SCN5A polymorphisms with QTc interval length in a healthy population. *European journal of human genetics: EJHG* **13**, 1213–22 (2005).
17. Cheng, J. et al. SCN5A rare variants in familial dilated cardiomyopathy decrease peak sodium current depending on the common polymorphism H558R and common splice variant Q1077del. *Clinical and translational science* **3**, 287–94 (2010).
18. Makielski, J. C. et al. A ubiquitous splice variant and a common polymorphism affect heterologous expression of recombinant human SCN5A heart sodium channels. *Circulation research* **93**, 821–8 (2003).
19. Ye, B., Valdivia, C. R., Ackerman, M. J. & Makielski, J. C. A common human SCN5A polymorphism modifies expression of an arrhythmia causing mutation. *Physiological genomics* **12**, 187–93 (2003).
20. Poelzing, S. et al. SCN5A polymorphism restores trafficking of a Brugada syndrome mutation on a separate gene. *Circulation* **114**, 368–76 (2006).
21. Frayling, I. & Beck, N. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proceedings of the ...* (1998).
22. Yang, Z. et al. Desmosomal dysfunction due to mutations in desmoplakin causes arrhythmogenic right ventricular dysplasia/cardiomyopathy. *Circulation research* **99**, 646–55 (2006).
23. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073–81 (2009).
24. Adzhubei, I. a et al. A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–9 (2010).



## IMPUTATION-BASED ASSESSMENT OF NEXT GENERATION RARE EXOME VARIANT ARRAYS

ALICIA R. MARTIN\*

*Department of Genetics & Biomedical Informatics Training Program, Stanford University  
Stanford, CA, 94305*

*Email: armartin@stanford.edu*

GERARD TSE

*Department of Computer Science, Stanford University  
Stanford, CA, 94305*

*Email: gerardtse@gmail.com*

CARLOS D. BUSTAMANTE

*Department of Genetics, Stanford University  
Stanford, CA, 94305*

*Email: cdbustam@stanford.edu*

EIMEAR E. KENNY\*

*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai  
New York, NY 10029*

*Email: eimear.kenny@mssm.edu*

A striking finding from recent large-scale sequencing efforts is that the vast majority of variants in the human genome are rare and found within single populations or lineages. These observations hold important implications for the design of the next round of disease variant discovery efforts—if genetic variants that influence disease risk follow the same trend, then we expect to see population-specific disease associations that require large sample sizes for detection. To address this challenge, and due to the still prohibitive cost of sequencing large cohorts, researchers have developed a new generation of low-cost genotyping arrays that assay rare variation previously identified from large exome sequencing studies. Genotyping approaches rely not only on directly observing variants, but also on phasing and imputation methods that use publicly available reference panels to infer unobserved variants in a study cohort. Rare variant exome arrays are intentionally enriched for variants likely to be disease causing, and here we assay the ability of the first commercially available rare exome variant array (the Illumina Infinium HumanExome BeadChip) to also tag other potentially damaging variants not molecularly assayed. Using full sequence data from chromosome 22 from the phase I 1000 Genomes Project, we evaluate three methods for imputation (BEAGLE, MaCH-Admix, and SHAPEIT2/IMPUTE2) with the rare exome variant array under varied study panel sizes, reference panel sizes, and LD structures via population differences. We find that imputation is more accurate across both the genome and exome for common variant arrays than the next generation array for all allele frequencies, including rare alleles. We also find that imputation is the least accurate in African populations, and accuracy is substantially improved for rare variants when the same population is included in the reference panel. Depending on the goals of GWAS researchers, our results will aid budget decisions by helping determine whether money is best spent sequencing the genomes of smaller sample sizes, genotyping larger sample sizes with rare and/or common variant arrays and imputing SNPs, or some combination of the two.

---

\* Corresponding authors

## 1. Introduction

The ability to measure human genetic variation on a genome-scale reliably and inexpensively in research settings has fueled and shaped the movement toward personalized medicine in health care. A prominent strategy for discovering genetic variants underlying disease susceptibility is through genome-wide association studies (GWAS), in which a subset of genetic variation is observed or inferred via linkage disequilibrium (LD), and correlated with disease state. GWAS have been successful in identifying thousands of reproducible associations with complex disease, which have had some utility in clinical practice<sup>1,2</sup>. However, most variants identified in GWAS with genotyping arrays are of small effect and fail to explain a large portion of genetic variation, even when the disease is estimated to be highly heritable<sup>3</sup>. Population genetics and neutral theory suggest that common variation might be less important than rare variation in these cases because selective pressure has had more time to eliminate deleterious alleles. With the advent of next generation sequencing technology, large consortia seeking to identify nonsynonymous coding changes have emerged. A salient result of these large-scale projects is that the vast majority of genetic variation is rare and exhibits little sharing among diverged populations<sup>4-6</sup>. The sequencing costs for an exome still outweigh those of genotyping arrays, however, and large sample sizes are required to detect rare variants. This creates a budget dilemma for GWAS researchers trying to explain the genetic basis of disease regarding the number of individuals they can afford to study with sequencing versus genotyping methods.

As a consequence of these findings, researchers have designed a next generation genotyping array that enriches for nonsynonymous rare coding variants. More than 15 labs with exome sequencing data from ~12,000 individuals contributed to the ascertainment of SNPs to include in the first rare variant array. The current design of the first publicly available next generation array, the Illumina Infinium HumanExome BeadChip, consists of only ~250,000 variants, a fraction of the sites that most common variant arrays currently assay. The vast majority of sites are rare coding variants; the remaining sites include randomly selected synonymous single nucleotide polymorphisms (SNPs), Native American and African ancestry informative markers, GWAS tag SNPs, HLA tags, common scaffold SNPs, and ~2,000 variants from other functional classes. A potential way to bolster the number of sites is through statistical inference of variants not molecularly assayed on the genotyping array through phasing and imputation guided by publicly available reference panels<sup>4,7,8</sup>. Phasing and imputation methods rely on the correlated inheritance between neighboring alleles or linkage disequilibrium (LD) between assayed alleles. LD is substantially reduced between variants on the rare exome array overall, however, because the number of scaffold SNPs is substantially reduced compared to other GWAS arrays (5,286 SNPs total compared to hundreds of thousands on common variant arrays). Admixture mapping, an approach often used when ancestry confounds GWAS associations, also relies heavily on a dense scaffold of linked markers. For example, results from HapMix, a method for inferring local ancestry across chromosomes, indicated that accuracy is reduced with fewer than 50,000 scaffold markers even when admixture is recent<sup>9</sup>.

In order to better understand the amenability of rare exome variant arrays to existing phasing and

imputation methods, we have performed evaluations of multiple LD-based methods as well as parameters that influence imputation accuracy, including sample size and population. We find that imputation with common variant arrays is more accurate across both the exomic and genomic regions of chromosome 22, highlighting the importance of contextual variants in imputation and suggesting that the Illumina Infinium HumanExome BeadChip is not ideal for imputation purposes.





## 2. Methods

### 2.1. Evaluation overview

We based all our evaluation on the data provided by the phase I 1000 Genomes project<sup>10</sup>, wherein 1,092 individuals from 14 distinct populations were genome sequenced, exome sequenced, and genotyped to produce an integrated variant call set. These populations include three African populations, three East Asian populations, five European populations, as well as three populations from the Americas. We created a pipeline (Figure 1) to perform phasing and imputation using three methods: BEAGLE v3.3.2<sup>11,12</sup> for both phasing and imputation, MaCH-Admix<sup>8</sup> v2.0.198 for both phasing and imputation, and ShapeIt<sup>13,14</sup> v2.r644 for phasing followed by Impute2<sup>15,16</sup> v2.2.2 for imputation (process abbreviated as SHAPEIT2/IMPUTE2).

To fairly evaluate phasing and imputation performance we compared one rare and one common variant array of approximately the same SNP density (the Illumina Infinium HumanExome BeadChip and Illumina Infinium HumanHap 300v1 containing ~250K and ~300K SNPs, respectively). To evaluate performance versus cost trade-offs, we also included two higher-cost, higher-density common variant arrays, the Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina Human Omni2.5 BeadChip containing 1M and 2.5M SNPs, respectively. To generate the phasing and imputation results for each array, we sampled individuals into a reference panel and a test set. The reference panel contained all of the sequence calls on chromosome 22, while the test set was further filtered to the markers on each of the corresponding arrays (Table 1). We generated a known truth set from the full phase I integrated call set and imputed set using the imputed sites not on each of the evaluated arrays for each run for accuracy evaluation.

Table 1 - Arrays evaluated in this study and number of sites across all of chromosome 22 versus exomic regions of chromosome 22. Exome sites were filtered using sites annotated with EXOME in the phase I 1000 Genomes integrated call set info fields and are a subset of Genome sites. Minor allele frequency (MAF) distributions are as assessed in the 1000 Genomes phase I samples across all chromosome 22 sites and are drawn for each array from a frequency of 0 – 0.5. “Dark sites” are the sites that are on the array but not in the 1000 Genomes phase I reference panel.

Array	Genome	Exome	MAF distributions	Mean MAF	Dark sites (%)
Illumina HumanOmni2.5 BeadChip	33,188	1,631		0.173	6.99
Affymetrix Genome-Wide Human SNP Array 6.0	11,739	262		0.208	1.01
Illumina Infinium HumanHap 300v1	5,376	240		0.272	0.99
Illumina Infinium HumanExome BeadChip	3,442	3,009		0.050	69.81
Total reference panel sites	475,372	16,885			

Simulated data from each of the four arrays were run through the phasing and imputation pipeline. The reference panel for each run was used as an input to the pipeline to inform the phasing and imputation algorithms. The pipeline first phased the incomplete genotypes in the test set, then imputed markers up to the reference panel markers using the same test set markers as in the phasing step as a scaffold (Figure 1). In order to speed up computational run time, we split the reference panel sites into 5 Mb windows with 250 kb flanking on either ends that were removed in post-processing to reduce edge effects between windows. We ran separate instances of imputation for each chunk in parallel, enabling the pipeline to run with reasonable memory and in reasonable time. At the end of each run, we extracted the imputed genotypes and each algorithm's confidence score ( $R^2$  in the cases of BEAGLE and MaCH-Admix and informative measure in the case of Impute2). We calculated diploid and haploid error for each imputed site from the known truth data.

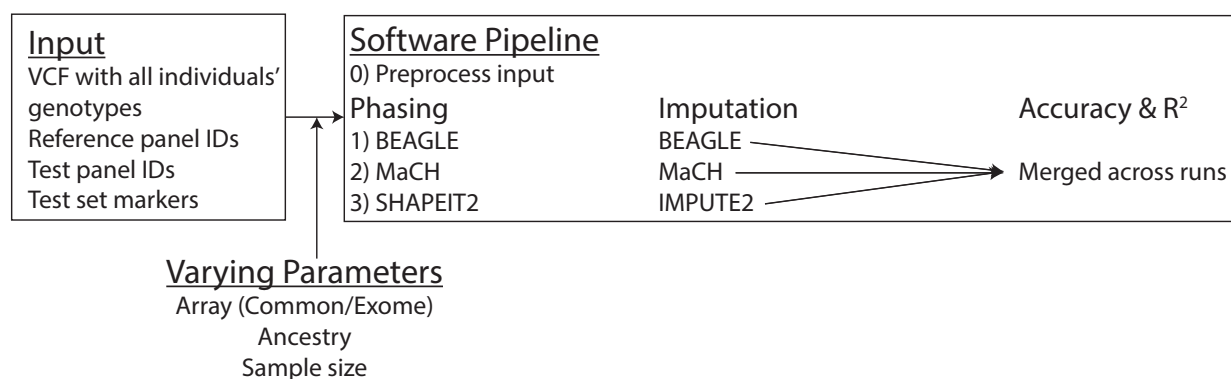


Figure 1 - Phasing and imputation pipeline. Inputs files are subsetting based on varying parameters specified, and for each set of parameters phasing and imputation was performed using three methods.

## 2.2. Sampling strategy for test/reference size analyses

Previous studies have assessed imputation accuracy on single chromosomes, including chromosomes 10 (~135 Mb), 20 (~62 Mb), and 22 (50 Mb), and have found highly consistent results<sup>7,15,16</sup>, indicating that they are representative. As such, we used full sequence data from chromosome 22 for computational efficiency from all 1,092 individuals and sampled them randomly into two groups: A reference panel and a test set. To study the effect of different reference panels and GWAS study sizes on the accuracy of imputed haplotypes, we investigated 13 different configurations of test set and reference panel sizes: a test set of size 92 with varying reference panel sizes of 63, 125, 250, 500, and 1000; and test panel sizes of 300 and 500, each with reference panels of 62, 125, 250, and 500.

Using the reference panel to inform phasing and imputation, we ran the pipelines for each of the three common variant arrays and the rare exome array and collected the results. The results were compared to the true calls found in the unfiltered genotypes of individuals in the test set.

### 2.3. *Sampling strategy for population analyses*

We used full sequence data from all of the 1,092 individuals and separated them into 14 populations. Four different sampling strategies were employed to identify biases when different reference sets are used for each of the 14 populations, resulting in 56 sets of samplings, as follows. The first two samplings assessed imputation accuracy when a test population is not or is included in the reference panel, respectively. We created a test set with all individuals in each population and sampled 900 individuals from the rest of the genomes available in the 1000 Genomes project (strategy A, Figure 3). As a control for the presence of a population from the reference panel, we created another test set with half of all the individuals in each population and put the remaining half of the population in the reference panel, then added individuals from other populations randomly until the reference panel contained 900 individuals (strategy B).

The other two population samplings focused on the significance of having individuals from the same continent in the reference panel. We created a test set with 33 individuals in the population and sampled 148 from all other individuals from the same continental group (strategy C). These numbers were chosen for uniformity across populations in order to represent the smallest continental group in the data. We performed this evaluation for each population and considered four continental groups: Africans, Asians, Europeans, and Native Americans. As a control, we created another test set with 30 individuals in the population and sampled 148 from all other individuals regardless of origin (strategy D).

### 2.4. *Phasing and imputation summaries and analysis*

Using the reference panel to inform phasing and imputation, we ran the pipelines for each of the three common variant arrays and the rare exome array. The imputed genotypes were compared to the true calls in the unfiltered sequences of individuals in the test set. Data summaries for all three algorithms reported an informative metric ( $R^2$ ), which were generated by the imputation algorithms. Because each algorithm calculates  $R^2$  differently, we calculated diploid and haploid error, as well as minor allele frequency (MAF), in order to fairly compare the algorithms directly. We define the diploid error as any discordance between the most likely imputed and true calls, which is affected by MAF and therefore only used to compare method performances. In this scenario, if the true variant is homozygous reference, heterozygous or homozygous non-reference imputation dosages count equally toward the error. We also calculated haploid error, where in the previous scenario, a heterozygous call counts half as much toward the error as a homozygous non-reference call, which was highly correlated (>99%) with diploid error. We note that the diploid and haploid errors are critical to examine but that they are highly influenced by MAF. For example, at a site where a very rare variant exists in the reference panel, error is very low because the imputation algorithm frequently fills in the major allele, even in the absence of any surrounding variants. In contrast, when a common variant exists, the imputation algorithms require more neighboring information to correctly impute the variant. For these reasons, we assess imputation accuracy as  $R^2$  as previously<sup>15</sup>, except where

otherwise noted. In order to compare MAF versus imputation accuracy, we performed local regression weighted by least squares. Unless otherwise noted, the span was 0.75.

### 3. Results

We first compared the performance of three phasing and imputation algorithms, BEAGLE, MaCH-Admix, and SHAPEIT2/IMPUTE2 under multiple conditions. The informative measure metrics are defined slightly differently for each algorithm<sup>7</sup>, and in all cases SHAPEIT2/IMPUTE2 reports the highest informative measures (data not shown). In order to determine which method was performing most accurately based on known truth data, we compared their performance via mean diploid error across all test panel sizes, reference panel sizes, and the four arrays we evaluated, as outlined in Methods. In each case, BEAGLE had the highest error, SHAPEIT2/IMPUTE2 performed comparably with MaCH-Admix, and MaCH-Admix resulted in the lowest error, which highlights the importance of using a directly comparable metric to assess method performance. Table 2 shows the average diploid error across chromosome 22 across all reference and test panel sizes using the Affymetrix Genome-Wide Human SNP Array 6.0 for each, which showed the same trends with other arrays (data not shown). Because MaCH-Admix resulted in the lowest imputation error, all following analyses show results using this method.

Table 2 - Diploid error across multiple sample sizes. Reported values are mean percentages across all variant sites in the phase I 1000 Genomes Project on chromosome 22 using sites on the Affymetrix Genome-Wide Human SNP Array 6.0 as test markers. Individuals in the test and reference panel are the same across methods for each comparison. Imputation  $R^2$  values are shown for each algorithm, which are defined differently for each algorithm. Note that BEAGLE  $R^2$  averages are calculated only for values that are not “NaN,” which likely increases the  $R^2$  reported with respect to other algorithms.

Test panel size	Reference panel size	BEAGLE (%)	MaCH-Admix (%)	Shapeit+Impute2 (%)	BEAGLE ( $R^2$ )	MaCH-Admix ( $R^2$ )	Shapeit+Impute2 ( $R^2$ )
500	500	6.36	4.21	4.35	.7349	<b>.3762</b>	<b>.5604</b>
500	250	6.37	4.27	4.38	.7329	.3333	.4735
500	125	6.63	4.41	4.56	.6820	.2959	.4048
500	62	6.77	4.63	4.74	.7403	.2464	.3175
300	500	6.31	4.16	4.32	.7387	.3724	.5348
300	250	6.60	4.39	4.56	.7392	.3279	.4567
300	125	6.57	4.37	4.53	.7344	.2954	.3928
300	62	6.87	4.66	4.79	.7331	.2513	.3191
92	1000	<b>6.36</b>	<b>4.13</b>	<b>4.30</b>	<b>.7653</b>	.3503	.4655
92	500	6.49	4.25	4.45	.7637	.3401	.4482
92	250	6.37	4.17	4.33	.7467	.3081	.3978
92	125	6.59	4.51	4.65	.7481	.2799	.3540
92	63	6.68	4.40	4.59	.7123	.2506	.3033

We next evaluated the impact of test and reference panel sizes on imputation accuracy, as assessed by  $R^2$ , for the four arrays described previously (Figure 2). We compared three test panel sizes (92, 300, and 500) and find that in all cases, larger test panels have greater imputation accuracy,

indicating that phasing and imputing a full study set together improves imputation accuracy. We also find that reference panel size has a greater impact on imputation accuracy than test panel size when the test panel contains greater than 92 individuals. These results indicate that large reference panels are necessary to accurately impute variants.

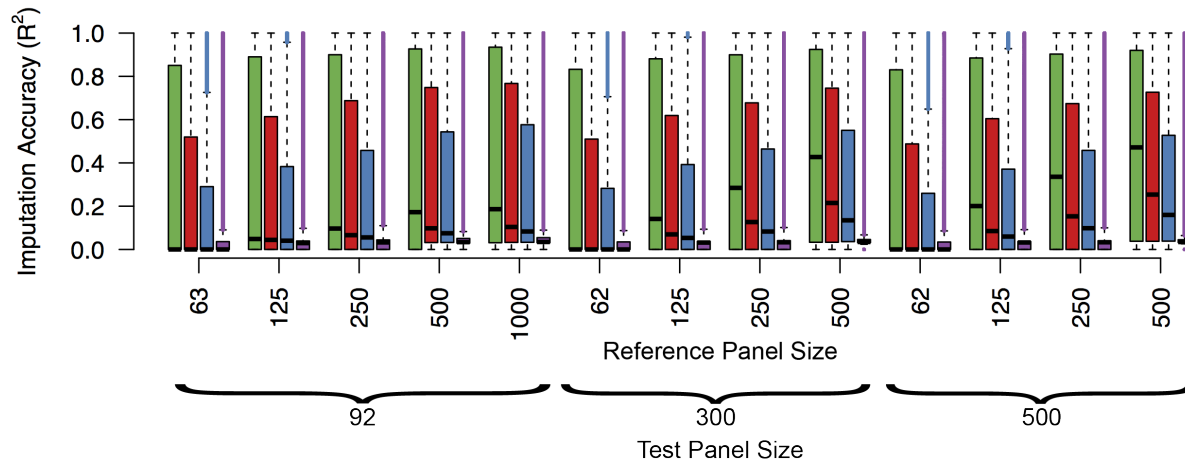


Figure 2 - Imputation accuracy across varying reference and test panel sizes. Phasing and imputation was performed using MaCH-Admix. Test panel markers were ascertained on chromosome 22 using sites from four arrays in the following colors: green – Illumina HumanOmni2.5 BeadChip, red – Affymetrix Genome-Wide Human SNP Array 6.0, blue – Illumina Infinium HumanHap 300v1, purple – Illumina Infinium HumanExome BeadChip. On the x-axis, the first number indicates the number of individuals included in the test panel, and the second number is the number of individuals included in the reference panel.

The effect of reference panel size on imputation accuracy is especially pronounced when fewer markers are assayed. For example, imputation accuracy is not substantially reduced for most common sites across chromosome 22 ( $MAF > 5\%$ ) when the reference panel size is reduced from 500 individuals to only 62 individuals using the dense Illumina HumanOmni2.5 BeadChip, and most common sites maintain an  $R^2$  of  $\sim 0.9$ . In contrast, the accuracy drops considerably between a reference panel size of 500 versus 62 with the sparser Illumina Infinium HumanHap 300v1 (e.g. reduction of 13% from  $R^2=0.772$  to  $0.669$  at  $MAF=0.3$ ) and Illumina Infinium HumanExome BeadChip arrays (e.g. reduction of 26% from  $R^2=0.146$  to  $0.108$  at  $MAF=0.3$ ). We also find that accuracy plateaus as a function of minor allele frequency ( $MAF$ ). Additionally, invariant reference panel SNPs likely drive the number of “dark sites” on each array (Table 1). Interestingly, the  $MAF$  at which accuracy peaks is array-specific. For example, the Illumina Infinium HumanHap 300v1 array has a similar number of sites on chromosome 22 as the Illumina Infinium HumanExome BeadChip (Table 1); however, accuracy peaks around  $MAF=0.3$  on the Illumina 300k array and around  $MAF=0.5$  on the exome array. Interestingly, imputed exome rare variant array sites from genome-wide arrays are imputed more accurately than across all chromosome 22 sites for varying allele frequencies (Figure 4A-C versus Figure 4I-K), likely because scaffold sites on genome-wide arrays are enriched near exonic regions, improving imputation accuracy.

Previous work has indicated that reference panels that share more haplotypes with the study panel improve imputation accuracy compared to a random panel<sup>17</sup>. We compared multiple population stratifications as described in Section 2.3 (Figure 3). In all scenarios, imputation performs the poorest in individuals of African descent. This is likely due to the reduced LD structure in African populations<sup>18</sup> and European ascertainment bias in genotyping arrays<sup>19</sup>. Imputation with both global reference panel strategies with a larger number of reference individuals, albeit from more distantly related populations overall (Figure 3A and Figure 3B), outperforms imputation with smaller continental reference panels (Figure 3C and Figure 3D). Low frequency alleles are imputed with greater accuracy when the reference panel includes individuals from the same population compared to when it does not (Figure 3B versus Figure 3A). This is especially true in European populations with the exception of TSI individuals, which likely arises from the greater genetic diversity and more complicated demographic history present in Italy compared to other European populations presented here<sup>20,21</sup>.

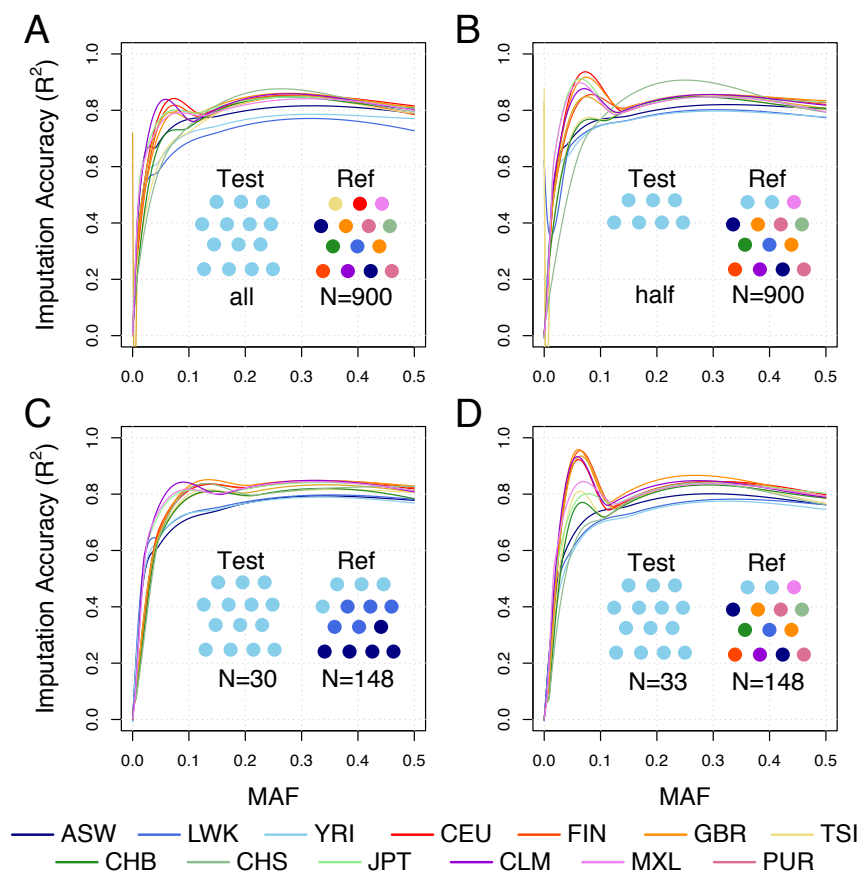


Figure 3 - Variability in imputation accuracy across populations. All simulations were performed using the Affymetrix Genome-Wide Human SNP Array 6.0 markers from chromosome 22 in the test set. Lines are local regression fits to the data, and local peaks near MAF=0 in A and B for the GBR and TSI, respectively, are simply due to smoothing edge



effects. A) Strategy A. B) Strategy B. C) Strategy C. D) Strategy D. Diagrams drawn under loess curves are cartoons of sampling strategies, as outlined in section 2.3. Abbreviations are as follows: ASW=HapMap African ancestry individuals from SW US, LWK=Luhya individuals, YRI=Yoruba individuals, CEU=CEPH individuals, FIN=HapMap Finnish individuals from Finland, GBR=British individuals from England and Scotland, TSI=Toscan individuals, CHB=Han Chinese in Beijing, CHS=Han Chinese South, JPT=Japanese individuals, CLM=Colombian in Medellin, Colombia, MXL=HapMap Mexican individuals from LA California, PUR=Puerto Rican in Puerto Rico.

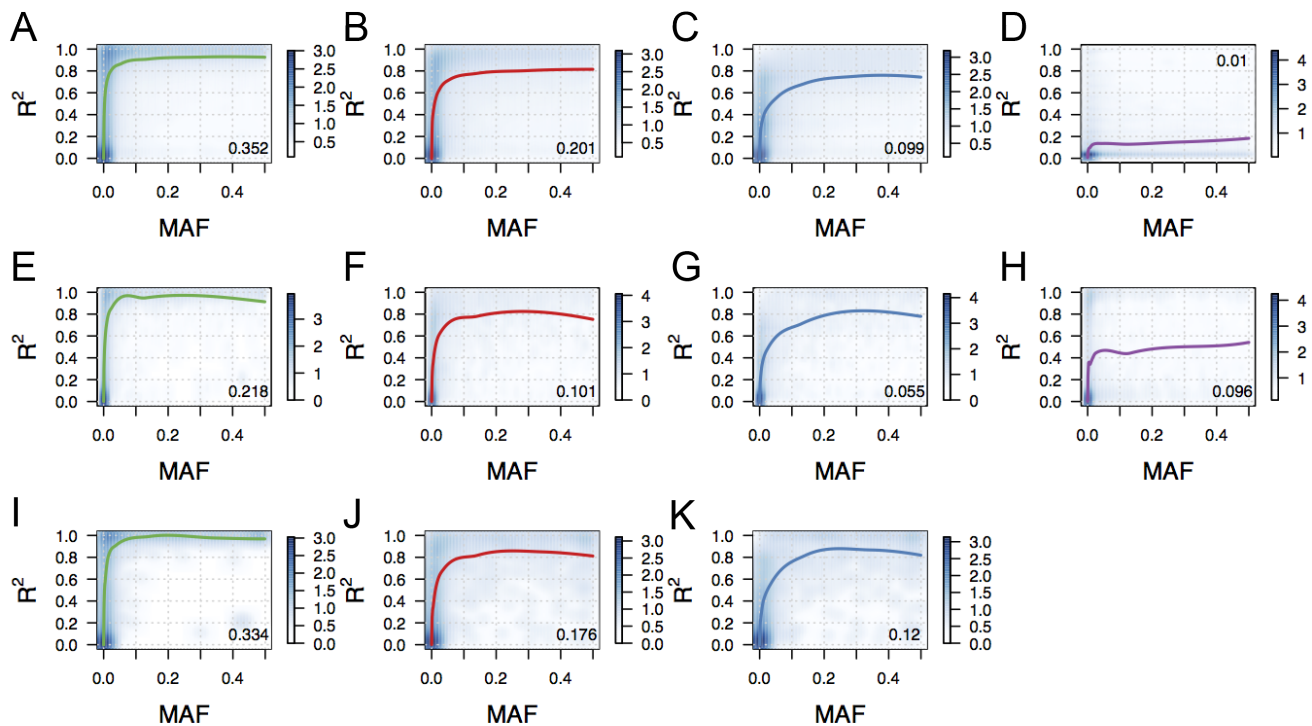


Figure 4 - Imputation accuracy across three common variant and one rare exome variant arrays in genomic, exomic, and imputable exome rare variant array regions of chromosome 22. Colors correspond with arrays, as in Figure 2. All subpanels show smoothed scatter plots with an overlaid local regression fit, and the proportion of sites imputed with  $R^2 > 0.8$  is reported, which are consistent with previous results<sup>22</sup>. Span was adjusted to 0.5 in order to keep the fits within the bounds of the data. A-D) genomic regions of chromosome 22; E-H) exomic regions of chromosome 22; I-K) Imputation accuracy for imputable exome rare variant sites using each of the genome-wide common variant arrays.

We next compared imputation accuracy across three common variant and one rare exome variant genotyping array platforms. As expected, the common variant arrays impute sites across chromosome 22 more accurately than the Illumina Infinium HumanExome BeadChip. Surprisingly, all three common variant arrays also outperform the exome array in imputing the exome-only regions, though their accuracy is substantially reduced in the exome compared to the genome (Figure 4). Imputation accuracy is the poorest with the rare variant exome array, even though the Illumina 300k common variant array has slightly fewer assayed variants on chromosome 22 (Table 1). Aside from the exome array, accuracy improves with arrays tagging more variants, as expected. The accuracy in the rare variant exome array is increased in the exomic regions compared to all chromosome 22 variants

(Figure 4H and Figure 4D, respectively). As shown in Figure 4, the imputable exome variant sites are imputed with similar accuracy as all sites across chromosome 22 with common genome-wide arrays as a scaffold. While the “dark sites” on the exome chip will be missed, other imputable sites, which are enriched for biomedically relevant SNPs, are imputed with similar accuracy as any similar frequency SNP.

#### 4. Discussion

We have evaluated multiple factors that influence imputation accuracy, including test and reference panel size, phasing and imputation methods, populations, and genotyping arrays. We find that both larger reference and test panels lead to greater imputation accuracy, and that reference panel size is more important than test panel size in most GWAS scenarios. Larger reference panels, regardless of population, aid imputation performance for common variants, while more closely related reference panels are critical for accurately imputing rare variants. Comparing three methods, our simulations revealed that BEAGLE was both the most computationally costly method (e.g. ~48 hours to run and 10.5G of memory for chromosome 22 with a reference size of 500 and test size of 500) and had the least accurate performance. SHAPEIT2/IMPUTE2 and MaCH-Admix were comparable in terms of computationally efficient (2 hours to run and 2G of memory versus 3.5 hours to run and 1G of memory with the same test and reference panel as in the BEAGLE case). These computational costs are consistent with previously reported values<sup>8</sup>.

It is important to note that there is an obvious bias in imputation accuracy across populations, with the lowest accuracy in African populations. Greater accuracy in out-of-Africa groups is likely due to ascertainment bias as well as longer haplotypes from the serial founder effect during the peopling of the globe. We see improved imputation accuracy at the rare end of the allele frequency spectrum when the reference panel includes the same population as the test panel. These results suggest that nearby reference panels are especially important for large outbred groups.

Imputation with common variant arrays substantially outperforms imputation with the Illumina Infinium HumanExome BeadChip. This reduction in accuracy is apparent for all frequencies, including rare alleles, suggesting that covariance between rare and nearby alleles is low, and alleles are tagged poorly. This is likely in part due to the uneven distribution of variants on the exome array across the chromosome, reducing LD on the array. A scaffold of genomic variants will likely aid imputation accuracy in exome arrays. One potential way to assay a large number of rare variants accurately without losing important rare variant information is to combine arrays, coupling the exome array with one of the common arrays we evaluated, for example. The improved imputation accuracy by the exome array in exomic regions is likely due to denser markers and greater LD in this region. The reduction of imputation accuracy in exomic regions with the common variant arrays may be due to greater sequencing depth in the 1000 Genomes Project in the integrated call set, which contains, genotyping, genome-, and exome-sequencing data, leading to more low frequency calls passing variant filters.

Finally, alternative algorithms for phasing<sup>23,24</sup> that rely on identity-by-descent (IBD) structure preferentially rather than LD have recently been published. These methods take advantage of haplotypic structure and will likely aid imputation differentially depending on the degree of sharing within a population and the potential to improve phasing accuracy. A question for future work, for example, might compare phasing accuracy using LD-based and IBD-based methods in endogamous African populations where imputation with traditional arrays performs poorly but where cryptic relatedness is more likely to exist.

## 5. Conclusions

The next generation of genotyping arrays intends to capture rare, coding variation that is likely to contain more pathogenic variation than randomly ascertained SNPs. Here, we assess the ability of a commercially available rare variant exome array to adequately tag variation that has not been directly assayed, compared to common variant arrays. We assess multiple methods, sample sizes, and populations, and find that imputation accuracy is substantially reduced with the rare variant exome array compared to common variant arrays. This result is true both in genomic and exomic regions of chromosome 22, although the difference in imputation accuracy between common and exome arrays is reduced in exomic regions. We also find that the European ascertainment bias in common variant arrays is reflected in imputation accuracy across populations, with most European variants imputed more accurately than those of other continental groups. Additionally, closely related populations are critical in reference panels for low frequency variants. Finally, we compare three phasing and imputation methods and find that BEAGLE is the least accurate, and SHAPEIT2/IMPUTE2 performs slightly less accurately than MaCH-Admix for all reference and test panel sizes. This research provides guidelines for GWAS researchers to avoid the current design of exome rare variant arrays when imputing genotype data. We acknowledge, however, that these next generation arrays have potential utility when fine-mapping a variant that is suspected to be coding and not tagged by common variant genotyping arrays.

## Acknowledgments

We thank the instructors, Russ B. Altman, Steven C. Bagley, and Hua Fan-Minogue, and students of the Stanford University Biomedical Informatics Project Course (BMI 212, Spring 2013) for their feedback. We also thank Xueheng Zhao for his helpful discussions. ARM was funded by the NIH-NIGMS Genetics & Developmental Biology Training Program (NIH GM007790).

## Appendix

All code written to run phasing and imputation simulations on a Sun Grid Engine can be downloaded here: [https://github.com/armartin/compare\\_impute](https://github.com/armartin/compare_impute).

## References

1. Hindorf, L. a *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–7 (2009).
2. Manolio, T. a. Bringing genome-wide association findings into clinical use. *Nature reviews. Genetics* **14**, 549–58 (2013).
3. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics* **11**, 415–25 (2010).
4. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
5. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11983–8 (2011).
6. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
7. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499–511 (2010).
8. Liu, E. Y., Li, M., Wang, W. & Li, Y. MaCH-admix: genotype imputation for admixed populations. *Genetic epidemiology* **37**, 25–37 (2013).
9. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5**, e1000519 (2009).
10. Project, G., Asia, E., Africa, S., Figs, S. & Tables, S. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135**, 0–9 (2012).
11. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics* **84**, 210–23 (2009).
12. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**, 1084–97 (2007).
13. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5–6 (2013).
14. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179–81 (2012).
15. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)* **1**, 457–70 (2011).
16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529 (2009).
17. Huang, L. *et al.* Haplotype variation and genotype imputation in African populations. *Genetic epidemiology* **35**, 766–80 (2011).
18. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5154–62 (2011).
19. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and evolution* **27**, 2534–47 (2010).
20. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *European journal of human genetics : EJHG* **21**, 659–65 (2013).
21. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS biology* **11**, e1001555 (2013).
22. Nelson, S. C. *et al.* Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. *G3 (Bethesda, Md.)* (2013). doi:10.1534/g3.113.007161
23. Palin, K., Campbell, H., Wright, A. F., Wilson, J. F. & Durbin, R. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology* **35**, 853–60 (2011).
24. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *American journal of human genetics* **91**, 238–51 (2012).

UTILIZATION OF AN EMR-BIOREPOSITORY TO IDENTIFY THE GENETIC PREDICTORS OF  
CALCINEURIN-INHIBITOR TOXICITY IN HEART TRANSPLANT RECIPIENTS

MATTHEW OETJENS<sup>†</sup>

*Center for Human Genetics Research,*  
*Email: [matthew.t.oetjens@vanderbilt.edu](mailto:matthew.t.oetjens@vanderbilt.edu)*

WILLIAM S. BUSH

*Department of Biomedical Informatics, Center for Human Genetics Research,*  
*Email: [william.s.bush@vanderbilt.edu](mailto:william.s.bush@vanderbilt.edu)*

KELLY A. BIRDWELL

*Department of Medicine,*  
*Email: [kelly.birdwell@vanderbilt.edu](mailto:kelly.birdwell@vanderbilt.edu)*

HOLLI H. DILKS

*Vanderbilt Technologies for Advanced Genomics Core Facility,*  
*Email: [holli.dilks@chgr.mc.vanderbilt.edu](mailto:holli.dilks@chgr.mc.vanderbilt.edu)*

ERICA A. BOWTON

*Office of Personalized Medicine,*  
*Email: [erica.a.bowton@vanderbilt.edu](mailto:erica.a.bowton@vanderbilt.edu)*

JOSHUA C. DENNY

*Department of Biomedical Informatics,*  
*Email: [josh.denny@vanderbilt.edu](mailto:josh.denny@vanderbilt.edu)*

---

<sup>†</sup> The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. A portion of the genome-wide genotyping was funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. This study was supported in part by RC2 GM092618 from NIGMS/NIH and U19 HL 065962 from NHLBI/NIH. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

RUSSELL A. WILKE

*Department of Medicine, Department of Pharmacology,*  
*Email: [russell.a.wilke@vanderbilt.edu](mailto:russell.a.wilke@vanderbilt.edu)*

DAN M. RODEN

*Department of Medicine, Department of Pharmacology, Office of Personalized Medicine,*  
*Email: [dan.roden@vanderbilt.edu](mailto:dan.roden@vanderbilt.edu)*

DANA C. CRAWFORD<sup>†</sup>

*Department of Molecular Physiology and Biophysics, Center for Human Genetics Research,*  
*Vanderbilt University, 2215 Garland Ave*  
*Nashville, TN 37212, United States of America*  
*Email: [crawford@chgr.mc.vanderbilt.edu](mailto:crawford@chgr.mc.vanderbilt.edu)*

Calcineurin-inhibitors CI are immunosuppressive agents prescribed to patients after solid organ transplant to prevent rejection. Although these drugs have been transformative for allograft survival, long-term use is complicated by side effects including nephrotoxicity. Given the narrow therapeutic index of CI, therapeutic drug monitoring is used to prevent acute rejection from underdosing and acute toxicity from overdosing, but drug monitoring does not alleviate long-term side effects. Patients on calcineurin-inhibitors for long periods almost universally experience declines in renal function, and a subpopulation of transplant recipients ultimately develop chronic kidney disease that may progress to end stage renal disease attributable to calcineurin inhibitor toxicity (CNIT). Pharmacogenomics has the potential to identify patients who are at high risk for developing advanced chronic kidney disease caused by CNIT and providing them with existing alternate immunosuppressive therapy. In this study we utilized BioVU, Vanderbilt University Medical Center's DNA biorepository linked to de-identified electronic medical records to identify a cohort of 115 heart transplant recipients prescribed calcineurin-inhibitors to identify genetic risk factors for CNIT. We identified 37 cases of nephrotoxicity in our cohort, defining nephrotoxicity as a monthly median estimated glomerular filtration rate (eGFR)  $<30 \text{ mL/min/1.73m}^2$  at least six months post-transplant for at least three consecutive months. All heart transplant patients were genotyped on the Illumina ADME Core Panel, a pharmacogenomic genotyping platform that assays 184 variants across 34 genes. In Cox regression analysis adjusting for age at transplant, pre-transplant chronic kidney disease, pre-transplant diabetes, and the three most significant principal components (PCAs), we did not identify any markers that met our multiple-testing threshold. As a secondary analysis we also modeled post-transplant eGFR directly with linear mixed models adjusted for age at transplant, cyclosporine use, median BMI, and the three most significant principal components. While no SNPs met our threshold for significance, a SNP previously identified in genetic studies of the dosing of tacrolimus *CYP3A5* rs776746, replicated in an adjusted analysis at an uncorrected p-value of 0.02 (coeff(S.E.) = 14.60(6.41)). While larger independent studies will be required to further validate this finding, this study underscores the EMRs usefulness as a resource for longitudinal pharmacogenetic study designs.

## 1. Introduction

Calcineurin-inhibitors (CI), such as tacrolimus and cyclosporine, are immunosuppressants prescribed to recipients of allografts to reduce the risk of rejection by the immune system. These drugs function by

dampening IL-2 signaling pathway in T-cells and avoid the vigorous inflammation and tissue damage typical of an alloresponse. While these drugs have led to dramatically improved survival among heart transplant recipients, the nephrotoxic side-effects of these drugs continue to diminish the long-term survival rates among patients [1;2]. CI are dosed in a narrow therapeutic window requiring close monitoring of serum drug levels to prevent allograft rejection while minimizing the risk of adverse events.

Post-transplant, patients undergo continuous monitoring of their serum creatinine and glomerular filtration rates (GFR) to determine impact of the immunosuppressants on kidney function. A decline in kidney function is nearly universal among heart transplant recipients with significant variability in the development of severe kidney disease. Patients are frequently faced with the development of chronic kidney disease (CKD) which is classified into 5 stages of increasing severity, each defined by the estimated GFR. In a retrospective study 352 heart transplant recipients, 3% developed end-stage renal disease or CKD Stage 5 by 5 years and 12% by 10 years [3]. Clinical risk factors for developing post-transplant CKD include pre-transplant GFR, pre-transplant diabetes mellitus, a female cardiac donor, gender of the recipient, and post-operative renal replacement therapy [3].

Despite vast structural differences, the pharmacokinetics of cyclosporine and tacrolimus are surprisingly similar, and both agents are targets of the P-gp efflux pump *ABCB1* and the cytochrome p450 *CYP3A* family of enzymes [4]. These genes are polymorphic for functional alleles, and variants have been examined in several pharmacogenetic studies of calcineurin-inhibitor dosing and nephrotoxicity in renal transplants [5-8]. Despite a large number of candidate gene studies on the effects of these variants on immunosuppression therapy, many of these analyses are narrow in their scope of genes tested. In this study, we explored the roles of other pharmacokinetic genes outside the *CYP3A* family and *ABCB1* on the development of calcineurin inhibitor nephrotoxicity CNIT. For our study, we identified 127 heart transplant recipients in BioVU, Vanderbilt University Medical Center's DNA biorepository linked to de-identified electronic medical records. From data collected in this patient population, we developed a longitudinal pharmacogenetic study to test the impact of ADME Core variants on the development of CNIT [9].

## 2. Methods

### 2.1. Study Population

As stated above, our study population of heart transplant recipients was obtained from BioVU. A full description of BioVU as a resource, including its ethical, privacy and other protections has been described in detail elsewhere [10]. In brief, BioVU extracts and stores DNA from blood collected from routine clinical testing that is scheduled to be discarded after a three-day waiting period at the Vanderbilt University Medical Center (VUMC) in Nashville, TN. DNA samples are linked to a de-identified version of the patient's electronic medical record, known as the Synthetic Derivative (SD), which can be accessed by investigators for research purposes after approval by the local internal review board and BioVU Review Committee. Patients eligible for possible inclusion into BioVU are those with an out-patient laboratory

blood draw, have signed the consent to treatment form, and have not made a formal indication to opt-out [11].

Using the SD, we identified initial candidates for our study by screening for patients who met the following criteria: a) a heart transplant documented with three or more ICD9 code V42.1 (heart replaced by transplant) and/or one CPT code 33945, b) one or more mention of an immunosuppressant, c) DNA available in the biorepository and genotyped on the Illumina ADME Core Panel, and d) the patient was over the age of 15 at the date of the transplant operation. This initial screen identified 152 potential candidates. We then manually extracted the date of the transplant operation from each record. We excluded 10 patients with an ambiguous transplant operation date in the record or miscoded with a kidney, lung, liver, or multiple heart transplants during his/her lifespan. We extracted immunosuppressant data from the de-identified records of this heart transplant sample population with MedEx. MedEx extracts medications and their signature mentions from free-text entries in the EMRs. We used only medications with at least one mention of a dose, route, frequency or strength to limit the medications to those the patient was actually prescribed. A more detailed description of the software has been published elsewhere [12].

We also extracted additional clinical information from the SD. For quantitative measurements such as body mass index (BMI,  $\text{kg/m}^2$ ), serum creatinine (mg/dl), and systolic and diastolic blood pressure (mmHg), monthly medians were calculated. Prior to transplant chronic kidney disease and diabetes mellitus were defined by ICD9 codes before the transplant date. Chronic kidney disease was defined by three or more mentions of the following ICD9-codes: 403, 585.1, 585.2, 585.3, 585.4, 585.5, 585.6, and 585.9. Patients were considered to have diabetes mellitus pre-transplant if they had three or more mentions of the following ICD-9 codes: 250.3, 250.32, 250.2, 250.22, 250.9, 250.92, 250.8, 250.82, 250.7, 250.72, 250.6, 250.62, 250.5, 250.52, 250.4, 250.42, 250, and 250.02. Pre-transplant hypertension was defined as median systolic blood pressure  $> 140$  mmHg, systolic and /or  $> 90$  mmHg diastolic, or prescribed one of the following hypertension medications: hydralazine, minoxidil, renin antagonist, central alpha agonists, ACE inhibitors (ACEI)/angiotensin receptor blockers (ARB), aldosterone antagonists, diuretics, K-sparing diuretics, loop diuretics, alpha antagonists, calcium channel blockers (CCB), beta blockers (BB), thiazide/BB, thiazide/ACEI/ARB, thiazide/aldosterone antagonist, thiazide/renin antagonist, and diuretic combinations, all before the transplant date.

## 2.2. Phenotype Definition

The outcome of interest was time to develop severe nephrotoxicity clinically attributed to CNIT, which we defined in our patient population as the development of CKD stage 4 or 5 in the setting of CI use. To assess kidney function over the course of immunosuppression therapy, we estimated the glomerular filtration rate from the “four variable” Modification of Diet in Renal Disease formula [13]:

$$186 \times \text{Serum Creatinine}^{-1.154} \times \text{Age}^{-0.203} \times [1.212 \text{ if Black}] \times [0.742 \text{ if Female}] \quad (1)$$

All patients who entered into the SD by the time of their transplant date were included in this study. Patients who entered the SD post-transplant were included if their initial eGFR measurement upon entering the SD was  $> 30 \text{ mL/min/1.73m}^2$ , this included patients with CKD stages 1, 2, and 3. These patients were assumed not to have CKD 4 or 5 in the setting of CI prior to their entry into BioVU and were entered into the analysis at their heart transplant date. Patients who entered the SD after their heart transplant date with



an eGFR < 30 were excluded from the analysis. Our definition of severe chronic kidney disease 4 was a monthly median eGFR of < 30 mL/min/1.73m<sup>2</sup> for three consecutive months. This threshold is adapted from the National Kidney Foundation's definition for CKD stage 4: GFR of 15-29 and CKD Stage 5: GFR <15 or dialysis [14].

### **2.3. Genotyping**

DNA samples from a total of 115 heart transplant recipients were genotyped on Illumina's ADME Core Panel as part of Vanderbilt Electronic Systems for Pharmacogenomic Assessment (VESPA). In short, Illumina's pharmacogenetic-targeted ADME Core panel is designed for the genotyping of 184 markers in 34 genes. A full description of the panel's content and performance has been published elsewhere [9]. Genotyping for this study was conducted at the Center for Human Genetics Research DNA Resources Core at Vanderbilt University. Genotype calling was performed with ADME Module Version 1.0.0.3. Formatting of the ADME Core Panel data set and quality control of the markers was performed with PLATO and PLINK [15;16]. SNPs were filtered from the analysis if the allele frequency was below 5%, genotyping efficiency <95%, or a statistically significant deviation from Hardy Weinberg expectations ( $p < 0.001$ ) in the European American population. After filtering, 49 SNPs remained in our analysis. A principal components analysis (PCA) was performed with the Eigensoft software using available genome-wide data in the full dataset and in the subset of European Americans [17]. We tested for relatedness of individuals in subsets of samples stratified by race/ethnicity. One sample from a related pair of European Americans was removed. The genome-wide inflation factor for this study was 1. We extracted 333,804 overlapping markers from the samples' genotype data from the following platforms: 18 individuals on Illumina's HumanOmni5-Quad, 109 on the HumanOmni1-Quad, and four on Illumina's 1M-Duo BeadChip.

### **2.4. Statistical Analysis**

Cox proportional hazard models were calculated using the date of the heart transplant as the starting time in a time-to-event analysis. Genotypes were modeled additively against development of CKD stage 4. Factors that were associated with renal function in univariate analyses ( $p < 0.05$ ) were included in the final multivariable model. Patients who did not develop CKD stage 4 were censored from the analysis at their final eGFR measurement. For the linear mixed effects modeling of post-transplant eGFR, we used the R package, nlme. SNPs and covariates that met a 0.05 threshold in univariate analyses were included as fixed effects and the subject identifier was included as a random effect. The within subject correlation was 0.70 and we chose to account for it in our models with an autoregressive-moving average model with one autoregressive and one moving average parameters. Plots were generated with STATA 11 and RStudio Version 0.97.551

### 3. Results

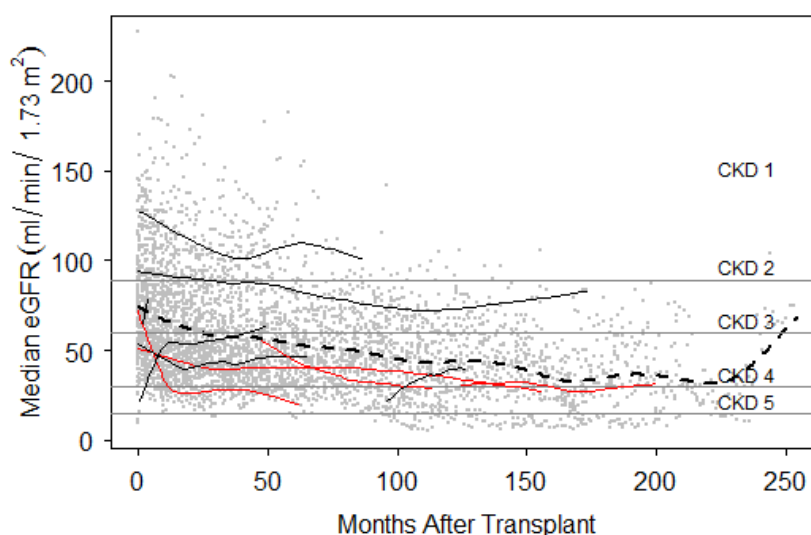
#### 3.1. Demographics

**Table 1:** Clinical Characteristics of heart transplant samples.

Patients	115
European Descent (%)	86.0
Female (%)	33.9
Transplant Operation at VUMC (%)	80.8
Pre-transplant Diabetes Mellitus (%)	10.4
Median Systolic (mmHg)	100.2, IQR: 94.3-107.0
Median Diastolic (mmHg)	64.0 IQR: 59.9-66.9
Pre-transplant Hypertension (%)	66.0
Pre-transplant Chronic Kidney Disease	9.56
Median Age at Tx (years)	52.5, IQR: 40.5-58.1
Required Dialysis Post-Transplant (%)	18.2
Median Post Tx Follow up Time (years)	8.8, IQR: 4.8 – 12.2
Median Pre-eGFR (mL/min/1.73m <sup>2</sup> )	68.0, IQR: 57.4-87.2
Median Body Mass Index (kg/m <sup>2</sup> )	27.4, IQR:24.6-31.1
Died (%)	21.7
Cyclosporine Only (%)	35.7
Tacrolimus Only (%)	25.2
Cyclosporine and Tacrolimus (%)	39.1

Table 1 presents the clinical characteristics of our study population identified in BioVU. Overall, this is an ancestrally cosmopolitan cohort where 80.8% of the patients were administratively assigned [18] as of European descent, while the remainder was reported as African American with the exception of one

sample reported as Hispanic. The median age at transplant was 52.5 years of age. This is a slightly overweight population with the median body mass index of  $27.4 \text{ kg/m}^2$ . Prior to transplant, 10.4% and 60.6% patients had evidence of diabetes mellitus and hypertension, respectively. A majority of patients (52.7%) had their heart transplant at VUMC. Twenty-five patients died during post-transplant follow up. All patients were prescribed a calcineurin-inhibitor: 35.7% were prescribed cyclosporine alone, 25.2% tacrolimus alone, and 39.1% were prescribed a combination of the two (at different times).



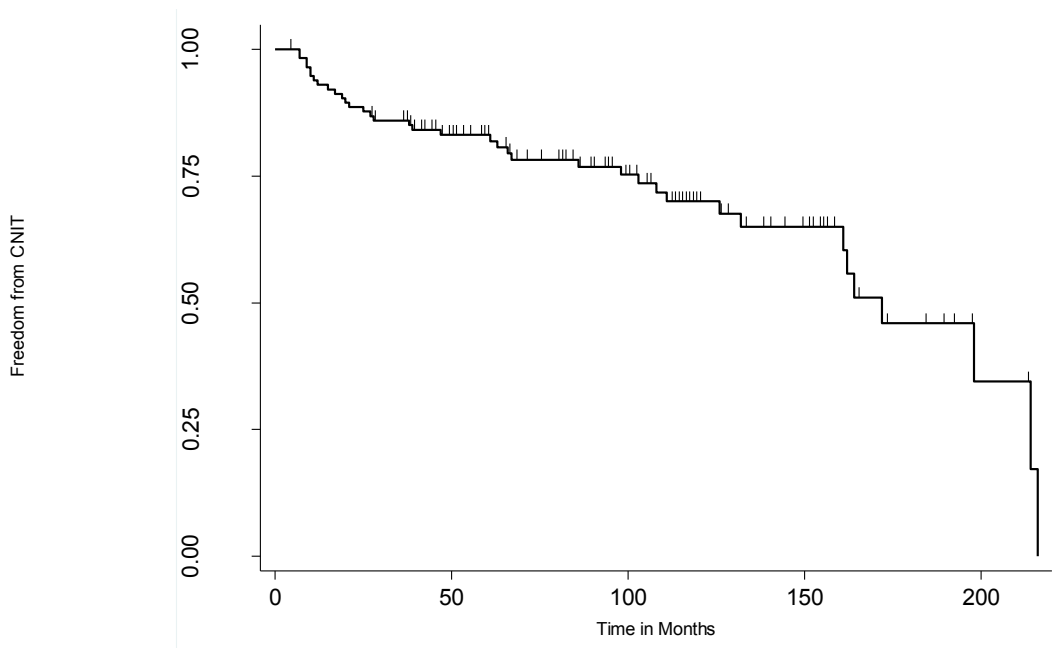
**Figure 1. Post-transplant eGFR measurements plotted on the thresholds of the five stages of chronic kidney disease.** Individual post-transplant eGFR measurements are plotted on the y-axis against time in months after transplant on x-axis as grey dots. The dashed line represents a polynomial function fit to all eGFR measurements collected in the study. Ten randomly selected patient's eGFR profiles have fitted with loess lines and colored in red if the patient developed Chronic Kidney Disease (CKD) Stage 4 or below. Thresholds for the 5 stages of CKD are indicated: CKD1 >90, CKD2 60-89, CKD3 30-59, CKD4 15-29, and CKD5 <15 mL/min/1.73m<sup>2</sup>.

As expected for this patient population, the eGFR prior to transplant was lower than would be expected for a healthy population (median =  $68.0 \text{ mL/min/1.73m}^2$ ). Follow up time for these patients varied (Figure 1): median time to the final eGFR measurement in the SD was 8.8 years, and the median frequency of follow-up was 5.5 (IQR: 4.2-7.5) eGFR measurements per year. Kidney function continued to decline over time (Figure 1). In the second year (12-24 months) post-transplant 14.0, 31.4, 50.0, and 4.6 percent of individuals had median eGFR measurements that corresponded with the first four stages of CKD, respectively. By the fifth year (60-72 months), the distribution shifted towards lower median eGFR levels: 3.4, 22.4, 62.0, and 12.0 percent of individuals were observed with median eGFRs in range with the first

four stages of CKD. At year ten, 11.7 and 11.7 percent of patients median eGFR measurements corresponded to CKD stages four and five, respectively.

### 3.2 Time to CKD Stage 4 and 5 Survival Analysis

Figure 2 displays the development of CNIT in this study population in months post-heart transplant. Thirty-seven out of 115 patients (25.2) in this heart transplant cohort met the CNIT case definition. By twelve months, eight individuals (7.0%) met the criteria for CNIT, 19 (16.5%) by 60 months, and 28 (24.3%) by 120 months. From among the various clinical variables tested for an association with CNIT (the three most significant PCAs, gender, systolic and diastolic blood pressure, pre-transplant diabetes, pre-transplant hypertension, pre-transplant chronic kidney disease, age at transplant, pre-transplant eGFR, BMI, and prescribed calcineurin inhibitor), only pre-transplant eGFR, pre-transplant CKD status, pre-transplant diabetes mellitus status, and age at transplant met a significance threshold of  $p < 0.05$  (Table 2).



**Figure 2. Kaplan-Meier plot describing the proportion of non-nephrotoxic heart transplant recipients over time.** The y-axis indicates the proportion of event-free subjects and tick marks on the plot indicate where individuals are censored from the analysis.

**Table 2: Results of CNIT Analysis in European Americans**

Predictor	Hazard Ratio (95% CI)	P-value
<u>Univariate Clinical Variable Model</u>		
Recipient Age per year	1.05 (1.01-1.08)	$9.85 \times 10^{-3}$
Pre-transplant CKD	3.69 (1.36-10.01)	0.01
Pre-Transplant eGFR per ml/min/1.73m <sup>2</sup>	0.96 (0.94-0.98)	$1.03 \times 10^{-3}$
Prior Diabetes Mellitus	6.92(2.64-18.54)	$8.33 \times 10^{-5}$
<u>Multivariable Genetic Model</u>		
<i>DPYD</i> rs1801265	0.45 (0.22-0.93)	0.03
<i>UGT2B17</i> rs1902023	2.23 (1.21-4.11)	0.01
<i>SLCO1B1</i> rs4149056	0.38(0.14-0.96)	0.03
<i>SLC22A1</i> rs34305973	2.14(1.18-3.90)	0.01

First, in the European American subset (n=99 heart transplant recipients with 35 cases of CKD stages 4 and 5) we tested the 49 Illumina ADME Core Panel markers that passed quality control for association with CNIT outcome. In unadjusted analysis, no markers were associated with CNIT after adjustment for multiple testing ( $p < 1.02 \times 10^{-3}$ ). Variants in *SLC22A1* rs34305973 and *UGT2B17* rs1902023 trended toward significance in the unadjusted model ( $p = 0.02$  and  $p=0.02$ , respectively). In models adjusted for pre-transplant CKD, pre-transplant diabetes mellitus, age at transplant, and the three most significant PCAs, *UGT2B17* rs1902023 was the most significant ( $p = 0.01$ ) among all the tested ADME Core Panel markers (Table 2). Secondly, we expanded our analysis to the full dataset regardless of race/ethnicity (n=115 heart transplant recipients with 37 cases of CKD stage 4 and 5) and the results were largely unchanged (data not shown). In the adjusted models for the full dataset, *DPYD* rs1801265 was the most significant ( $p = 9.24 \times 10^{-3}$ , HR: 0.39, CI: 0.19-0.79) among all the tested ADME Core Panel markers. No marker was associated with CNIT in unadjusted or adjusted models after correction for multiple testing when the data were limited to cyclosporine treated only patients (n=95 heart transplant recipients with 27 cases of CKD stage 4 or 5) or tacrolimus treated only patients (n=79 heart transplant recipients with 18 cases of CKD stage 4 or 5; data not shown).

### 3.2 Modeling Post-Transplant eGFR

As a secondary analysis of post-transplant kidney function, the repeated eGFR measurements were analyzed directly using mixed effects models to account for the within subject correlation. In univariate analyses of covariates among European Americans, only cyclosporine use (coef(S.E) = -17.05(7.13),  $p = 0.02$ ), median BMI (coef(S.E) = -1.27(0.62),  $p < 0.05$ ), and age at transplant (coef(S.E) = -1.01(0.15),  $p = 1.55 \times 10^{-8}$ ) were associated with eGFR over time. No SNP met the significance threshold for multiple testing in unadjusted or adjusted analyses. However, in unadjusted analyses, two of the three SNPs that met a threshold of 0.05 have previously been associated with post-transplant renal function: *CYP2C19* rs4244285 (coef(S.E) = 13.28(6.17),  $p = 0.03$ ) and *CYP3A5* rs776746 (coef(S.E) = 21.94(8.37),  $p = 0.01$ ). SNP *CYP2A6* rs28399433 also met the 0.05 threshold (coef(S.E) = 20.91(3.46),  $p = 0.02$ ) in unadjusted analyses. Two of these associations maintained significance at the 0.05 threshold in the multivariate models *CYP3A5* rs776746 (coef(S.E) = 14.60(6.41),  $p = 0.03$ ) and *CYP2A6* rs28399433 (coef(S.E) = 17.14(8.24),  $p = 0.04$ ) [19]. In analyses extended to the full dataset regardless of race/ethnicity, only *CYP2A6* rs28399433 (coef(S.E) = 17.46(6.70),  $p = 0.01$ ) approached significance in the adjusted analysis (data not shown).

## 4. Discussion

### 4.1. Summary and Relevance

We used a biorepository linked to de-identified electronic medical records to identify heart transplant patients for pharmacogenomic studies. The two outcomes of interest in the present pharmacogenomics study was the development of advanced nephropathy (CKD Stage 4 or 5) in the setting of calcineurin-inhibitor therapy post-transplant and post-transplant eGFR over time. In this study, we have demonstrated that EMR-based cohorts linked to DNA samples provide ample opportunity to identify adverse drug reactions (ADR). This specific study focused on a common ADR to calcineurin-inhibitor therapy among heart transplant recipients. While there are several studies that have explored the relationship between a patient's genetic profile and calcineurin-inhibitor dosing [5;20;21], this is the first study of our knowledge utilizing an EMR-based cohort of heart transplant patients to examine the pharmacogenetics of calcineurin-induced nephrotoxicity.

Our most significant result in the time to CNIT survival analysis was *DPYD* rs1801265, which approached our corrected p-value ( $p = 9.24 \times 10^{-3}$ ) in the full dataset regardless of race/ethnicity. *DPYD* rs1801265 defines the *DPYD* \*9A haplotype and encodes a cysteine to arginine missense mutation in the 29<sup>th</sup> position of the protein that some studies have suggested to be without significant enzymatic activity [22]. The gene is located in the centromeric region of chromosome one between 1p22 and 1q21 [23]. While the variant did not meet our multiple-testing threshold, larger studies may confirm its role in CNIT. It is interesting to note that *CYP3A5* variants, which have been strongly associated with tacrolimus dosing in multiple studies [5], were not associated with CNIT but one marker in this gene trended towards significance in modeling eGFR directly. This marker rs776746 defines the *CYP3A5*\*3 allele, a non-expressing variant of the gene found a high frequency in populations of European descent [24]. In this study we found the functional *CYP3A5*\*1 allele, which we found at comparable frequency to other studies (MAF = 0.06), to be positively associated with eGFR post-transplant [25].

The application of a heart transplant cohort for the pharmacogenetics of calcineurin-inhibitor nephrotoxicity has advantages over kidney and liver transplantations, as it eliminates the potential for donor-recipient gene interactions. The donor genetic information of kidney and liver transplant may play crucial roles in the susceptibility of nephrotoxicity. The liver is the primary site of drug metabolism, and in the case of liver transplants, the donor's genome becomes the driver of metabolism. Its own unique genetic variation may lead to a different pharmacokinetic profile of calcineurin-inhibitor metabolism compared with the recipient. The donor genome in the case of kidney transplant may also be a factor in developing nephrotoxicity [26]. Therefore studies designed at identifying these interactions are presented with experimental design challenges unlikely to be overcome in blood sample focused biorepository [27].

## **4.2. Limitations**

Small sample size is a pervading challenge to pharmacogenetic study design. Even in an immense resource such as BioVU with over 160,000 samples as of July 2013, we were only able to identify 167 patients who met the study criteria, and of those, only 35 of those samples developed CKD stage 4 over the course of calcineurin drug therapy. This finding highlights the need for very large repositories when studying uncommon outcomes of medical interest. While survival analysis did afford us more power opposed to treating the data as strict case-control and performing logistic regression, we were still underpowered to detect an association. For example, assuming a dominant genetic model with an allele with a frequency of 0.5, a sample size of 191 cases of CKD stage 4 would be required to detect an association with a moderately sized hazard ratio of 1.5 at an alpha of 0.05 [28].

Heterogeneity marked another challenge when defining this study population and modeling the association. Clinically, heart transplant recipients are a very diverse population in regards to co-morbidities and medications. Further complicating the issue is that CNIT is not the only cause of CKD in this population: other factors include the decline of kidney function with age, diabetes, hypertension, heart disease, other medication exposures, and latent infection of the BK virus [29]. In this study, we ignored phenotypic heterogeneity to increase the sample size and overall power of the study. Also, to avoid increasing the type II error rate, we were parsimonious in our covariate selection for our statistical model to maximize statistical power [30]. Indeed, large multi-center studies may be required to fully model the relationship between heart disease and kidney function. Large studies will also be required to fully address the phenotype heterogeneity problem or to explore more susceptible subpopulations such as high dose patients, a strategy successfully used to identify genetic variants associated with statin-induced ADRs [31].

## **4.3. Conclusions**

Despite the relatively small sample size for a genetic association study, the current study represents a fairly large sample size for pharmacogenomics studies of ADRs. We have demonstrated here that the EMR, rich in clinical data, is an excellent and logical resource to establish pharmacogenomics studies for

less common ADRs such as CNIT. While the genetic association results presented here require replication and downstream functional and biological interpretation, the existence of other biobanks linked to DNA samples in the United States [32] and across the world [33] makes this future direction possible for CNIT as well as other ADRs with a suspected genetic risk factor.

## 5. References

1. Radovancevic B, Konuralp C, Vrtovec B et al., *J. Heart Lung Transplant.* **24**, 156 (2005).
2. Naesens M, Kuypers DR, Sarwal M, *Clin. J. Am. Soc. Nephrol.* **4**, 481 (2009).
3. Hamour IM, Omar F, Lyster HS, Palmer A, Banner NR, *Nephrol. Dial. Transplant.* **24**, 1655 (2009).
4. Murray B, Hawes E, Lee RA, Watson R, Roederer MW, *Pharmacogenomics.* **14**, 783 (2013).
5. Birdwell KA, Grady B, Choi L et al., *Pharmacogenet. Genomics.* **22**, 32 (2012).
6. Haufroid V, Mourad M, Van K, V et al., *Pharmacogenetics.* **14**, 147 (2004).
7. Hesselink DA, van GT, van Schaik RH, *Pharmacogenomics.* **6**, 323 (2005).
8. Kuypers DR, Naesens M, de JH, Lerut E, Verbeke K, Vanrenterghem Y, *Ther. Drug Monit.* **32**, 394 (2010).
9. Oetjens MT, Denny JC, Ritchie MD et al., *Pharmacogenomics.* **14**, 735 (2013).
10. Roden DM, Pulley JM, Basford MA et al., *Clin. Pharmacol. Ther.* **84**, 362 (2008).
11. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR, *Clin. Transl. Sci.* **3**, 42 (2010).
12. Xu H, Jiang M, Oetjens M et al., *J. Am. Med. Inform. Assoc.* **18**, 387 (2011).
13. Poggio ED, Wang X, Greene T, Van LF, Hall PM, *J. Am. Soc. Nephrol.* **16**, 459 (2005).
14. Abboud H, Henrich WL, *N. Engl. J. Med.* **362**, 56 (2010).
15. Purcell S, Neale B, Todd-Brown K et al., *Am. J. Hum. Genet.* **81**, 559 (2007).
16. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD, *Pac. Symp. Biocomput.*, 315 (2010).
17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D, *Nat. Genet.* **38**, 904 (2006).
18. Dumitrescu L, Ritchie MD, Brown-Gentry K et al., *Genet. Med.* **12**, 648 (2010).
19. Boso V, Herrero MJ, Bea S et al., *Drug Metab Dispos.* **41**, 480 (2013).
20. Ware N, MacPhee IA, *Curr. Opin. Mol. Ther.* **12**, 270 (2010).
21. Ferrarresso M, Tirelli A, Ghio L et al., *Pediatr. Transplant.* **11**, 296 (2007).
22. Vreken P, Van Kuilenburg AB, Meinsma R, Van Gennip AH, *Hum. Genet.* **101**, 333 (1997).
23. Takai S, Fernandez-Salguero P, Kimura S, Gonzalez FJ, Yamada K, *Genomics.* **24**, 613 (1994).
24. Lamba J, Hebert JM, Schuetz EG, Klein TE, Altman RB, *Pharmacogenet. Genomics.* **22**, 555 (2012).
25. de DS, Zakrzewski M, Barhdadi A et al., *J. Heart Lung Transplant.* **30**, 326 (2011).
26. Hauser IA, Schaeffeler E, Gauer S et al., *J. Am. Soc. Nephrol.* **16**, 1501 (2005).
27. Ritchie MD, Denny JC, Crawford DC et al., *Am. J. Hum. Genet.* **86**, 560 (2010).
28. Schoenfeld DA, *Biometrics.* **39**, 499 (1983).
29. Bloom RD, Reese PP, *J. Am. Soc. Nephrol.* **18**, 3031 (2007).
30. Pirinen M, Donnelly P, Spencer CC, *Nat. Genet.* **44**, 848 (2012).
31. Link E, Parish S, Armitage J et al., *N. Engl. J. Med.* **359**, 789 (2008).
32. McCarty CA, Chisholm RL, Chute CG et al., *BMC. Med. Genomics.* **4**, 13 (2011).
33. Harris JR, Burton P, Knoppers BM et al., *Eur. J. Hum. Genet.* **20**, 1105 (2012).



# ROBUST REVERSE ENGINEERING OF DYNAMIC GENE NETWORKS UNDER SAMPLE SIZE HETEROGENEITY

ANKUR P. PARIKH, WEI WU, ERIC P. XING

*School of Computer Science, Carnegie Mellon University,  
Pittsburgh, PA 15213, USA*

*\*E-mail: apparikh@cs.cmu.edu, weiwu2@cs.cmu.edu, epxing@cs.cmu.edu*

Simultaneously reverse engineering a collection of condition-specific gene networks from gene expression microarray data to uncover dynamic mechanisms is a key challenge in systems biology. However, existing methods for this task are very sensitive to variations in the size of the microarray samples across different biological conditions (which we term *sample size heterogeneity in network reconstruction*), and can potentially produce misleading results that can lead to incorrect biological interpretation. In this work, we develop a more robust framework that addresses this novel problem. Just like microarray measurements across conditions must undergo proper normalization on their magnitudes before entering subsequent analysis, we argue that networks across conditions also need to be "normalized" on their density when they are constructed, and we provide an algorithm that allows such normalization to be facilitated while estimating the networks. We show the quantitative advantages of our approach on synthetic and real data. Our analysis of a hematopoietic stem cell dataset reveals interesting results, some of which are confirmed by previously validated results.

*Keywords:* gene network reconstruction, dynamic, sample size heterogeneity

## 1. Introduction

Capturing and understanding the differential usage (i.e. rewiring) of cellular pathways and regulatory structures as a result of various biological processes and responses to external stimuli is an important problem in systems biology. Some examples include embryonic development, cell cycle, differentiation, and carcinogenesis. One promising technique to help uncover complex gene interactions governing these processes is to use computational methods to reverse engineer gene networks from microarray data. The macro-topology of the recovered network as well as the individual interactions among the genes can then be analyzed to shed more light into the underlying regulatory mechanisms.

To model the evolving nature of these phenomena, it often does not suffice to reconstruct one static snapshot of the underlying regulatory structure since this cannot uncover dynamic functional roles played by various genes in different cellular stages or at different times. Consider an example of the human hematopoietic system shown in Figure 1. Hematopoietic stem cells (located at the root) differentiate into more specialized cells along the lineages, eventually becoming red blood cells, platelets, or white blood cells. It would be inappropriate to pool together various samples to reconstruct a single network representing a common regulatory structure for different cell states, e.g., red and white blood cells, since they have distinct morphologies and play completely different roles in biological systems, and thus their respective regulatory structures must also be considerably different. Instead it is more suitable to reconstruct a *collection* of networks, one for each cell state. Different functional roles of various genes across the different cell states can then be analyzed.

However, the problem of simultaneously recovering a collection of networks over different cell states poses unique challenges that do not appear in the static recovery case. The key challenge we face in this work is that different cell states have different numbers of microarray samples, which we term *sample size heterogeneity in network reconstruction*. This phenomenon is quite common in biological datasets due to a variety of reasons such as samples having to be discarded if the quality of the microarrays is poor, or constraints on acquisition of certain biomedical samples.

Even though sample size heterogeneity can pose considerable challenges for many existing network reconstruction methods in different ways, in this work we choose to focus on addressing its effect on a class of state-of-the-art methods that are based on sparse, regularized regression.<sup>1-3</sup> These methods are designed for the high dimensional setting common in biology, where the number of genes can be substantially larger than the number of samples, and allow us to uncover more sophisticated dependencies than can be obtained by measuring simpler quantities such as correlation or mutual information. Building upon the regularized regression based network learning paradigm, several methods<sup>4-6</sup> have recently proposed leveraging similarities of multiple networks corresponding to biological conditions considered to be related for more accurate multi-network joint estimation, under evolving network scenarios. This strategy is very valuable in the scenario we consider in this work, where the number of samples for each cell state is small (e.g., as few as 4 per cell state, clearly statistically insignificant for inferring a network alone), and thus information sharing between related cell states is crucial and can increase the effective sample size and consequently the power of network learning. Such methods have helped reveal the dynamic interactions in embryonic development<sup>4</sup> as well as cancer progression and reversion.<sup>6</sup>

Despite being statistically powerful, network learning approaches based on regularized regression can suffer from sample size heterogeneity, which can substantially bias the density of the networks recovered. In particular, with existing sparse regression methods, cell states with more samples will tend to have considerably denser networks than those with fewer samples, a phenomena depicted in Figure 1. Intuitively, this is because the algorithm is more confident about estimating networks with more samples and thus these networks are denser.

The resultant artificial difference may be acceptable in certain applications (e.g. features for a downstream classifier). However, in many cases, we are interested in a comparative analysis of the networks, both in terms of macro-topology (e.g. density, centrality) or micro-topology (e.g. neighborhoods of individual genes). In this scenario, sample size heterogeneity can lead to misleading biological conclusions, since it will be unclear which differences among the networks are manifestations of the actual changes in regulatory mechanisms across different cell states and which are the artifacts due to sample size heterogeneity.

One simple approach to handle sample size heterogeneity is to make each cell state have the same number of samples by discarding excess samples in some states. The downside of this approach is the waste of the precious data in the small-sample-size scenarios common in biological studies. For example, in the hematopoietic stem cell dataset we consider, using this strategy would lead to a reduction of the total sample size by approximately 40 percent.

Another approach is to post-process the networks to be more calibrated, e.g. normalizing

all the edge weights across the cell states and then applying some threshold. However, this may produce adverse effects. Namely, since edges can only be *deleted*, and not *added* during post-processing, the original networks learned using sparse regression have to be denser than desired, and then further sparsified via post-processing. The resulting edge set from this procedure would then be suboptimal compared to the edge set constructed by just learning a sparser network with the regularized regression.

### 1.1. Our Contribution

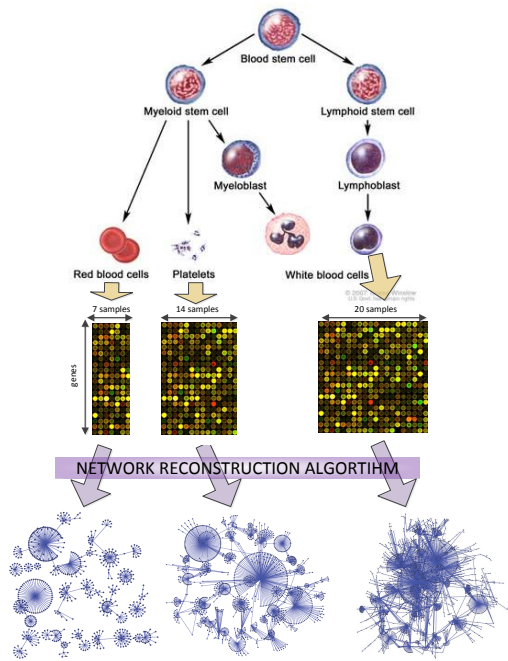


Fig. 1. Illustration of a hematopoietic stem cell genealogy and how more samples bias existing reconstruction methods to give artificially denser networks. <sup>a</sup>

estimator, and therefore more effective and statistically justifiable.

The rest of the work is outlined as follows. We first present the general framework of reconstructing gene networks via sparse regression methods and concretely illustrate the problem that sample size heterogeneity poses. We then present our robust method. Lastly, we evaluate our approach on synthetic data as well as on a human hematopoietic stem cell dataset.

## 2. Background: Recovering Gene Networks via Gaussian Graphical Models

Consider the problem of modeling a set of gene regulatory networks, denoted by  $\mathcal{Z}$  (where  $|\mathcal{Z}| = Z$ ), each corresponding to a different cell state  $z \in \mathcal{Z}$  with  $S_z$  *i.i.d.* microarray measurements of all genes in cell state  $z$ .  $\mathcal{Z}$  could represent a set of networks over time or over a genealogy. Let  $\mathcal{G}^{(z)} = (\mathcal{V}, \mathcal{E}^{(z)})$  represent a network in cell state  $z$ , where  $\mathcal{V}$  denotes the set of  $p$  genes (fixed for all  $z$ ), and  $\mathcal{E}^{(z)}$  denotes the set of edges over vertices. An edge  $(u, v) \in \mathcal{E}^{(z)}$  can

<sup>a</sup><http://www.siteman.wustl.edu/CancerDetails.aspx?id=661&xml=CDR257990.xml>

represent a relationship (e.g., influence or interaction) between genes  $u$  and  $v$  in cell state  $z$ . Let  $\mathbf{X}^{(s,z)} = (X_1^{(s,z)}, \dots, X_p^{(s,z)})'$ , where  $s \in \{1, \dots, S_z\}$ , be a vector of gene expression values that are real valued and standardized, such that each dimension has mean 0 and variance 1.

A gene network can be represented by a probabilistic graphical model.<sup>7,8</sup> While there are many other ways to represent gene networks, the advantage of using graphical models is that the graph structure encodes conditional independence relations among the genes, and is thus able to model more nuanced relationships than simple statistical quantities such as correlation or mutual information. In this work, we assume that  $\mathbf{X}^{(z)}$  follows a multivariate Gaussian distribution with mean 0 and covariance matrix  $\Sigma^{(z)}$ , so that the conditional independence relationships among the genes can be encoded as a Gaussian graphical model (GGM).<sup>9</sup> It is well known that for GGMs, edges in the graph correspond to non-zero elements in the inverse covariance matrix (known as the precision matrix), which we denote by  $\mathbf{\Omega}^{(z)} := (\omega_{uv}^{(z)})_{u,v \in [p]}$ . Thus, estimating the graph structure is equivalent to selecting the non-zero elements of the precision matrix.

As commonly done, instead of directly estimating the precision matrix elements  $\omega_{uv}^{(z)}$ , we estimate the partial correlation coefficients  $\rho^{(z)}$ , which are proportional to the precision matrix elements:  $\rho_{uv}^{(z)} = -\frac{\omega_{uv}^{(z)}}{\sqrt{\omega_{uu}^{(z)}\omega_{vv}^{(z)}}}$ . Thus,  $\rho_{uv}^{(z)}$  is zero if and only if  $\omega_{uv}^{(z)}$  is zero. Thus the network resultant from the non-zero  $\rho_{uv}^{(z)}$  is equivalent to that from the nonzero  $\omega_{uv}^{(z)}$ . Furthermore, the partial correlation is intuitive in the sense that a high positive value of  $\rho_{uv}^{(z)}$  indicates that the genes  $u$  and  $v$  are strongly positively correlated (conditioned on the other genes), while a low negative value indicates the genes are strongly negatively correlated (conditioned on the other genes), and  $\rho_{uv}^{(z)} = 0$  for all  $(u, v) \notin \mathcal{E}^{(z)}$ . As a result, we simply consider estimating the partial correlation coefficients and designate these as the edge values in  $\mathcal{G}^{(z)}$ :  $\mathcal{E}^{(z)} = \{\rho_{uv}^{(z)} : |\rho_{uv}^{(z)}| > 0\}$ .

## 2.1. Neighborhood Selection

Estimating  $\rho_{uv}$  is challenging because biological data is often high dimensional (tens of thousands of genes) while the number of samples is small (in the tens). One approach is neighborhood selection<sup>2</sup> based on  $\ell_1$ -norm regularized regression, which has strong theoretical guarantees and also works well in practice. We first discuss it in the context of estimating a collection of networks independently, which is also the foundation of existing approaches on time-varying network estimation that leverage information among similar states.<sup>4-6</sup>

Here the neighborhood of each gene  $u$  is estimated independently and the neighborhoods are then combined to form a network. In every neighbor estimation step, gene  $u$  is treated as a response variable, all the other genes are the covariates, and the regression weights are proportional to the partial correlation coefficients between the other genes and  $u$ . More formally, let  $\mathbf{X}_{\setminus u}$  indicate the  $p - 1$  vector of the values of all genes except  $u$ . Similarly,  $\beta_{\setminus u} := \{\beta_{uv} : v \in \mathcal{V} \setminus u\}$ . It is a well known result, that the partial correlation coefficients can be related to the following regression model<sup>10</sup>:  $X_u^{(z)} = \sum_{v \neq u} X_v^{(z)} \beta_{uv}^{(z)} + \epsilon_u^{(z)}$ ,  $u \in [p]$ , where  $\epsilon_u^{(z)}$  is uncorrelated with  $\mathbf{X}_{\setminus u}^{(z)}$  if and only if  $\beta_{uv}^{(z)} = -\frac{\omega_{uv}^{(z)}}{\omega_{uu}^{(z)}} = \rho_{uv}^{(z)} \sqrt{\frac{\omega_{vv}^{(z)}}{\omega_{uu}^{(z)}}}$ . Some algebra gives that  $\rho_{uv}^{(z)} = \text{sign}(\beta_{uv}^{(z)}) \sqrt{\beta_{uv}^{(z)} \beta_{vu}^{(z)}}$ . The above equations basically indicate that we can solve for the regression coefficients using a linear regression, where the response variable corresponds to

$X_u$  and the covariates correspond to  $\mathbf{X}_{\setminus u}$ . The corresponding partial correlation coefficients can be recovered via the algebraic relations. An  $\ell_1$  penalty is applied to encourage a sparse solution, as in the lasso.<sup>1</sup> We can estimate the neighborhood of gene  $u$  for all cell states  $z \in \mathcal{Z}$  using this strategy, as depicted in Eq. 1.

$$\hat{\beta}_{\setminus u}^{(1)}, \dots, \hat{\beta}_{\setminus u}^{(Z)} = \underset{\beta_{\setminus u}^{(1)}, \dots, \beta_{\setminus u}^{(Z)}}{\operatorname{argmin}} \sum_{z \in \mathcal{Z}} \mathcal{L}^u(\mathbf{X}^{(z)}, \beta_{\setminus u}^{(z)}) + \lambda \sum_{z \in \mathcal{Z}} \|\beta_{\setminus u}^{(z)}\|_1 \quad (1)$$

where  $\mathcal{L}^u(\mathbf{X}^{(z)}, \beta_{\setminus u}^{(z)}) := \sum_{s=1}^{S_z} \left( x_u^{(s,z)} - \sum_{v \neq u} \beta_{uv}^{(z)} x_v^{(s,z)} \right)^2$ . Note that the optimization problem decouples into  $Z$  separate problems. This procedure is repeated to estimate the neighborhood of every gene  $u \in \mathcal{V}$ . It has been shown that under certain conditions, one can obtain an estimator of the edge set  $\mathcal{E}$  that is *sparsistent*,<sup>2,11</sup> i.e. the correct network structure can be attained as a function of the number of genes, samples, and topology of the network.

## 2.2. Neighborhood Selection and Sample Size Heterogeneity

However, applying the same  $\lambda$  to all  $z \in \mathcal{Z}$  such as in Eq. 1 can be problematic under sample size heterogeneity. Consider two cell states  $z_1$  and  $z_2$  and assume that  $S_{z_1} > S_{z_2}$ . This implies that  $\mathcal{L}^u(\mathbf{X}^{(z_1)}, \beta_{\setminus u}^{(z_1)} = \mathbf{0})$  will generally be larger than  $\mathcal{L}^u(\mathbf{X}^{(z_2)}, \beta_{\setminus u}^{(z_2)} = \mathbf{0})$ . Applying the same  $\lambda$  to both of them will then tend to lead to a more sparse solution for  $z_2$  than  $z_1$ . This is because networks with different sample sizes should be learned with different amounts of regularization.

At first glance, it seems simple scaling/normalization (such as dividing  $\mathcal{L}^u(\mathbf{X}^{(z)}, \beta_{\setminus u}^{(z)})$  by  $S_z$ ) would be sufficient. Asymptotic theory<sup>12</sup> dictates that in addition to dividing each  $\mathcal{L}^u(\mathbf{X}^{(z)}, \beta_{\setminus u}^{(z)})$  by  $S_z$ ,  $\lambda$  should be divided by  $\sqrt{S_z}$  as shown in Eq 2:

$$\hat{\beta}_{\setminus u}^{(1)}, \dots, \hat{\beta}_{\setminus u}^{(Z)} = \underset{\beta_{\setminus u}^{(1)}, \dots, \beta_{\setminus u}^{(Z)}}{\operatorname{argmin}} \left( \sum_{z \in \mathcal{Z}} \frac{1}{S_z} \mathcal{L}^u(\mathbf{X}^{(z)}, \beta_{\setminus u}^{(z)}) + \sum_{z \in \mathcal{Z}} \frac{\lambda}{\sqrt{S_z}} \|\beta_{\setminus u}^{(z)}\|_1 \right) \quad (2)$$

However, this scaling is based on several theoretical assumptions on the underlying model. As a result, it may behave erratically in practice on microarray data as we show in Section 7. Even when all the theoretical assumptions hold, the  $\sqrt{S_z}$  factor is correct only *asymptotically*, and not necessarily for smaller sample sizes. To illustrate the problem, we present an example shown in Figure 2. (More quantitative results will be given in Section 6.) Here, a single network with 100 vertices and 200 edges was randomly generated. Then, 10 sets with 20 samples, 10 sets with 30 samples, and 10 sets with 40 samples were generated, all from the same network. We vary the sparsity parameter  $\lambda$ , and plot the mean edge count for each sample size. Figure 2(a) shows the results of optimizing Eq. 1 without scaling<sup>a</sup>. As one can see, although all the samples were generated from the same network, the networks learned from the 40 samples have many more edges than those from fewer samples. Figure 2(b) shows the results for optimizing Eq. 2 (with scaling). This works better, but networks learned from the 40 samples still have considerably more edges than those from 20.

One possible strategy is to assign each network a different regularization parameter and tune these manually according to known biological interactions. Unfortunately, this requires

<sup>a</sup>MB stands for Meinshausen and Bühlmann who proposed neighborhood selection<sup>2</sup> for GGMs.

that we have enough prior knowledge about *all* the networks, which is unlikely for many systems. Instead, it is preferable to develop an approach that only requires prior knowledge about a small subset of the networks for the purposes of parameter tuning.

### 3. A More Robust Formulation

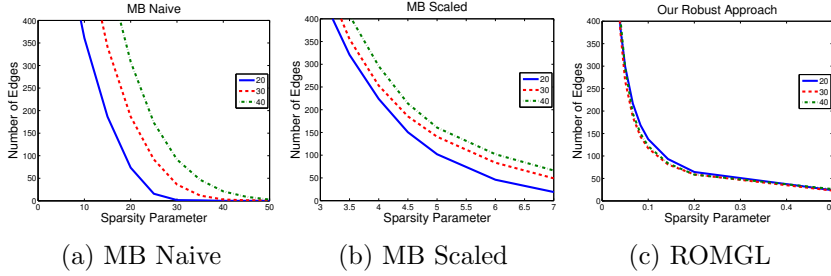


Fig. 2. Comparison of non-robust vs. robust approaches on a simple example. Our robust approach, *ROMGL*, produces networks that are much more balanced than the naive and scaled methods (MB Naive and MB Scaled). See text for details.

the assumptions made in microarray data pre-processing via *normalization* which rely on less than ideal yet necessary assumptions in order to remove systematic dye bias from the data, e.g., quantile normalization in RMA assumes an identical distribution of gene expression values in all samples in a dataset.<sup>13,14</sup>

Rather than post-processing the networks, we integrate this assumption into our network algorithm, thus allowing for a more principled and effective approach.

Unfortunately, it is difficult to directly modify neighborhood selection described in the previous section to incorporate this assumption, because we are constraining the *entire* networks to have the same sum of absolute edge weights, rather than the individual neighborhoods. The former assumption is much more realistic, since the latter implies all the nodes have similar degrees. However, since neighborhood selection estimates each neighborhood independently, it cannot incorporate this assumption in its procedure. Instead, we build our solution from SPACE<sup>15</sup> which is a procedure that simultaneously performs neighborhood selection on all neighborhoods. First define,

$$\mathcal{M}(\mathbf{X}^{(z)}, \boldsymbol{\rho}^{(z)}, \boldsymbol{\sigma}^{(z)}) := \sum_{u \in \mathcal{V}} \sum_{s=1}^{S_z} \left( x_u^{(s,z)} - \sum_{v \neq u} \beta_{uv}^{(z)} x_v^{(s,z)} \right)^2 = \sum_{u \in \mathcal{V}} \sum_{s=1}^{S_z} \left( x_u^{(s,z)} - \sum_{v \neq u} \rho_{uv}^{(z)} \sqrt{\frac{\sigma_{vv}}{\sigma_{uu}}} x_v^{(s,z)} \right)^2 \quad (3)$$

Then, using SPACE to estimate each network  $z \in \mathcal{Z}$  separately will give the following optimization problem:

$$\begin{aligned} \hat{\boldsymbol{\rho}}^{(1)}, \dots, \hat{\boldsymbol{\rho}}^{(Z)} = & \underset{\boldsymbol{\rho}^{(1)}, \dots, \boldsymbol{\rho}^{(Z)}}{\operatorname{argmin}} \left( \sum_{z \in \mathcal{Z}} \frac{1}{S_z} \mathcal{M}(\mathbf{X}^{(z)}, \boldsymbol{\rho}^{(z)}, \boldsymbol{\sigma}^{(z)}) \sum_{z \in \mathcal{Z}} \frac{\lambda}{\sqrt{S_z}} \|\boldsymbol{\rho}^{(z)}\|_1 \right) \\ \text{subject to } & \rho_{uv}^{(z)} = \rho_{vu}^{(z)} \quad \forall z, \forall u \neq v \end{aligned} \quad (4)$$

Similar to the previous sections, the objective above decouples into  $Z$  separate problems.

In order to calibrate the networks to mitigate the artifacts caused by sample size heterogeneity, we propose the following approach. We require that the sum of the absolute edge weights to be the same for all networks reconstructed. This is in some sense similar to

Here  $\sigma_{uv}^{(z)} = 1/\text{var}(\epsilon_u^{(z)})$ , where  $\epsilon_u^{(z)}$  was defined in Section 2.1. Note that SPACE estimates  $\rho$  directly instead of  $\beta$ . This is because while  $\rho_{uv}^{(z)} = \rho_{vu}^{(z)}$ ,  $\beta_{uv}^{(z)} \neq \beta_{vu}^{(z)}$  due to the relation in Section 2.1.

Note that SPACE has the same problem as neighborhood selection with varying sample sizes. However, because we estimate all the neighborhoods jointly, we can propose a new formulation that enforces our assumption. This can be done by requiring the  $\ell_1$  norm of the absolute value of the edge weights to be equal to  $C$  for all  $z \in \mathcal{Z}$ .

$$\begin{aligned} \hat{\rho}^{(1)}, \dots, \hat{\rho}^{(Z)} &= \underset{\rho^{(1)}, \dots, \rho^{(Z)}}{\operatorname{argmin}} \sum_{z \in \mathcal{Z}} \frac{1}{S_z} \mathcal{M}(\mathbf{X}^{(z)}, \rho^{(z)}, \sigma^{(z)}) \\ \text{subject to } \rho_{uv}^{(z)} &= \rho_{vu}^{(z)} \quad \forall z, \forall u \neq v, \quad \|\rho^{(1)}\|_1 = C, \|\rho^{(2)}\|_1 = C, \dots, \|\rho^{(Z)}\|_1 = C \end{aligned} \quad (5)$$

The formulation above represents the foundation of our approach, which we call *ROMGL* (*RObust Multi-network Graphical Lasso*). Note that this formulation is different than that in Eq. 4, because if we write it in Lagrangian form with  $\lambda$ 's instead of constraints, then it is equivalent to a different  $\lambda$  for each constraint

$$\begin{aligned} \hat{\rho}^{(1)}, \dots, \hat{\rho}^{(Z)} &= \underset{\rho^{(1)}, \dots, \rho^{(Z)}}{\operatorname{argmin}} \left( \sum_{z \in \mathcal{Z}} \frac{1}{S_z} \mathcal{M}(\mathbf{X}^{(z)}, \rho^{(z)}, \sigma^{(z)}) + \sum_{z \in \mathcal{Z}} \lambda_z \|\rho^{(z)}\|_1 \right) \\ \text{subject to } \rho_{uv}^{(z)} &= \rho_{vu}^{(z)} \quad \forall z, \forall u \neq v \end{aligned} \quad (6)$$

Moreover, without solving the optimization problem, the correspondence between  $C$  and the set of equivalent  $\{\lambda_z\}_{z \in \mathcal{Z}}$  is unknown. Thus, the advantage of our approach is that we only have to explicitly set one parameter  $C$  instead of a different  $\lambda$  for each  $z \in \mathcal{Z}$  (since  $|\mathcal{Z}|$  might be quite large). We demonstrate our approach in Figure 2. Unlike the non-robust methods, our approach returns edge counts that are more similar across the different sample sizes.

#### 4. Sharing Information Across States

So far, we have discussed robustly estimating a collection of networks without sharing information among different cell states. However, in the small-sample-size scenarios prevalent in regulatory genomics, this can result in poor estimation quality of the networks. For example, in the hematopoietic stem cell dataset we consider, some of the cell states have only 4 microarray samples, which is clearly statistically insufficient for reliable network estimation. However, since in many cases the gene networks are related, such as in a time series or a genealogy, we can leverage this interconnectedness of the networks for more accurate network reconstruction.

We assume we have *prior knowledge* of which networks are biologically related, and this information is encoded as a graph over the cell states  $\mathcal{Z}$ , which we denote by  $\mathcal{H} = (\mathcal{Z}, \Gamma)$ .  $\mathcal{H}$  is constructed such that cell states closer to one another in the graph are assumed to be more biologically similar than those farther apart. For cells over a tree genealogy (e.g. stem cell differentiation),  $\mathcal{H}$  represents a tree, and cell state  $z$  is connected to its parent and sibling cell states. As stated earlier, several methods<sup>4-6</sup> have recently proposed leveraging similarities of multiple networks for more accurate multi-network estimation. KELLER<sup>4</sup> proposes kernel smoothing, which estimates a given network by pooling a weighted average of related samples. TESLA and Treegl propose total variation regularization.<sup>5,6</sup>

However, these methods do not account for sample size heterogeneity. In fact, when sharing information among related states, robustness to sample size heterogeneity is even more crucial. This is because different cell states may have different numbers of neighbors in  $\mathcal{H}$ , and thus some may be able to share more information than others.

For simplicity, we only discuss how our robust formulation can be incorporated with kernel smoothing. Consider a smoothing kernel  $K_h(z, y)$  that defines a similarity between cell state  $z$  and cell state  $y$ . We use the Epanechnikov kernel:  $K_h(z, y) = 1 - \left(\frac{d(z, y)}{h}\right)^2$  if  $\frac{d(z, y)}{h} \leq 1$ , and 0 otherwise. Here we define  $d(z, y)$  to be the shortest path from  $z$  to  $y$  in  $\mathcal{H}$ . Intuitively, this means that cell states closer to one another in the graph are assumed to be more biologically similar than those farther apart. Note that this is a more general setting than Song et al.,<sup>4</sup> who merely consider smoothing over time. We can then estimate a network for a cell state using a weighted average of samples from all cell states via the kernel:

$$\begin{aligned} \hat{\rho}^{(1)}, \dots, \hat{\rho}^{(Z)} = \underset{\rho^{(1)}, \dots, \rho^{(Z)}}{\operatorname{argmin}} \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Z}} K_h(z, y) \mathcal{M}(\mathbf{X}^{(y)}, \rho^{(z)}, \sigma^{(z)}) \\ \text{subject to } \rho_{uv}^{(z)} = \rho_{vu}^{(z)} \quad \forall z, \forall u \neq v, \quad \|\rho^{(1)}\|_1 = C, \dots, \|\rho^{(Z)}\|_1 = C \end{aligned} \quad (7)$$

We term this approach *ROMGL-Smooth* (an abbreviation for *Kernel-Smoothed ROMGL*).

## 5. Optimization

We briefly describe how to optimize Eq. 7. The objective is separable in  $z \in \mathcal{Z}$ , and thus each  $\{\rho^{(z)}, \sigma^{(z)}\}$  pair can be optimized separately from the other  $z' \neq z$ . However, Eq. 7 is not jointly convex in both  $\rho^{(z)}$  and  $\sigma^{(z)}$ . Fortunately, given a fixed  $\sigma^{(z)} = \bar{\sigma}^{(z)}$ , the problem is convex in  $\rho^{(z)}$ . Similarly, given a fixed  $\rho^{(z)} = \bar{\rho}^{(z)}$  we can update  $\sigma^{(z)}$ . Thus, we proceed by alternatively updating  $\rho^{(z)}$  and  $\sigma^{(z)}$ .

To optimize  $\rho^{(z)}$  given a fixed  $\bar{\sigma}^{(z)}$ , we use a projected gradient method, where after updating the current value of  $\rho^{(z)}$  in the direction of the gradient, it is projected back onto the constraint set. For our constraint, the projection can be done very efficiently in  $O(n \log n)$  time using the method of Duchi et al.<sup>16</sup> Updating  $\sigma^{(z)}$  given a fixed  $\bar{\rho}^{(z)}$  can be done using a similar update to traditional SPACE:  $\frac{1}{\bar{\sigma}_{uv}^{(z)}} \leftarrow \frac{1}{\sum_{y \in \mathcal{Z}} K_h(z, y)} \sum_{y \in \mathcal{Z}} K_h(z, y) \mathcal{M}^u(\mathbf{X}^{(y)}, \bar{\rho}^{(z)}, \bar{\sigma}^{(z)})$ .

## 6. Synthetic Evaluation

We first focus on synthetic data where the modelling assumptions hold. Our *ROMGL-Smooth* (Eq. 7) can naturally be compared with a Gaussian Graphical Model (GGM) version of KELLER<sup>4</sup> which also uses kernel smoothing. We find that in this case the  $\sqrt{S_z}$  scaling (Eq. 2) performs better than the naive approach (Eq. 1), and therefore only compare our approach to GGM KELLER with scaling (which we refer to as *MB-Smooth Scaled*) in this section.

We performed the experiments with two types of graphs: Erdos Renyi random graphs and sparse graphs with hubs. For each type, we generate a sequence of graphs of length 25. Each graph in the sequence has 100 vertices and 200 edges, and is created by randomly deleting and adding 10 edges from the previous graph. The sample size is 30 for the first five graphs, 35 for the next 5, and so on up to 50 for the last 5 graphs. Note that all graphs have the same number of edges (even though they are not identical). We run both methods for  $h = \{2, 3\}$ , for a variety of regularization parameters, and repeat each experiment for 5 different graph



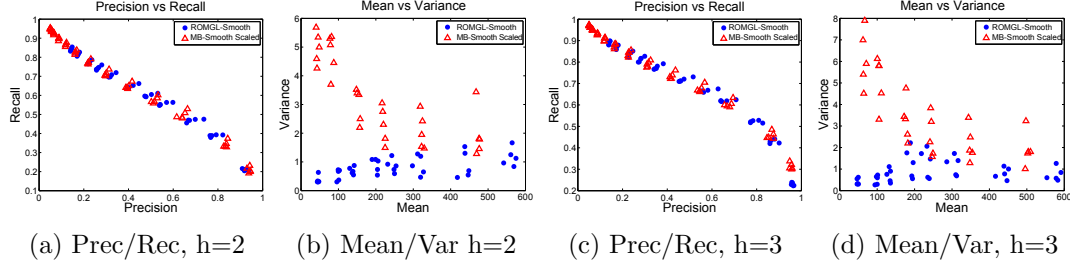


Fig. 3. Comparison of our robust approach, *ROMGL-Smooth* (blue circles), with an existing non-robust method, *MB-Smooth Scaled* (red triangles), on synthetic Erdos Renyi random graphs. See text for details.

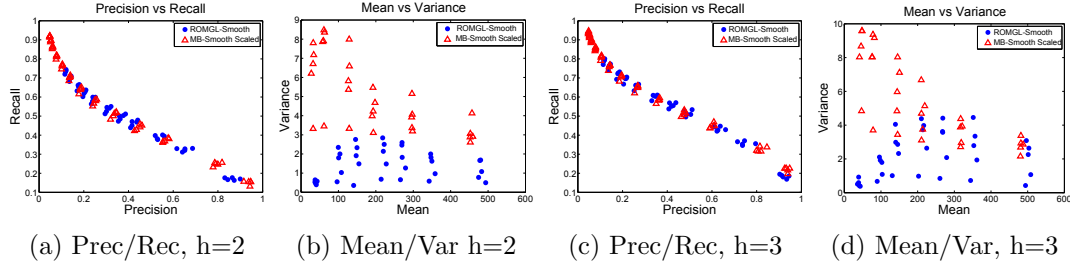


Fig. 4. Comparison of our approach, *ROMGL-Smooth* (blue circles), with an existing non-robust method, *MB-Smooth Scaled* (red triangles), on synthetic sparse graphs with hubs. See text for details.

sequences. The methods are evaluated on two different criteria. To measure accuracy of the approaches in recovering the structures we plot precision/recall curves. The precision is defined as  $prec = \frac{1}{Z} \sum_{z \in Z} \frac{|\hat{\mathcal{E}}^{(z)} \cap \mathcal{E}^{(z)}|}{|\hat{\mathcal{E}}^{(z)}|}$  and the recall is defined as  $rec = \frac{1}{Z} \sum_{z \in Z} \frac{|\hat{\mathcal{E}}^{(z)} \cap \mathcal{E}^{(z)}|}{|\mathcal{E}^{(z)}|}$ .

We also propose a quantitative measure of robustness. Let  $\hat{e} = (|\hat{\mathcal{E}}^{(1)}|, \dots, |\hat{\mathcal{E}}^{(Z)}|)$  be the vector of edge counts of the networks recovered by a method. Intuitively, if a method is robust to sample size heterogeneity, the variance of  $\hat{e}$  should be small, since all the true graphs have the same number of edges. Thus, we propose the quantity  $var(\hat{e})/mean(\hat{e})$  as a measure of robustness (scaling by  $mean(\hat{e})$  provides for easier comparison).

The precision/recall curves show that both methods perform comparably according to this metric (Figures 3(a), 3(c), 4(a), and 4(c)), indicating that our new robust approach generates results with comparable accuracy as the scaling method. However, our new approach yields results with considerably lower variance, indicating that it is more robust than the scaling method (Figures 3(b), 3(d), 4(b) and 4(d)). This is especially true when the recovered graphs are sparser, since *MB-Smooth Scaled* has very high variance in this case. This is the most prevalent scenario, since on many real biology datasets, the sample size is small, so we are more likely to select sparse graphs. Furthermore, as we will see, the scaling method performs much worse on real data than synthetic data.

## 7. Application to the Hematopoietic Stem Cell Dataset

We applied our method to the human hematopoietic stem cell dataset analyzed in Novershtern et al.<sup>17</sup> There are 38 cell states in the tree-shaped multi-lineage stem cell genealogy. We focus on a subset of 732 genes from the entire dataset for the experiments in this section.

First, we quantitatively compare our approach (*ROMGL-Smooth*) to the non-robust approaches: naive (*MB-Smooth Naive*) and scaling (*MB-Smooth Scaled*). The bandwidth for

these algorithms was fixed to 5. For a given setting of the regularization parameter ( $\lambda$  or  $C$ ), we plot the average edge count over all the 38 cell states on the x-axis and the difference between the largest edge count and the smallest edge count on the y-axis. As shown in Figure 5, the non-robust methods produce networks with very different sizes, e.g., some of the networks have less than 100 edges while others have thousands. Our robust approach produces much more calibrated results.

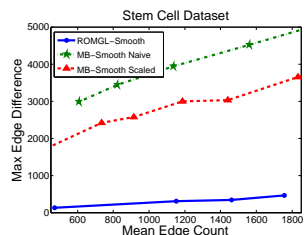


Fig. 5. Our approach, denoted by *ROMGL-Smooth* (blue) compared with *MB-Smooth Scaled* (red) and *MB-Smooth Naive* (green) on the hematopoietic stem cell dataset. Our approach returns networks that are much more calibrated with more similar edge counts.

see, for the naive approach (Figures 6(a) and 6(b)), sample size heterogeneity is such a problem that the GRAN3 network has zero edges while the CMP network has 4532. Similarly, the scaling approach also performs poorly. The GRAN3 network has only 72 edges (Figure 6(c)) while the CMP network has 2944 edges (Figure 6(d)). Thus, with both of these approaches, it is practically impossible to analyze the GRAN3 network in relation to the other networks. In contrast, our approach gives much more balanced results; the GRAN3 network has 1269 edges (Figure 6(e)) while the CMP network has 1614 edges (Figure 6(f)).

Next, we examined the results generated by our robust approach in more detail. Novershtern et al.<sup>17</sup> discovered various gene modules and their corresponding regulators active in different cell states in the hematopoietic stem cell dataset. It is unknown, however, how genes in these modules interact with one another. We compare and contrast our results to theirs on the two modules 721 and 817 described in Novershtern et al.<sup>17</sup> The former module is induced in granulocytes and monocytes (GRAN/MONO), while the other in B cells, T cells, and granulocytes (BCELL/TCELL/GRAN).

The subnetworks corresponding to the GRAN/MONO 721 module we recovered in the granulocytes and monocytes are shown in Figure 7 (a) and (b). It can be seen that we recovered all the genes in the module for both subnetworks, which include both experimentally verified ones (shown in dark purple and dark green) and unverified ones (light green). Note almost all of the proposed genes in the module are within 2-3 hops from the regulators CEBPD and MNDA in the GRAN3 and MONO2 subnetworks. Moreover, our results reveal interaction patterns of the genes in these subnetworks (only a list of genes in the module was shown in Novershtern et al.<sup>17</sup>). A closer examination of the two subnetworks reveals that they contain

To examine these differences further, we show cell-specific networks for two cell states, granulocytes (GRAN3) and common myeloid progenitors (CMP), recovered by the three approaches in Figure 6. GRAN3 is a leaf in the cell genealogy; it has few neighbors and the lowest effective sample size (14.92) when the smoothing kernel is applied. In contrast, CMP is an internal node in the genealogy that can differentiate into megakaryocytes, erythrocytes, granulocytes, and monocytes, and thus has many neighbors; it has the highest effective sample size (60.52). As one can

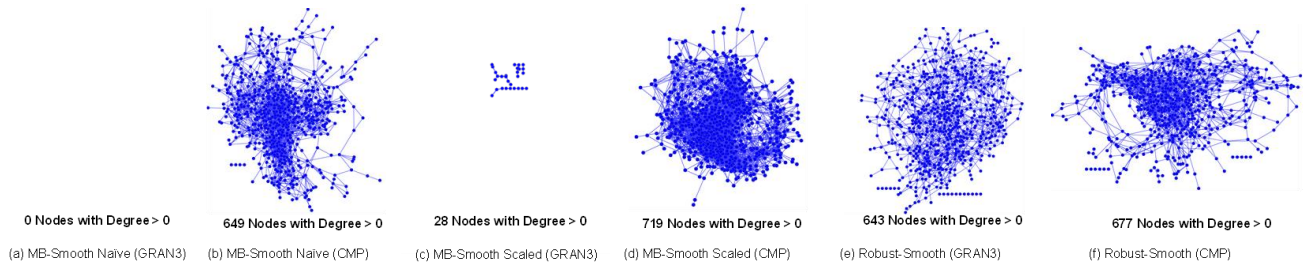


Fig. 6. The cell-state-specific networks for granulocytes (GRAN3) and common myeloid progenitors (CMP) recovered by the three approaches. The robust approach (*ROMGL-Smooth*), shown in (e) and (f), produces substantially more balanced networks than the other two approaches.

two modules with similar gene interaction patterns, one is a large 10-gene module with MNDA, CREB5, VDR, RAB31, NOD2, CEBPD, CFP, MYCL1, WDFY3, and VENTX, and the other is a small 2-gene module with HBEGF and ATF3. Interestingly, 7 out of these 12 genes were also proposed by Novershtern et al.

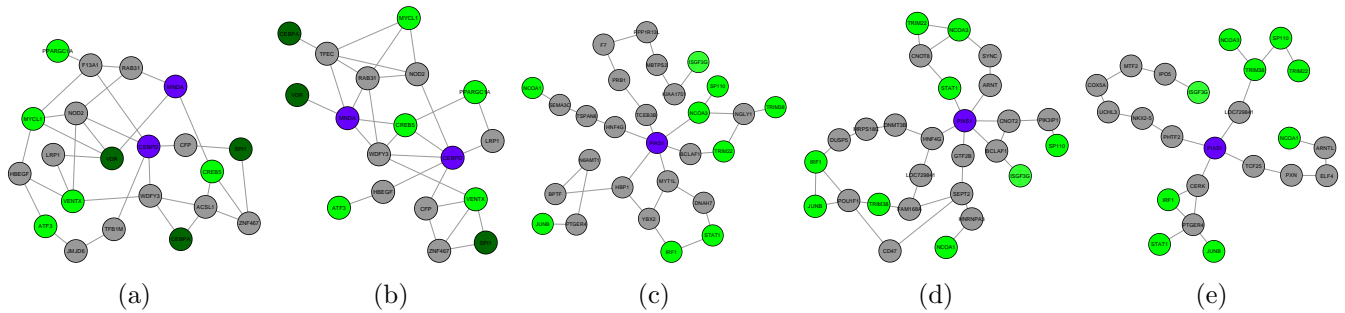


Fig. 7. The *ROMGL-Smooth* reconstructed subnetworks corresponding to (a) module 721 in granulocytes (GRAN3), and (b) module 721 monocytes (MONO2) (c) PIAS1 module in B cells (BCELLa3), (d) PIAS1 module in T cells (TCELL3), and (e) PIAS1 module in granulocytes (GRAN3). Purple represents genes that are regulators of the module and were experimentally validated in Novershtern et al.<sup>17</sup> Dark green represents other genes in the module that were experimentally validated. Light green represents the genes in the module which were not experimentally validated. All the other genes are colored gray.

Finally, we examined the reconstructed subnetworks in B cells (BCELLa3), T cells (TCELL3), and granulocytes (GRAN3) corresponding to the BCELL/TCELL/GRAN 817 module in Novershtern et al.<sup>17</sup> (Figure 7 (c),(d),(e)). In this case, the topologies of the subnetworks are very different. The only gene module shared between the BCELLa3 and TCELL3 subnetworks is HNF4G–PIAS1–BCLAF1. In addition, the topology of the GRAN3 subnetwork corresponding to the BCELL/TCELL/GRAN 817 module is distinctly different from the BCELLa3 and TCELL3 subnetworks. These findings are consistent with the fact that both B cells and T cells are lymphocytes and closer in the genealogy than granulocytes.

## 8. Discussion

In conclusion, we have identified the problem of sample size heterogeneity in multi-network reconstruction and proposed a principled solution that works well in practice. Our method assumes that all networks have approximately the same number of edges. However, more

complex assumptions are possible if we have prior knowledge about the network densities. For example, we can assume cell states in a certain category each have sum of absolute edge weights equal to  $C_1$ , while cell states in another category are associated with parameter  $C_2$ .

**Acknowledgements** This research was made possible by Grants NIH 1R01GM093156 and NIH 1R01GM087694, and an NSF Graduate Fellowship (Grant No. 0946825) to APP

## References

1. R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267 (1996).
2. N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, *Annals of Statistics* **34**, 1436 (2006).
3. J. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **9**, 432 (2008).
4. L. Song, M. Kolar and E. Xing, Time-Varying Dynamic Bayesian Networks, *Bioinformatics* **25**, p. i128 (2009).
5. A. Ahmed and E. Xing, Recovering time-varying networks of dependencies in social and biological studies, *Proceedings of the National Academy of Sciences* **106**, p. 11878 (2009).
6. A. Parikh, W. Wu, R. Curtis and E. Xing, TREEGL: reverse engineering tree-evolving gene networks underlying developing biological lineages, *Bioinformatics* **27**, i196 (2011).
7. E. Segal, H. Wang and D. Koller, Discovering molecular pathways from protein interaction and gene expression data, *Bioinformatics-Oxford* **19**, 264 (2003).
8. A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao and M. West, Sparse graphical models for exploring gene expression data, *Journal of Multivariate Analysis* **90**, 196 (2004).
9. D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques* (MIT press, 2009).
10. S. Lauritzen, *Graphical models* (Oxford University Press, USA, 1996).
11. M. Wainwright, P. Ravikumar and J. Lafferty, High-Dimensional Graphical Model Selection Using  $\ell_1$ -Regularized Logistic Regression, *Advances in Neural Information Processing Systems* **19**, p. 1465 (2007).
12. M. Wainwright, Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using Constrained Quadratic Programs, *Information Theory, IEEE Transactions on* **55**, 2183 (2009).
13. R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf and T. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics* **4**, 249 (2003).
14. B. Bolstad, R. Irizarry, M. Åstrand and T. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* **19**, 185 (2003).
15. J. Peng, P. Wang, N. Zhou and J. Zhu, Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association* **104**, 735 (2009).
16. J. Duchi, S. Shalev-Shwartz, Y. Singer and T. Chandra, Efficient projections onto the  $\ell_1$ -ball for learning in high dimensions, in *Proceedings of the 25th international conference on Machine learning*, 2008.
17. N. Novershtern, A. Subramanian, L. Lawton, R. Mak, W. Haining, M. McConkey, N. Habib, N. Yosef, C. Chang, T. Shay *et al.*, Densely interconnected transcriptional circuits control cell states in human hematopoiesis, *Cell* **144**, 296 (2011).

**VARIANT PRIORIZATION AND ANALYSIS INCORPORATING PROBLEMATIC REGIONS OF THE  
GENOME**

**ANIL PATWARDHAN**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [apatwardhan@personalis.com](mailto:apatwardhan@personalis.com)*

**MICHAEL CLARK**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [michael.clark@personalis.com](mailto:michael.clark@personalis.com)*

**ALEX MORGAN**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [alex.morgan@personalis.com](mailto:alex.morgan@personalis.com)*

**STEPHEN CHERVITZ**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [schervitz@personalis.com](mailto:schervitz@personalis.com)*

**MARK PRATT**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [mark.pratt@personalis.com](mailto:mark.pratt@personalis.com)*

**GABOR BARTHA**

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [gabor.bartha@personalis.com](mailto:gabor.bartha@personalis.com)*

GEMMA CHANDRATILLAKE

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [gemma.chandratillake@personalis.com](mailto:gemma.chandratillake@personalis.com)*

SARAH GARCIA

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [sarah.garcia@personalis.com](mailto:sarah.garcia@personalis.com)*

NAN LENG

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [nan.leng@personalis.com](mailto:nan.leng@personalis.com)*

RICHARD CHEN

*Personalis Inc., 1350 Willow Road, Suite 202  
Menlo Park, CA, 94025, USA  
Email: [richard.chen@personalis.com](mailto:richard.chen@personalis.com)*

In case-control studies of rare Mendelian disorders and complex diseases, the power to detect variant and gene-level associations of a given effect size is limited by the size of the study sample. Paradoxically, low statistical power may increase the likelihood that a statistically significant finding is also a false positive. The prioritization of variants based on call quality, putative effects on protein function, the predicted degree of deleteriousness, and allele frequency is often used as a mechanism for reducing the occurrence of false positives, while preserving the set of variants most likely to contain true disease associations. We propose that specificity can be further improved by considering errors that are specific to the regions of the genome being sequenced. These problematic regions (PRs) are identified a-priori and are used to down-weight constitutive variants in a case-control analysis. Using samples drawn from 1000-Genomes, we illustrate the utility of PRs in identifying true variant and gene associations using a case-control study on a known Mendelian disease, cystic fibrosis(CF).

## 1. Introduction

Exome sequencing is a potentially powerful tool in detecting variants and genes responsible for both simple and complex diseases. Recent successes in identifying the causal variants of several Mendelian or monogenic disorders<sup>1-4</sup> have highlighted the utility of heuristic methods of variant filtering and prioritization in the discovery process. These methods often preferentially retain or prioritize variants based on novelty, functional impact, putative effects in the protein coding regions (i.e. missense/nonsense substitutions, coding indels, and splice site-acceptor and donor sites), population frequency, and/or concordance with a subjective assessment of phenotypic features<sup>5</sup>. This biologically informed reduction in the number of variants helps maintain statistical power by reducing the number of formally tested hypotheses and the subsequent impact of multiple testing correction procedures required in high-throughput experiments<sup>6</sup>.

While these strategies may enrich the set of disease-associated variants based on variant/functional-level information and disease phenotype, they do not directly address the occurrence of false positives stemming from sequencing inaccuracies. Exome sequencing coverage varies greatly across the genome<sup>7-8</sup> with some regions under-covered due to areas of low-complexity, areas of high GC content, and the occurrence of segmental duplications and homopolymers<sup>9-10</sup>. In case-control studies investigating variant-disease associations, alignment and mapping errors in these problematic regions (PRs) reduces the sensitivity to detect true associations in these regions and may introduce false positive associations in instances where cases and controls have differential coverage depths<sup>11</sup>. The integration of PR information in a case-control analysis may help identify false discoveries not readily identified by other commonly used methods of variant prioritization.

Using a well-characterized set of samples drawn from 1000-Genomes<sup>12</sup> we illustrate the utility of PRs in resolving known causal variants in cystic fibrosis (CF). Combined with other variant prioritization methods, the use of PRs improves the specificity of both standard variant association tests and gene-level collapsing methods in identifying true associations despite limited sample sizes.

## 2. Methods

### 2.1. Subject Samples

All DNA samples were drawn from the 1000Genomes project. Samples were drawn from a pool of subjects broadly identified as Caucasian and known to be affected (cases) or unaffected (controls) with CF. Information regarding ethnicity, sex, and known mutations in this group of samples, including those samples harboring the  $\Delta F508$  common founder mutation, were obtained from the *CFTR* Human Gene Mutation Panel records at the Center for Disease Control<sup>13</sup> and Coriell Institute for Medical Research<sup>14</sup> websites. Cases and controls were sequenced separately using identical platforms and technologies. Raw sequencing data were aligned and variants were called simultaneously for all case and control samples.

## **2.2. Genomic Library Construction, Exome Sequencing, Alignment and Variant Calling**

DNA libraries were prepared using Illumina TruSeq Genomic DNA High throughput Sample Prep Kits (Illumina, San Diego, CA) and exome enrichment (targeting 62Mb) was accomplished using the TruSeq Exome Target Enrichment kit (Illumina, San Diego, CA) according to manufacturer's protocols. Sequencing was performed using Illumina HiSeq2000 or HiSeq2500 sequencers with single lane, paired-end 2X100bp reads. DNA fragments were generated and amplified using Clonal Single Molecule Array technology (Illumina, San Diego, CA). The sequences were determined using the Clonal Single Molecule Array and Sequencing-by-Synthesis using Illumina's proprietary instrumentation and Reversible Terminator Chemistry. Sequencing reads of at least 2x100bp in length for a total of at least 8Gb of sequence data per sample were generated for each sequenced sample.

Raw sequence data were in FASTQ format and were analyzed in multisample mode with standard (Sanger) Phred-scale quality scores. The Pipeline then uses an integrated set of proprietary and public analysis tools to align and variant call genomic sequencing data. Gapped alignment is performed using the popular Burrows-Wheeler Aligner (BWA) combined with Picard and the Genome Analysis Toolkit (GATK) to improve sequence alignment and to correct base quality scores. Data was aligned to the hg19 genome, producing standard, compressed Binary Alignment Map (BAM) format files.

GATK's Unified Genotyper module provides the Pipeline's core set of SNV calls and their accompanying quality metrics. Calls are enhanced by proprietary SNV accuracy software which incorporates both genomic context and sequence alignment information into a model that corrects miscalled loci. All calls are made on BAM files that have been recalibrated by GATK's base quality score recalibration (BQSR). SNV and small indels are reported in VCF format. Reference calls and no-call information is returned in BED files.

Variants were annotated using the Personalis Annotation Engine, which applied population frequencies, genetic region information, effect on genes, protein impact, protein-protein interactions and additional structural and functional features to the variants.

## **2.3. Problematic Regions of the Genome**

Based on a previous study of discordant variant calls among multiple sequencing platforms<sup>8,15</sup> and further work in elucidating the mechanisms underlying these errors<sup>16</sup>, a database of PRs was constructed. PRs are comprised of regions having >3X the average error rate seen among variant calls deemed high-quality by VQSR (i.e. largely PASS calls). PRs included those regions of the genome with high GC content, low coverage, degeneracy due to redundant paralogous sequences, low complexity repetitive elements, segmental duplications, and compression regions<sup>17</sup> for which large amounts of discordance in variant calls were previously observed. It also includes HLA regions and breakpoint library regions for structural variants (BreakSeq<sup>18</sup>). While PR regions are not always mutually exclusive in terms of their categorization, the bulk of PRs (~70%) are due to 100bp regions having >70%



GC content, degenerate 100bp single reads, and simple repeats > 100bp long. Variants called in the case-control analyses were mapped onto the PR database and were flagged as potentially problematic variant calls if they fell into a PR region.

#### **2.4. Case Control Analysis**

Eighteen unrelated subjects with CF were matched to 54 unrelated and unaffected subjects based on sex and broad ethnic category (i.e. Caucasians) to form a 1:3 case-control study design. In a second case-control analysis, the case-group was redefined to only include the subset of CF-affected individuals without the  $\Delta F508$  founder mutation. These 8 case-subjects were again compared to the same 54 unaffected control subjects to form a ~1:7 case-control match. Analysis was performed independently in each of these case-control studies to investigate variant and gene associations with CF.

In each study, variants were removed from analysis if they failed our internal QC requirements. These QC standards required that 1) no more than 10% of the data was missing across case samples and/or control samples and 2) the multi-sample variant call from GATK's Variant Quality Score Recalibration (VQSR) was "PASS"- indicating that there was sufficient evidence that the site was really variant in one or more samples. In order to reduce the likelihood of false discoveries when reporting CF-associated variants and genes, variants were also filtered to retain only those that were protein-coding. These filtering criteria were used when reporting variant-level associations with CF and as input criteria when testing for gene-level associations.

Remaining variants were assessed for association with disease-status using Fisher's Exact Test. Effect size was summarized as the Odds Ratio (OR) calculated from the conditional maximum likelihood estimate of a 2x2 contingency table containing alternative and reference allele counts in cases and controls assuming an additive model. Significance testing of the null of conditional independence (OR=1) used a two-tailed test.

Analysis of the second case-control study, in which all cases with the  $\Delta F508$  founder mutation were removed, was done to investigate the occurrence of PRs in studies where smaller effect sizes among causal variants could be expected. This required detection strategies that could accommodate the genetic heterogeneity of the remaining affected individuals- since known causal variants were interspersed throughout the *CFTR* gene<sup>13-14</sup>. Given the challenges in detecting rare variant enrichment with a limited number of heterogeneous case samples, we collapsed the variant-level associations based on gene-membership. An implementation of the Combined Multivariate and Collapsing (CMC) method<sup>19</sup> was used to assess the combined association of variants within the same gene to CF. Variants were binned into groups based on their respective gene membership and further binned (rare vs. common) based on a 1000-Genome derived minor allele frequency (MAF) cutoff of 5%. A multivariate test, Hotelling T-squared, was performed on the counts within all bins to determine differences among the cases and controls with asymptotic p-

values calculated based on the F-distribution. The method of Storey<sup>20</sup> was used to calculate FDR-adjusted p-values (i.e. q-values).

### 3. Results

The application of filtering criteria related only to QC-criteria (i.e. variant-call quality and missing data) among the 18 cases and matched controls, resulted in 541,119 variants for which association with CF was tested. Distribution of observed  $-\log_{10}(\text{p-value})$  revealed departure from the expected distribution and severe inflation of type-1 error (Figure 1). Filtering of variants to include only those that were protein-coding reduced the number of variants 10-fold (54,178) and improved data characteristics.

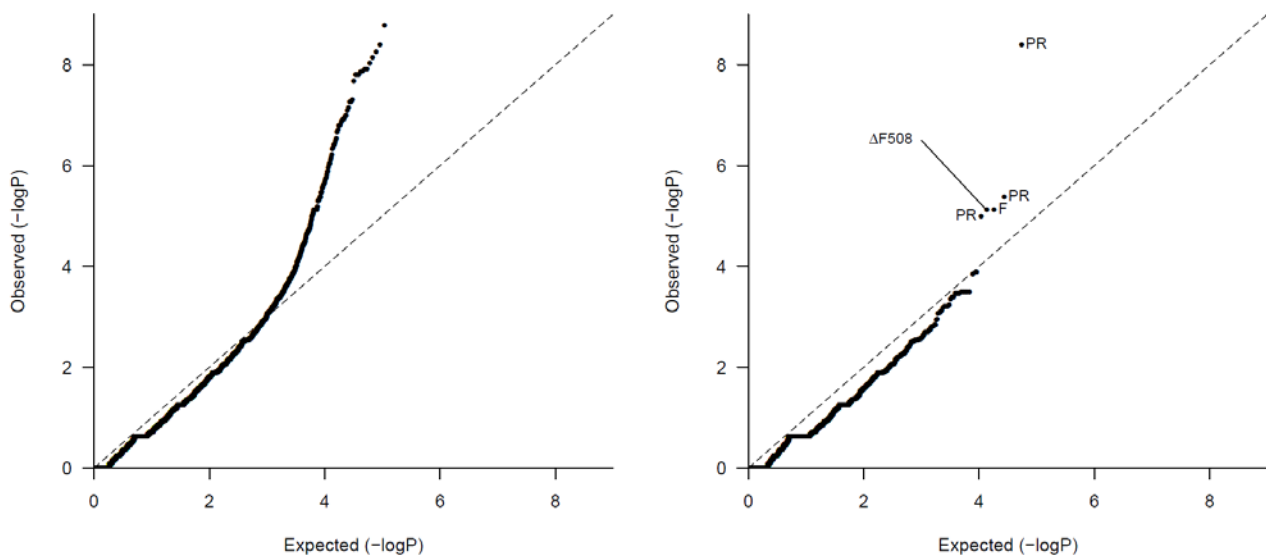


Figure 1. Q-Q Plot comparing the expected normal distribution of  $-\log(\text{p-values})$  to the observed distribution revealed inflated Type-I error when only data quality filters are applied (left). Filtering variants to include only those that are protein-coding (right) improves the data characteristics and revealed significant ( $p < 10^{-5}$ ) true associations ( $\Delta F508$ ), false positives occurring in problematic regions (PR), and false positives that would have been removed based on low allele frequency requirements (F).

Given the number of variants available after QC-criteria and protein-coding filters were applied, an exome-wide significance threshold was set at a p-value of  $10^{-5}$ . At this level, five variants were significantly associated with CF-status, including the known causal variant  $\Delta F508$  (rs199826652) that was present in eight affected individuals. Three variants, all indels, occurred in PR regions (Figure 1, "PR"), and one SNP was a missense mutation in the gene *DOK3* (Table 1). For variants occurring in PRs of the genome, the underlying presence of simple repeats (*POU4F2*, *KIAA0664*) or interspersed repeats (*COPB1*) caused sequencing

errors. The SNP, rs3749728, had an allele frequency of 14% according to 1000-Genomes, and would have been identified as a likely false positive based on low frequency assumptions often used for rare, Mendelian disorders (Figure 1, “F”).

Table 1. Five variants significantly ( $p < 10^{-5}$ ) associated with CF-status after applying QC criteria and protein-coding filters. Also shown are the associated frequencies (MAF) and occurrences in PRs

dbSNP/Gene	Chromosome Position	Ref/Alt Allele	MAF	PR	p-value
<i>POU4F2</i>	Chr4: 147560457	TGGCGGCGGCGGC/ TGGC,TGGCGGCGGCGGCGGC,TGGCGGC,T		Yes	$4.0 \times 10^{-9}$
<i>COPB1</i>	Chr11: 14521144	CGTA/C		Yes	$4.2 \times 10^{-6}$
rs3749728/ <i>DOK3</i>	Chr5: 176936819	C/G	14%	No	$7.7 \times 10^{-6}$
rs199826652/ <i>CFTR</i>	Chr7: 117199644	ATCT/A	1%	No	$7.7 \times 10^{-6}$
<i>KIAA0664</i>	Chr17: 2595272	GCCCCGCCACGCCCCGCCGCGCACCTG/ G,GCCCCGCCGCGCACCTG		Yes	$1.0 \times 10^{-5}$

Aside from the  $\Delta F508$  mutation, no other variants in the *CFTR* gene occurred in more than 3 case samples, reflecting the genetic heterogeneity of CF. Since an analysis of only case-samples not harboring the  $\Delta F508$  founder mutation would be severely underpowered to detect the smaller effect sizes of the remaining *CFTR* variants, we aggregated variant effects based on gene-membership (i.e. collapsing). Subsequent association testing of 10522 genes with CF-status revealed 15 genes with FDR-controlled p-values (q-values)  $< .05$ . Of these, *CFTR* was ranked the 4<sup>th</sup> gene by p-value. Table 2 summarizes these 15 genes, the nominal p-values derived from the CMC test-statistic, the number of variants contributing to the test statistic, the percentage of those variants found in PRs and the predominant PR type. Collectively, out of the 27 variants occurring in PRs and contributing to these collapsing results, the majority (14) occurred in areas of high-GC content, 10 occurred among segmental duplications, and the remaining occurring among areas of low complexity/simple repeats. Notably one gene association listed in Table 2, *ATF7IP2*, had no constitutive variants in PRs, yet was ranked higher than the known causal gene (i.e. *CFTR*). Further examination of this result revealed good coverage in this area across samples indicating that this was likely reflecting a true difference between cases and controls. However, 3 out of 4 constitutive variants had MAFs  $> 5\%$ , indicating that these differences are unlikely to be causally related to CF and would be typically excluded using MAF threshold filters.

Table 2. Collapsing results using only CF-affected samples without the  $\Delta F508$  mutation revealed 15 genes with q-values < 0.05. The nominal p-values, the percentage of those variants in PRs, and the predominant PR type is shown.

Gene	p-value	Number of variants	Percentage of variants in PR	PR types
<i>POU4F2</i>	$1.5 \times 10^{-9}$	2	100%	Repetitive sequence, High GC
<i>MSX1</i>	$4.9 \times 10^{-8}$	5	40%	High GC
<i>ATF7IP2</i>	$2.7 \times 10^{-7}$	4	0%	--
<b><i>CFTR</i></b>	<b><math>4.5 \times 10^{-7}</math></b>	<b>29</b>	<b>0%</b>	--
<i>FUZ</i>	$1.2 \times 10^{-6}$	3	33%	High GC
<i>C8orf74</i>	$1.2 \times 10^{-6}$	5	40%	High GC
<i>TRIM10</i>	$1.2 \times 10^{-6}$	7	0%	--
<i>COL6A1</i>	$3.4 \times 10^{-6}$	6	33%	High GC
<i>PTK2B</i>	$6.1 \times 10^{-6}$	8	0%	--
<i>FAM108A1</i>	$1.2 \times 10^{-5}$	2	100%	Segmental Duplication
<i>MAP7D1</i>	$1.6 \times 10^{-5}$	7	43%	High GC
<i>SCN10A</i>	$1.8 \times 10^{-5}$	13	0%	--
<i>DIDO1</i>	$3.5 \times 10^{-5}$	4	24%	High GC
<i>FLG</i>	$4.0 \times 10^{-5}$	9	89%	Segmental Duplication
<i>COPB1</i>	$8.7 \times 10^{-5}$	2	50%	Repetitive sequence

#### 4. Discussion

In retrospective observational studies of disease association, where disease-affected samples (cases) may be compared to previously sequenced shared controls, alignment and mapping errors can create false evidence for polymorphisms when there are differences in coverage and read depth between groups. Recent evidence has shown that these types of errors can persist when the same genome is sequenced twice under identical analytical environments<sup>16</sup>. Even in carefully designed case-control studies, where samples are matched appropriately and are collected, sequenced, and analyzed together to avoid experimental bias, these errors reduce statistical power for detecting true disease associations.<sup>20</sup> Reduction in these errors are essential for many diseases in which it is a challenge to sufficiently power a case-control study, and is particularly important for complex diseases in which filtering based on frequency thresholds and functional impact may not be appropriate, and where expected effect sizes for a single variant/gene are small or moderate.

CF and the associated study samples used here provide a dataset well-suited to testing the effects of PRs on detection specificity, given that the underlying causal gene and mutations are well-known. Even with a limited number of case samples, we are sufficiently powered to detect variants or genes known to be associated with CF, but suffer from an inflated Type-I error rate. While the effects of these errors can be mitigated through the use of commonly used filtering criteria using a-priori knowledge of the disease (e.g. rare, Mendelian, monogenic), their presence indicates a likely underlying source of bias occurring in the study. No evidence of population stratification was observed when the variance across samples was summarized using principal components- largely discounting biases that might have arisen during the case-control matching process. A potential source of this high error rate may be due to the use of a multi-sample VQSR variant-quality call. In multi-sample mode, a VQSR filter call of "PASS" denotes that the variant call is likely correct in at least one sample- but does not insure it is of sufficient quality across all samples. Variants in which a subset of samples contain low quality calls may introduce false positives associations when those calls occur disproportionality in either the case or control groups. The use of sample-specific (rather than multi-sample) variant-quality calls may help target only those variants of sufficient quality across all samples, providing a higher quality set of variants for association testing in downstream analysis.

Even with the use of filtering criteria, sequencing errors that occur in PRs of the genome cause several false-discoveries to persist. While the variants in Table 1 included those related to errors in covering repeat sequences, examination of PRs in Table 2 revealed that the majority of errors were related to areas of high-GC content and the occurrences of segmental duplications. A comprehensive database integrating these regions provides a mechanism to identify and experimentally or statistically address these potential sources of error.

While the rational use of variant prioritization and/or filtering can enrich the pool of variants likely to be associated with disease, the concomitant reduction in detection sensitivity often increases the Type-II error rate. Filtering variants based on PRs would be particularly problematic in this regard, given that these occur throughout the genome and are not directly related to disease characteristics. Alternative strategies have used probabilistic models incorporating read-specific quality scores and/or sequencing training data in an effort to distinguish true variants from sequencing errors<sup>22-23</sup>. The outcome is typically a decision rule designed to improve false-positive or false-negative error rates in variant detection, or a scoring system in which variants can be differentially weighted in subsequent analysis. While these approaches are certainly improvements over simple filtering of variants, they do not explicitly model all sources of errors inherent in the sequence data itself, including areas of degeneracy, high GC content or areas of low-complexity.

Regardless of the strategy used to distinguish sequencing errors from true discoveries, the errors in the sequence data still exist. The greatest potential impact of a database of PRs is in the identification of areas in the genome that should be targeted for improved coverage—the result being reductions in sequencing error rates<sup>16</sup> regardless of the underlying cause. Improvements in coverage can have beneficial effects on sensitivity; and will improve specificity in large-scale studies where the error rates can differ across samples.

## References

1. Ng, S. B. et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
2. Ng, S. B. et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
3. Ng, S. B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
4. Hoischen, A. et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
5. Ku, C.-S., Naidoo, N. & Pawitan, Y. Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* **129**, 351–370 (2011).
6. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9546–9551 (2010).
7. Hedges, D. J. et al. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* **6**, e18595 (2011).
8. Clark, M. J. et al. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
9. Wang, W., Wei, Z., Lam, T.-W. & Wang, J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* **1**, 55 (2011).

10. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* **8**, e62856 (2013).
11. Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol.* (2011).
12. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
13. [http://wwwn.cdc.gov/clia/Resources/GetRM/pdf/CF\\_Characterized\\_Other.pdf](http://wwwn.cdc.gov/clia/Resources/GetRM/pdf/CF_Characterized_Other.pdf)
14. <http://ccr.coriell.org/Sections/BrowseCatalog/Diseases.aspx?a=C&coll=&PgId=3>
15. Lam, H. Y. K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).
16. West, J. *et al.*, Analytical Validity of Genome Sequencing Platforms for Medical Interpretation. Poster presentation at AGBT (2013).
17. Glusman, G. *et al.* Compressions in Human Reference Sequences Identified Using Genome Sequences From Multiple Pedigrees (8 families, 45 complete genomes). Poster presentation at AGBT (2011).
18. Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library
19. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
20. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 479–498 (2002).
21. Zhi, D. & Chen, R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. *PLoS ONE* **7**, e31358(2012).
22. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* **20**, 273–280 (2010).
23. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).

# BAGS OF WORDS MODELS OF EPITOPE SETS: HIV VIRAL LOAD REGRESSION WITH COUNTING GRIDS

ALESSANDRO PERINA, PIETRO LOVATO and NEBOJSA JOJIC\*

*Microsoft Research, One Way Microsoft, Redmond WA, 98052.*

*E-Mail: [jojic@microsoft.com](mailto:jojic@microsoft.com), Tel: +1 (425) 705-5865*

The immune system gathers evidence of the execution of various molecular processes, both foreign and the cells' own, as time- and space-varying sets of epitopes, small linear or conformational segments of the proteins involved in these processes. Epitopes do not have any obvious ordering in this scheme: The immune system simply sees these epitope sets as disordered "bags" of simple signatures based on whose contents the actions need to be decided. The immense landscape of possible bags of epitopes is shaped by the cellular pathways in various cells, as well as the characteristics of the internal sampling process that chooses and brings epitopes to cellular surface. As a consequence, upon the infection by the same pathogen, different individuals' cells present very different epitope sets. Modeling this landscape should thus be a key step in computational immunology. We show that among possible bag-of-words models, the counting grid is most fit for modeling cellular presentation. We describe each patient by a bag-of-peptides they are likely to present on the cellular surface. In regression tests, we found that compared to the state-of-the-art, counting grids explain more than twice as much of the log viral load variance in these patients. This is potentially a significant advancement in the field, given that a large part of the log viral load variance also depends on the infecting HIV strain, and that HIV polymorphisms themselves are known to strongly associate with HLA types, both effects beyond what is modeled here.

*Keywords:* Gene expression, Modeling host-pathogen interactions, Bag of Peptides

## 1. Introduction

The mammalian immune system consists of a number of interacting subsystems employing various infection clearing paths, with cellular presentation playing a central role in many of them. Most of the cells present a sample of peptides derived from cellular proteins as a means of advertising their states to the immune system. This facilitates globally coordinated action against viral infection.

The input to the cellular immune surveillance is illustrated in Fig.1. We show a simplified illustration of an infected cell which expresses both self (black) and viral (red) proteins (Fig.1A). Major histocompatibility complex (MHC) type I molecules bind to a small fraction of peptides from these proteins, created by proteasomal cleavage (Fig.1B). Inside these MHC complexes, the peptides are transported to the surface of the cell, where they may be detected by the cytotoxic T cells (CTL), which then may send self-destruct signals to the infected cell, thus stopping further infection (Fig.1C). Peptides that are a target of immune surveillance are often referred to as *epitopes*. As the sampled peptides do not appear in a particular spatial organization on the surface, the immune system effectively sees the infection as a bag of MHC molecules loaded with different viral peptides. Depending on the application, this representation may be further simplified into a *bag of viral peptides* (Fig.1D), under the assumption that the main effect of the MHC molecules is the peptide selection (e.g. choosing conserved vs non-conserved targets<sup>6</sup>).



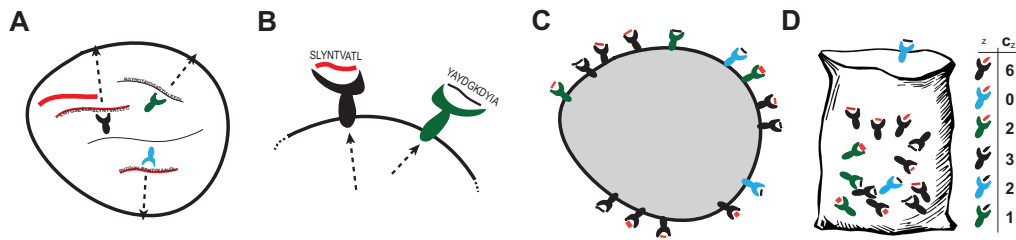


Fig. 1. Modeling immune surveillance input as a bag of words. **A** An Infected cell. **B** MHC binds to a fraction of peptides. **C** Sampled peptides appear without particular order on the cell surface. **D** A bag of peptides represents the relative counts  $c_z$  of the features seen on cellular surface.

This paper has a dual purpose: *i*) it argues for the new application of *bag of words* models,<sup>2,9</sup> which have already been successfully applied in various other areas of machine learning, as a set of tools for capturing correlations in the immune target abundances in humoral and cellular immune surveillance, and *ii*), it proposes a novel way of modeling bags of words which differs from PCA-like approaches not only in its treatment of observed epitope abundances as counts, but also moves away from the traditional componential structure towards a spatial embedding that captures smooth changes in cellular presentation.

In the experimental section, we restrict to the analysis of the links between the HIV viral load and the patients HLA types, leading to significant improvement with respect to the state of the art. Beyond the particular application tackled here, a good probability model of the epitope co-presentation has several direct applications, from correcting association studies, to detecting patients or populations that are likely to react similarly to an infection, to the rational vaccine design.

**Related Work** Explaining the differences in viral loads in different HIV patients has received a lot of attention from the HIV community, ever since the early longitudinal studies showed that changes in viral load occur in synchrony with the emergence of new HLA class I epitopes in immune assays.<sup>4</sup>

However, in case of the highly polymorphic HIV, a handful of epitopes usually fail to control the infection, and so researchers turned to population studies in search for optimal immune targets. Early studies failed to detect significant links between patients HLA types and viral load as the straightforward statistical approaches could not handle small dataset sizes (typically around 200 patients or less). But the evidence of HLA pressure on HIV was recognized in strong associations between viral mutations and patients' HLA types.<sup>5</sup> Viral load is highly variable and it may depend on numerous factors, such as gender, age, prior infections and general health of the individual. Thus it seemed likely that only the strongest MHC-driven effects would be visible through the noise. Still, any statistically significant result has been seen as having important consequences to HIV research. Eventually, larger cohorts allowed researchers to detect links between HLA types and viral load. Certain HLA B types, esp. B57 and B5801 were found to strongly associate with low viral load in a cohort of over 700 HIV patients in southern Africa.<sup>7</sup> In these studies, despite the statistically strong associations, the viral load in B57 or B5801 positive and negative patients still had such large variance that each of these HLA types alone could only explain less than 2% of the total log viral load variance in the

Table 1. The percentage of viral load (VL) explained in literature as the square of the Pearson's linear correlation coefficient (See Tab.2)

Ref.	Major Result
5	VL considered too noisy. Associations with mutations found
7	1-2% of VL variance explained through individual allele association
6	4% of VL variance explained through by targeting efficiency
10	4.3%-9% of VL variance explained by combinations of epitopes
<b>This Paper</b>	Up to 13.5% of VL variance explained by embedding into Counting Grids

population.

Multiple hypothesis testing issues and linkage disequilibrium among HLA loci complicated this research and the employed straightforward statistical approach did not present obvious ways to move from singular features (such as a binary labeling of patients as having B57 or not) to combinations of features that would provide higher explanatory power. However, by analyzing the tendency of the HLA molecules to bind to conserved targets in the HIV, it is possible to create a patient score (dubbed targeting efficiency) that captures binding characteristics of all 6 HLA molecules relative to HIV proteins.<sup>6</sup> At least on one cohort,<sup>5</sup> targeting efficiency explained a little less than 4% of the log viral load variance<sup>a</sup>. On the same cohort, another recent method deals with multiple features and their correlations, the *correlation sifting*,<sup>10</sup> explaining 4.3% of the log viral load variance by patients' HLA types. We show here that the bag of words models<sup>3,9</sup> lead to even better regression to viral load. This is especially the case for the new counting grid model<sup>9</sup> that efficiently captures correlations in cellular presentation by embedding patients in a grid, where the embedding coordinates can be used to explain 13.5% of log viral load variance, more than twice the current state of the art.

To put these numbers into perspective, it is important to make two observations. First, even weak signals, had the tendency to move the entire field,<sup>5,7</sup> as valuable characteristics of the interaction between HIV and the host immune system were revealed, informing both the research on HIV drugs and the research on HIV vaccine. Second, in addition to high variation of the viral load due to factors that relate to age and general health, it is known that the set point viral load depends strongly on the infecting strain,<sup>8</sup> and as HIV was found to mutate in its reactions to HLA presentation, this variation in fitness in the infecting strains may itself be due to the HLA pressure from previous hosts. Thus the increase in explanatory power of HLA types from around 4% of the log viral load to around 13.5% is potentially of great importance. Further analysis in selected combinations of features in the counting grid may lead to further advances in understanding the evolutionary arms race between HLA and the human immune system.

## 2. Bag of words models

In machine learning research, data samples are often represented as bags of features without a particular order. This choice is typically motivated by the difficulty or computational effi-

<sup>a</sup>Note again that the original analysis based on individual alleles failed to detect significant links with viral load there

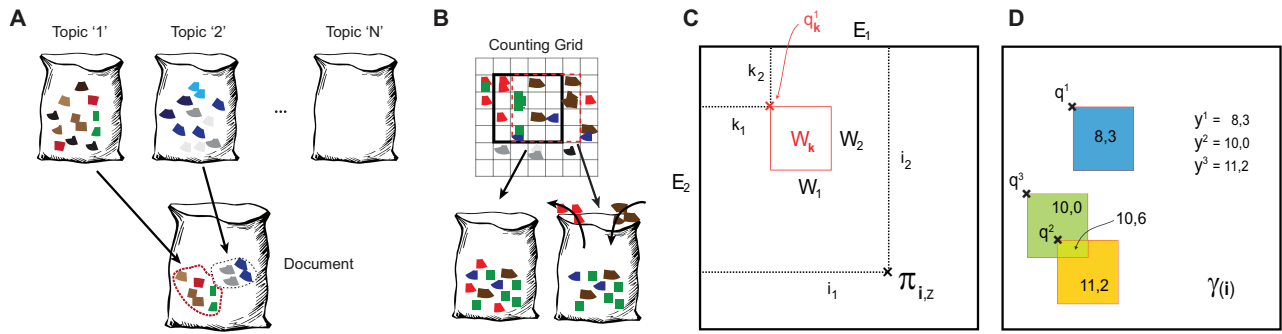


Fig. 2. Capturing dependencies in bags of words.

ciency of modeling the feature structure. Computational biology is abundant with examples of data where the structure is truly unknown, rather than just sacrificed for computational efficiency: for example, a gene expression array has been modeled as a bag of genes with expression levels simply corresponding to counts because most of the time little is known about the cellular pathways that employ these genes.<sup>11–14</sup> Without such knowledge there is no clear gene ordering. But biology is also abundant with situations where the raw data of interest actually has no (known or unknown) structure. In particular, in this paper we develop models of the sets of immune system targets.

Topic models<sup>1,2</sup> were introduced by the text analysis community and have been particularly successful in representing text documents. These simplified models of text assume that a text document has been generated simply by mixing words from a subset of possible topics. In typical applications, the number of possible topics is large, and these topics are inferred from the data by analyzing word co-occurrence patterns, and so the topic scope can vary from very narrow to quite broad, e.g., from near homonyms, to words found in most stories on US politics. An individual document is assumed to use only a fraction of all possible topics, and so the resulting bags of words will exhibit strong co-occurrence patterns: when the president is mentioned, so is the congress, as both appear in the same topic.

These models can be used in other domains by simply replacing words with some other set of features of interest. In bioinformatics, for example, words are replaced by genes and their counts by expression levels<sup>3,12</sup> to model microarray experiments. Visual descriptors are extracted from salient points of brain images and clustered into “visual words” replacing traditional words in bags of words and these representations were then used to classify schizophrenic patients from controls.<sup>18</sup> Peaks in nuclear magnetic resonance (NMR) spectrometry were also clustered and used as words.<sup>19</sup> Finally, protein sequences are sometimes broken into segments or *fragments*, which serve as words for comparing protein structures.<sup>20</sup>

Among topic models one of the best known is the Latent Dirichlet Allocation (LDA).<sup>2</sup> To formally define this model, we will index possible words (features) by  $z$  and denote the set of observed word (or feature) counts in the  $t$ -th bag of words by  $\{c_z^t\}$ . The latent (hidden) variables describe the choice of topics indexed by  $k$ . The choice of topics follows a distribution  $p(k|\theta) = \theta_k$ , and each topic has its own distribution over all the words  $p(z|k, \beta) = \beta_z|k$ . The

vector that depicts the topic distribution for one document  $\theta$  is sampled from a Dirichlet distribution with parameters  $\alpha$ . The following probability of generating a particular document is induced by this simple generative process (after picking the topic distribution  $\theta$ , pick a topic, then pick a word from the topic, then pick a topic and a word from it again and again till all the words in the document are generated):

$$p(\{c_z^t\}|\alpha, \beta) = \int p(\theta|\alpha) \cdot \prod_z \left( \sum_k (p(z|k, \beta) \cdot p(k|\theta))^{c_z^t} \right) d\theta \quad (1)$$

The model parameters are estimated based on a training set so as to maximize the product of probabilities of all training documents. The topic proportions  $\theta$  for individual documents can be used as a compact representation of the bag of words that discards the superfluous aspects of the data. For example, the HIV viral load can be regressed directly to these hidden variables in patient cohorts that are too small for the full representation of the viral presentation. Modeling cellular peptide presentation as a mixture of topics can capture some of the presentation patterns discussed above. Upon model fitting, the topics may correspond to individual MHC molecules that are more frequent in the patient cohort, or entire families of MHC types that have similar presentation (sometimes referred to MHC supertypes). In this case, all viral peptides would be indexed by  $z$ , and the topic probability distribution would reflect the probabilities of binding of a particular MHC (super)type to these different peptides. Some topics may also capture the HIV clade structure as mutations in each clade alter the MHC binding patterns.

### ***Estimating bags of peptides for individual HIV patients***

The concentration of any viral peptide on the cellular surface depends on the source protein's expression level. But different HIV proteins are expressed at different times in the HIV's infection and reproduction cycle. Instead of trying to estimate appropriate weighting factors, we simply considered each of the HIV proteins in isolation in our experiments.

As most epitopes are of length 9, for each analyzed protein we created a vocabulary of all *9-mers* that exist in this protein, indexed by  $z$ . Each human host has up to 6 different MHC I molecules (two from each of the three ancient duplicated and highly polymorphic loci A, B, C in the HLA region). In addition, in our experiments we dealt with a cohort in which we had the HLA types for each patient and we had access to an MHC I - peptide complex prediction algorithm that can estimate the *binding energy*  $E_b(z, m)$  for each of the peptides  $z$  and the different patient's HLA molecules indexed by  $m$ .<sup>21</sup> Finally, we also used a *cleavage energy*<sup>22</sup> estimate  $E_c(z)$  and turned the total energy into a count (concentration) as follows

$$c_z = e^{-E_c(z) - \min_m [E_b(z, m)]} \quad (2)$$

In a simplified model, the individual's immune system sees this variation in peptide counts (with many counts close to zero), and thus needs to recognize a virus not as a whole but as a set of disordered viral peptides.

Estimation of surface peptide (relative) counts could use any number of other epitope prediction techniques recently developed in computational biology.<sup>23</sup> Here we used the adaptive

double threading technique,<sup>21</sup> as it provides prediction for arbitrary MHC types simply defined by their protein sequence. NET MHC Pan<sup>24</sup> predictors provides similar functionality.

The counts  $c_z$  are not independent. The MHC system, as well as viral mutations, create links among the abundances of different viral peptides in the observed bag. Each MHC molecule has its binding preferences that lead to selection of only one of a hundred to a thousand of peptides. The human leukocyte antigen (HLA) region (human MHC) is the most polymorphic region of the human genome. As a result, two patients infected by the same virus, e.g. HIV, are highly unlikely to have the exact same MHC molecules. Each of their molecules will select specific targets from HIV proteins, and the patients' sets of immune targets will likely overlap only partially. The variation of the HIV epitope sets found in different patients exhibits strong co-occurrence patterns where a high count of one peptide often implies inclusion of several others, as they are all good binders to a particular MHC allele (families of different alleles can also share binding preferences). These links in epitope presentations are further expanded by weak linkage disequilibrium among MHC types as well as viral adaptation, which is itself correlated across sequence sites.

This all means that good models of bags of epitopes that constitute the immune surveillance input need to capture these correlations and this is precisely what the probability models of bags of words were meant to do for text documents.

### 3. The Counting Grid model

In the counting grid model, individual distributions over words are arranged on a grid (see Fig.2). Each of these distributions is relatively tight, with only a few features having significant probability. To generate a bag of words, instead of mixing topics, it is assumed simply that a window into the grid is opened, and the feature counts in the cells inside the window are combined to create the appropriate words in appropriate abundance. The window floating over the grid captures well variation in certain types of documents where we can see slow evolution of the topics, where certain words are dropped and new ones introduced: think for example to news stories over time, as interest in certain news slowly vanes in favor of new ones. Although traditional topics have been embedded in time or space and made slowly varying in certain directions, these variations do not quite capture the simple constraints present in CG models where a small window shift in the grid simply drops certain words and adds new ones. Furthermore, the counting grids are learned from the data for which the embedding in time or space is *not available*; this is the case for epitope bags. As we will show shortly, counting grids for this data can never the less be produced by iteratively estimating the grid distributions and inferring the mapping of the data to appropriate windows in it, thus resulting in the embedding of the data to a grid.

Formally, the basic counting grid  $\pi_{\mathbf{i},z}$  is a set of normalized counts of words / features indexed by  $z$  on the  $D$ -dimensional discrete grid indexed by  $\mathbf{i} = (i_1, \dots, i_D)$  where each  $i_d \in [1 \dots E_d]$  and  $\mathbf{E} = (E_1, \dots, E_D)$  describes the extent of the counting grid. Since  $\pi$  is a grid of distributions,  $\sum_z \pi_{\mathbf{i},z} = 1$  everywhere on the grid. A given bag of words/features, represented by counts  $\{c_z\}$  is assumed to follow a count distribution found somewhere in the counting grid. In particular, using windows of dimensions  $\mathbf{W} = [W_1, \dots, W_D]$ , each bag can be generated by

first averaging all counts in the hypercube window  $W_{\mathbf{k}} = [\mathbf{k} \dots \mathbf{k} + \mathbf{W}]$  starting at  $D$ -dimensional grid location  $\mathbf{k}$  and extending in each direction  $d$  by  $W_d$  grid positions to form the histogram  $h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z}$ , and then generating a set of features in the bag. In other words, the position of the window  $\mathbf{k}$  in the grid is a latent variable given which the probability of the bag of features  $\{c_z\}$  is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \frac{1}{\prod_d W_d} \prod_z \left( \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z} \right)^{c_z} \quad (3)$$

Relaxing the terminology, we will refer to  $\mathbf{E}$  and  $\mathbf{W}$  respectively as the counting grid and the window size. We will also often refer to the ratio of the window volumes,  $\kappa$ , as a capacity of the model in terms of an *equivalent number of topics*, as this is how many non-overlapping windows can be fit onto the grid. Fine variation achievable by moving the windows in between any two close by but non-overlapping windows is useful if we expect such smooth thematic shifts to occur in the data, and we illustrate in our experiments that indeed they do. Finally, with  $W_{\mathbf{k}}$  we indicate the particular window placed at location  $\mathbf{k}$  (see Fig.2C). To learn a counting grid we need to maximize the likelihood of the data:

$$\log P = \sum_t \log \left( \sum_{\mathbf{k}} \prod_z (h_{\mathbf{k},z}^{c_z^t}) \right) \quad (4)$$

The sum over the latent variables  $\mathbf{k}$  makes it difficult to perform assignment to the latent variables while also estimating the model parameters. The problem is solved by employing an iterative variational EM procedure. The E step aligns each bag of features  $\{c_z^t\}$  to grid windows, to match the bag's histograms. In this way we compute the posterior distribution  $q_{\mathbf{k}}^t$  over all windows  $\mathbf{k}$  so that a better match between  $\{c_z^t\}$  and  $h_{\mathbf{k},z}$  across all features  $z$  yields a higher value for the match. In other words,  $q_{\mathbf{k}}^t$  is probabilistic mapping of the  $t$ -th bag to the grid windows  $\mathbf{k}$ . This mapping is usually peaky, i.e., each bag tends to map to a few nearby locations in the grid. In the M-step we re-estimate the counting grid so that these same histogram matches are even better. To avoid severe local minima it is important to consider the counting grid as a torus, and perform all windowing operation accordingly. For details on the learning algorithm and on its efficiency see the original CG paper.<sup>9</sup>

### ***Regression of continuous values***

Once a CG is learned, we show here how one may embed continuous values  $y^t$  on the grid (e.g., HIV viral load). This is achieved using the posterior probabilities  $q_{\mathbf{k}}^t$  for each bag already inferred and embedding the corresponding viral load inside the entire mapped window(s), and then averaging all overlapping windows (Fig.2D), which is similar to how M step re-estimates the distributions  $\pi$ :

$$\gamma(\mathbf{i}) = \frac{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q_{\mathbf{k}}^t \cdot y^t}{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q_{\mathbf{k}}^t} \quad (5)$$

The function  $\gamma$  can then be used for regression, in what is essentially a nearest-neighbor strategy: when a new data point is embedded based on its bag of words, the target is simply read out from  $\gamma$ , which is dominated by the training points which were mapped in the same region.

In Fig.4A we show a couple of  $\gamma$ s, estimated from the dataset we used in the experiments. The window  $\mathbf{W}$  is shown with a dotted line in the figure.

#### 4. Experiments

In this section we first discuss what aspects of the epitope bags the counting grids may capture. Then we show that counting grids outperform not only traditional bag of words models, which have previously not been applied to this task, but also the state of the art in biomedical and computational biology literature<sup>5-7,10</sup> on analysis of the links between the HIV viral load and the patients HLA types (see Tab.1).

##### **Types of correlations in epitope bags that can be captured with counting grids**

There are reasons why a counting grid model may be a more appropriate model of variation in epitope bags and perhaps more generally in many computational biology applications. These reasons have to do with the manner in which biological entities interact and adapt to each other leading to patterns of slow evolution characterized by genetic drift, local co-adaptation, as well as punctuated equilibrium. In case of cellular presentation, for example, millions of years of evolution created certain typical variants of MHC as well as minor variation on each of these major types. These variations are at least in part due to the interaction with viruses,<sup>6</sup> and similarly the genetic variation in viruses reflect some of this evolutionary arms race, too. Thus, the HIV clade constraints, as well as MHC binding characteristics may be so interwoven that a rigid view of cellular presentation as a mix of a small number of topics may be inappropriate. In the counting grid, the major variants of cellular presentation can be modeled as far away windows, while minor variations would be captured by slight window shifts in certain regions of the grid. To illustrate this we analyzed the cellular presentation of HIV patients from the Western Australia cohort.<sup>5</sup> We represented each patient's cellular presentation by a set of 492 counts over that many 9-long peptides from the Gag protein, previously found to be targeted by the immune system. The counts were calculated based on the patients MHC class I types (or HLA types, as they are called in humans) and the HLA-peptide binding estimation procedure discussed in Sec.2. This provides us with bags of peptides (*BoP*, counts over the 492 words) that represent GAG in different patients. We used the same process for two more proteins, POL and VPR, resulting in counts matrices of respectively  $88 \times 135$  and  $939 \times 118$  words  $\times$  samples. We analyzed only the clade B infected patients.

**Cellular presentation of viral peptides and viral load** As the immune pressure depends on cellular presentation, the variation in cellular presentation across patients is expected to reflect on the variation in viral load, at least to some extent.<sup>5,6</sup> Viral load is expected to depend on the cellular presentation for various reasons. If the targeted peptides are conserved, this indicates inability of the virus to escape immune pressure. Even binding to some relatively variable peptides may lead to good outcomes for the patient (low viral load), as long as the CTLs can crossreact effectively across the peptide variants. In addition, there is a possibility that additional qualities of the peptides render some immune responses more effective than the others, or that certain immune responses trigger different viral behaviors. In bio-medical

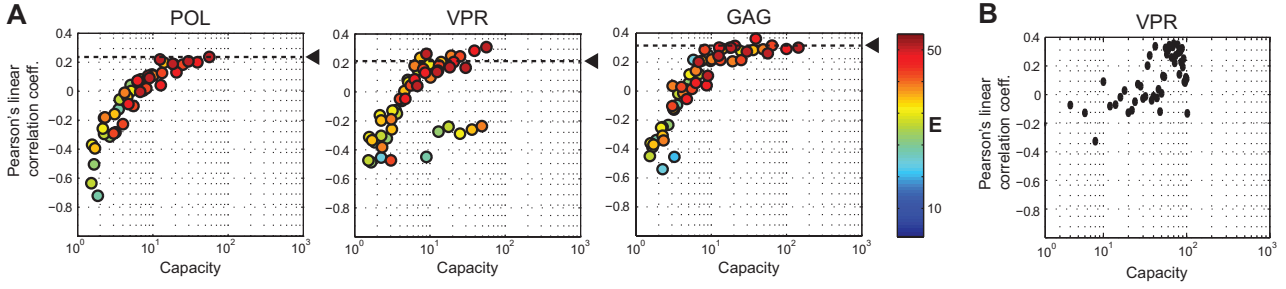


Fig. 3. HIV viral load regression. The variation of the correlation factor  $\rho$  for CG and LDA models of different complexities. Color code is used represent the square CG size  $E$  as a single capacity can be obtained with different  $E/W$  combinations.

literature, analysis of this type of data targeted individual peptides and the discovery of those peptides that have significant association with viral load. However, these results do not explain nearly as much of viral load variance as what follows.

As general procedure, we first trained a CG using the bags-of-peptides  $c_z$  but *without* using the regression targets  $y^t$  (log viral load). Then, in a leave-out-out fashion, we held out a sample  $\hat{t}$  and estimated the regression function  $\gamma$  (see Eq. 5, with  $t \neq \hat{t}$ ) using all the other epitope bag/viral load pairs, and finally, read out  $\gamma$  in the appropriate (probabilistic) location  $q_k^{\hat{t}}$  to obtain the viral load prediction for  $\hat{t}$  sample as  $y_{CG}^{\hat{t}} = \sum_k q_k^{\hat{t}} \cdot \gamma(k)$ . Once we computed the estimated regression target for all the samples, we computed  $\rho$ , the pairwise correlation coefficient between the true and the estimated viral load, comparing CGs with LDA,<sup>3</sup> and a technique based on phlogenetic trees<sup>15</sup> meant to established how much can the viral laod be predicted simply from the patient's dominant HIV sequence, as different strains may vary in fitness. We considered counting grids of various complexities  $E = [12, 15, 18, 21, 25, 30, 35, 40, 50]$  and  $W = [2, 3, 4, \dots]$ . We tested only the combinations with capacity  $\kappa$  between 1.5 and  $T/2$ , where  $T$  is the number of samples available.

Rogers' LDA adaptation,<sup>3</sup> LPD originally designed for modeling microarray data was evaluated in a similar fashion. We learned as single model (without using the target) and we predicted the viral load for the left out sample using linear regression based on the topic proportions  $\theta$ .

To compare with a sequence-based regression, we used the maximum likelihood approach<sup>15</sup> to estiamte a phylogentic tree for all patients' HIV sequences. Few parameters have to be tuned when computing such trees: In our experiments, we pick as a rate substitution matrix the WAG model,<sup>16</sup> and we allowed for rate variations across sites, setting 4 discrete gamma categories.<sup>17</sup> To predict the viral load  $\hat{y}$  for a test sequence  $x$  using the estimated tree, we detected the training sequences that lie near by in the tree and averaged their viral loads accdordign to their distance. If  $t$  indexes the training sequences  $x_t$  and their associated viral load value  $y_t$

$$\hat{y} = \sum_t e^{-C \cdot \text{dist}(x, x_t)} \cdot y_t \quad (6)$$

The parameter  $C$  has been found with crossvalidation on the training set. Fig.3, summarizes the performance of CG and LDA across a range of capacities  $\kappa$  for CGs and the number of topics  $K$  for LDA. LDA and CGs reach similar results of POL and VPR, while CGs have a clear advantage on GAG. It is important to note that for the Counting Grids, the correlation



factor varies much more regularly with the capacity  $\kappa$ , since this indicates that the complexity can be chosen on the training set through crossvalidation, which then allow us to properly calculate the percent of viral load explainable by the model. For each protein, we performed leave-one-out crossevaluation on the training set, to pick the best model complexity ( $\mathbf{E}/\mathbf{W}$  for Counting Grids, or the number of topics  $K$  for LDA) and we compared the results with the tree regression discussed above.

In leave-one-out experiments, the training set was each time used as a full set for another set of leave-one-out experiments on training data alone, plotting the graphs as above, and picking the best complexity. Then for the test sample we predicted the viral load using this best complexity. It is important to note that in this scheme *i*) the viral load of different patients can in principle be predicted using different complexities, and *ii*) the test sample does not contaminate the prediction model in any way. Results are shown in Tab.2. For Latent Dirichlet Allocation, this process failed and we could not obtain statistically significant results because of severe overtraining issues.

Finally, we also combined CG predictions with the idea of regressing the reconstruction error  $E_z^t = \tilde{c}_z^t - R_z^t$  on residual viral load  $y_{RED}^t = y^t - y_{CG}^t$ ,<sup>10</sup> where  $y_{CG}^t$  is the viral load prediction using the counting grid, and  $\tilde{c}_z^t$  the normalized feature count. We used a regularized linear regression with  $L_1$  norm using as before leave-one-out crossevaluation to choose the best model complexity. We computed the correlation factor  $\rho$ , setting final viral load prediction to be equal to the sum of  $y_{CG}^t$  and the prediction of  $y_{RED}^t$ . The idea here is that the deviation from the norm may be detecting viral adaptation and can predict further the modulation of viral fitness. As can be seen in Tab.2, column  $CGs \rightarrow^{10}$ , this improved the performance in all the cases.

Interestingly, the model complexities chosen by each round of leave-one-out, though they could in principle be different for each patient, did not in fact differ that much. Regardless of the protein considered, for more than 89% of the data points the same complexity was typically chosen, as reported in the last column of Tab. 2.

Table 2. Pearson's linear correlation (after crossevaluation where applicable). Crossevaluation for LDA was found not statistically significant (NS) for **GAG** and **POL**. The last column reports the most common CG's complexity chosen in the rounds of leave-one-out crossevaluation.

<b>Protein</b>	CGs $\rho$	CGs $\rightarrow$ 10 $\rho$	Trees $\rho$	LDA $\rho$	Ridge Regr. $\rho$	Complexity Chosen
<b>GAG</b>	0.3301	0.3674	0.3519	NS	0.1835	[ <b>30,5</b> ] - 89%
<b>VPR</b>	0.2011	0.2546	0.1061	0.1202	NS	[ <b>50,8</b> ] - 94%
<b>POL</b>	0.2338	0.2443	0.1812	NS	NS	[ <b>40,11</b> ] - 97%

The medical literature has other results obtained by analyzing GAG protein as shown in Tab.1, but the results reported here outperform all these methods, too.

We have one final note on the embedding function  $\gamma$ . The bags of peptides are mapped to the counting grid iteratively as the grid is estimated as to best model the bags, but the regression target, the viral load, was not used during the learning of CGs or LDA models. However, the inferred mapping after each iteration can be used to visualize how the embedded

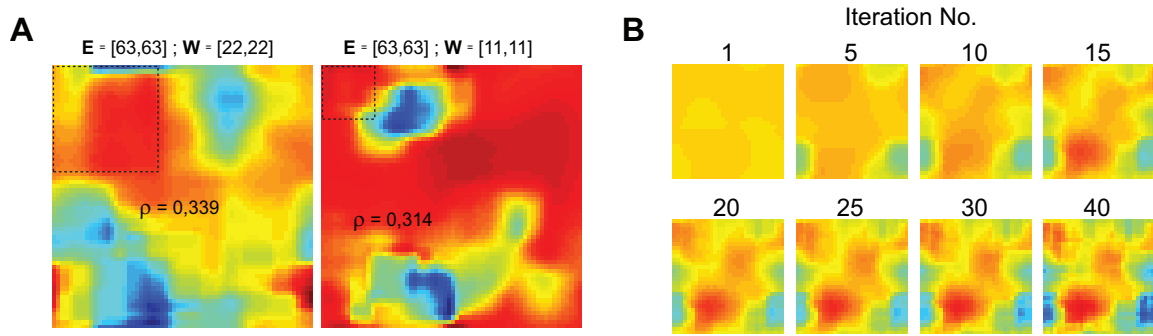


Fig. 4. **A** HIV viral load embedding in the 2D. The window is shown with a dotted line in the figure. **B** Evolution of the viral load across the iterations.

viral load  $\gamma$  evolves. This is illustrated in Fig.4B for a model of complexity  $\mathbf{E} = [30 \times 30]$ ,  $\mathbf{W} = [8 \times 8]$ . The emergence of areas of high (red) and low (blue) viral load indicates that as the structure in the cellular presentation is discovered, it does indeed reflect the variation in viral load.

## 5. Conclusions

We propose the use of bag of words models to capture cellular presentation, and more generally the view that the immune system has of the invading pathogens. Furthermore, we demonstrate that the newest of these models, the counting grid, seems to be especially well suited to this task, providing stronger predictions than what can be found in bio-medical literature.

It remains to be understood exactly why CGs exhibit such a strong advantage over topic models (LDA). One intuitive explanation is that the slow smooth variations in count data that can be captured in counting grids better represent the dependencies that were produced by millions of years of coevolution between the HLA system and various invading pathogens.<sup>6</sup> This process involved numerous mixing of both the immune types and the viral strains, and may have produced the sort of thematic shifts in cellular representation that CGs are designed to represent. A more speculative possibility is that the immune system, through some unknown mechanism, collates the reports from circulating CTLs into an immune memory of a similar structure, though this summarization would obviously be performed over different invading pathogens in one patient, while our CGs depict one virus in a population of patients. Our experiments showed that cellular presentation of the Gag protein explains more than 13.5% of the log viral load. Although viral load varies dramatically across patients for a variety of reasons, e.g. gender, previous exposures to related viruses, etc., detection of statistically significant links between cellular presentation and viral load is expected to have important consequences to vaccine research.<sup>7</sup>

## References

1. S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, Latent Semnatical Indexing *Journal of the American Society for Information Science* **41**, 391 (1990)
2. D. Blei, A. Ng and M. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* **3**, pp.993 (2003)
3. S. Rogers, M. Girolami, C. Campbell, R. Breitling, The latent process decomposition of cdna

- microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**, Vol. 2, 143–156 (2005)
4. A. McMichael *et al.*, Cellular immune responses to HIV. *Nature* **410**, 980-987 (2001)
  5. C. Moore *et al.*, Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science* **296**, 436 (2002)
  6. T. Hertz *et al.*, Mapping the Landscape of Host-Pathogen Coevolution: HLA Class I Binding and Its Relationship with Evolutionary Conservation in Human and Viral Proteins. *Journal of Virology* **85**, 1310 (2010)
  7. P. Kiepiela *et al.*, Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769 (2004)
  8. S. Alizon *et al.*, Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens* **6:9** (2010)
  9. N. Jojic, and A. Perina, Multidimensional Counting Grids: inferring words order from disordered bags of words. *Proceedings of Uncertainty in Artificial Intelligence* (2011)
  10. J. Huang and N. Jojic, Variable selection by correlation sifting. *International Conference on Research in Computational Molecular Biology* (2011)
  11. A. Perina, P. Lovato, V. Murino, and M. Bicego, Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray. *Proceedings of the IAPR international conference on Pattern recognition in bioinformatics*, (2010)
  12. M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, V. Murino Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, **9**, No. 6, pp. 1831-1836 (2012)
  13. F. Terrence, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics Journal*, **16**, No. 10, pp. 906-914, (2000)
  14. I. Guyon, J. Weston, S. Barnhill, V. Vapnik Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, **46**, No. 1-3, pp. 389-422 (2002)
  15. K. Tamura *et al.*, MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods, *Mol. Biol. Evol.*, **28**, (2011)
  16. S. Whelan and N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Mol. Biol. Evol.*, **18**, (2001)
  17. Z. Yang Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods *J. Mol. Evol.*, **39**, (1994)
  18. U. Castellani, A. Perina, V. Murino, M. Bellani, G. Rambaldelli, M. Tansella, P. Brambilla Brain morphometry by probabilistic latent semantic analysis *International Conference on Medical Image Computing and Computer Assisted Intervention* (2010)
  19. G. Brelstaff, M. Bicego, N. Culeddu, M. Chessa, Bag of Peaks: interpretation of NMR spectrometry *Bioinformatics Journal*, **25**, Vol. 5, pp. 258-274 (2009)
  20. I. Budowski-Tala, Y. Novb, R. Kolodnya, FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately *Proc. Natl. Acad. Sci.* **107**, pp. 3481-3486 (2010)
  21. N. Jojic, M. Reyes-Gomez, D. Heckerman, C.M. Kadie, O. Schueler-Furman Learning MHC I - peptide binding *Bioinformatics* **22**, Vol.14 pp.227-235 2006
  22. H. Nielsen, J. Engelbrecht, S. Brunak and G. Von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 1-6, 1997.
  23. G. Lan Zhang, V. Brusica *et al.* Machine Learning Competition in Immunology - Prediction of HLA class I molecules *J Immunol Methods*. 2011 Nov 30;374(1-2):1-4.
  24. I. Hoof *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans *Immunogenetics*. 2009 January; 61(1): 1-13.

# JOINT ASSOCIATION DISCOVERY AND DIAGNOSIS OF ALZHEIMER'S DISEASE BY SUPERVISED HETEROGENEOUS MULTIVIEW LEARNING

SHANDIAN ZHE<sup>1</sup>, ZENGLIN XU<sup>1</sup>, YUAN QI<sup>1,2</sup>, PENG YU<sup>3</sup>, FOR THE ADNI\*

<sup>1</sup>*Department of Computer Science, Purdue University,*

<sup>2</sup>*Department of Statistics, Purdue University,  
West Lafayette, IN 47907, USA*

*E-mail: {szhe,xu218,alanqi}@purdue.edu*

<sup>3</sup>*Eli Lilly and Company, Indianapolis, IN 46225, USA*

*E-mail: yu\_peng\_py@lilly.com*

A key step for Alzheimer's disease (AD) study is to identify associations between genetic variations and intermediate phenotypes (*e.g.*, brain structures). At the same time, it is crucial to develop a noninvasive means for AD diagnosis. Although these two tasks—association discovery and disease diagnosis—have been treated separately by a variety of approaches, they are tightly coupled due to their common biological basis. We hypothesize that the two tasks can potentially benefit each other by a joint analysis, because (i) the association study discovers correlated biomarkers from different data sources, which may help improve diagnosis accuracy, and (ii) the disease status may help identify *disease-sensitive* associations between genetic variations and MRI features. Based on this hypothesis, we present a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels based on Gaussian process ordinal regression; in return, the disease status is used to guide the discovery of relationships between the data sources. The sparse projection matrices not only reveal the associations but also select groups of biomarkers related to AD. To learn the model from data, we develop an efficient variational expectation maximization algorithm. Simulation results demonstrate that our approach achieves higher accuracy in both predicting ordinal labels and discovering associations between data sources than alternative methods. We apply our approach to an imaging genetics dataset of AD. Our joint analysis approach not only identifies meaningful and interesting associations between genetic variations, brain structures, and AD status, but also achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

*Keywords:* disease diagnosis, Alzheimer's disease, genetic variations, brain structures, multiview learning, ordinal regression.

## 1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder associated with aging. Although it accounts for 60-80% of age-related dementia cases, currently there is no cure for AD and its underlying mechanism remain elusive. To study AD mechanism, a crucial step is to identify associations between genetic variations and intermediate phenotypes (*e.g.*, endophenotypical traits). In other words, we want to discover cross linkages between genetic risk factors based on

---

\*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.ucla.edu](http://adni.loni.ucla.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.ucla.edu/wp-content/uploads/how\\_to\\_apply/ADNIAcknowledgement\\_List.pdf](http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNIAcknowledgement_List.pdf)

genomic data—such as single nucleotide polymorphisms (SNPs)—and indicative intermediate phenotypes—such as cortical thickness of different brain regions (based on magnetic resonance imaging (MRI)). This identification can help us locate a subset of polymorphisms which may have functional consequences on brain structures. Although GWAS studies have been applied to AD studies,<sup>1,2</sup> the association study between genetic variations and multiple intermediate phenotypes is still relatively scarce for AD. A similar task arises for expression quantitative trait locus (eQTL) analysis, where canonical correlation analysis (CCA) and its extensions<sup>3–6</sup> have been widely applied. Meanwhile, it has become increasingly important to develop a noninvasive means for AD diagnosis based on various biomarkers, including both genetic variations and MRI features. Because many of these biomarkers are irrelevant to the diagnosis, sparse models are needed to identify the relevant ones. For disease diagnosis, popular sparse models include lasso,<sup>7</sup> elastic net,<sup>8</sup> and automatic relevance determination.<sup>9</sup> Here we treat genotypes or intermediate phenotypes as biomarkers and the disease status as the response in a linear regression or classification setting. Non-zero regression or classification weights in our estimation indicate relevant biomarkers for the disease.<sup>10,11</sup>

Although these two tasks—association discovery and disease diagnosis—have been addressed separately in the previous works, they are closely related—due to their common underlying biological basis—and can potentially benefit each other by a joint analysis. To harness the natural synergy between the two tasks, we propose a new Bayesian approach that integrates multiview learning for association discovery with sparse ordinal regression for disease diagnosis. In the new approach, genetic variations and phenotypical traits are generated from common *latent* features based on separate sparse projection matrices and the common latent features are used to predict the disease status based on Gaussian process ordinal regression (See Section 2). To enforce sparsity in projection matrices, we assign spike and slab priors<sup>12</sup> over them; these priors have been shown to be more effective than  $l_1$  penalty to learn sparse projection matrices.<sup>13,14</sup> The sparse projection matrices not only reveal critical interactions between the different data sources but also identify *groups* of biomarkers in data relevant to disease status. Finding groups of biomarkers can avoid over-sparsification (*i.e.*, selecting one instead of multiple correlated features), thus boosting the accuracy for disease diagnosis. It can also help provide a better biological understanding because these groups may form biologically meaningful units (*e.g.*, pathways). Meanwhile, via its direct connection to the latent features, the disease status influences the estimation of the projection matrices. Hence we name this new method Supervised Heterogeneous Multiview Learning (SHML). In addition to enjoying the benefit of integrating the related tasks, two features of our model distinguish it from previous approaches:

- There is a severity order for AD, from being normal to mild cognitive impairment (MCI) and then to AD; and our ordinal regression component captures the AD severity order. Alternative sparse models, by contrast, use classification or regression likelihoods and do not consider the order of disease severity.
- The data are heterogeneous: SNPs values are discrete (or ordinal) and the imaging features are continuous. While popular CCA-type methods treat both of them as continuous data, our model captures the heterogeneous nature of the data.

To learn the model from data, we develop a variational Bayesian expectation maximization (VB-EM) approach (See Section 3). Maximizing this estimate enables us to automatically choose a suitable dimension for the latent features in a principled Bayesian framework.

In Section 4, we test our approach SHML on both synthetic and real datasets. On synthetic data, SHML achieves both higher estimation accuracy in recovering true associations between different views and higher prediction accuracy than alternative state-of-the-art methods. We then apply SHML to an AD study. SHML achieved highest prediction accuracy among all competing methods and yielded biologically meaningful relationships between genetic variations, brain atrophy, and the disease status.

## 2. Model

First, let us describe the data. We assume there are two heterogeneous data sources: one contains continuous data – for example, MRI features – and one discrete ordinal data – for instance, SNPs. Given data from  $n$  subjects,  $p$  continuous features and  $q$  discrete features, we denote the continuous data by a  $p \times n$  matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ , the discrete ordinal data by a  $q \times n$  matrix  $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$ , and the labels (*i.e.*, the disease status) by a  $n \times 1$  vector  $\mathbf{y} = [y_1, \dots, y_n]^\top$ . For the AD study, we let  $y_i = 0, 1$ , and 2 if the  $i$ -th subject is in the normal, MCI or AD condition, respectively.

To link two data sources  $\mathbf{X}$  and  $\mathbf{Z}$  together, we introduce common latent features  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]$  and assume  $\mathbf{X}$  and  $\mathbf{Z}$  are generated from  $\mathbf{U}$  by sparse projections. The common latent feature assumption is sensible for association studies because both SNPs and MRI features are biological measurements of the same subjects. Note that  $\mathbf{u}_i$  is the latent feature for the  $i$ -th subject and its dimension  $k$  is estimated by evidence maximization. In a Bayesian framework, we give a Gaussian prior over  $\mathbf{U}$ ,  $p(\mathbf{U}) = \prod_i \mathcal{N}(\mathbf{u}_i | \mathbf{0}, \mathbf{I})$ , and specify the rest of the model (see Figure 1) as follows: **Continuous data.** Given  $\mathbf{U}$ ,  $\mathbf{X}$  is generated from

$$p(\mathbf{X} | \mathbf{U}, \mathbf{G}, \eta) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}_i | \mathbf{G} \mathbf{u}_i, \eta^{-1} \mathbf{I})$$

where  $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_p]^\top$  is a  $p \times k$  projection matrix,  $\mathbf{I}$  is an identity matrix, and  $\eta^{-1} \mathbf{I}$  is the precision matrix of the Gaussian distribution. For  $\eta$ , we assign an uninformative diffuse Gamma prior,  $p(\eta | r_1, r_2) = \text{Gamma}(\eta | r_1, r_2)$  with  $r_1 = r_2 = 10^{-3}$ .

**Ordinal data.** For an ordinal observation  $z \in \{0, 1, \dots, R-1\}$ , its value is decided by which region an auxiliary variable  $c$  falls in  $-\infty = b_0 < b_1 < \dots < b_R = \infty$ . If  $c$  falls in  $[b_r, b_{r+1})$ ,  $z$  is set to be  $r$ . For the AD study, the SNPs  $\mathbf{Z}$  take values in  $\{0, 1, 2\}$  and therefore  $R = 3$ . Given a  $q \times k$  projection matrix  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_q]^\top$ , the auxiliary variables  $\mathbf{C} = \{c_{ij}\}$  and the ordinal data  $\mathbf{Z}$  are generated from

$$p(\mathbf{Z}, \mathbf{C} | \mathbf{U}, \mathbf{H}) = \prod_{i=1}^q \prod_{j=1}^n \mathcal{N}(c_{ij} | \mathbf{h}_i^\top \mathbf{u}_j, 1) \sum_{r=0}^2 \delta(z_{ij} = r) \delta(b_r \leq c_{ij} < b_{r+1})$$

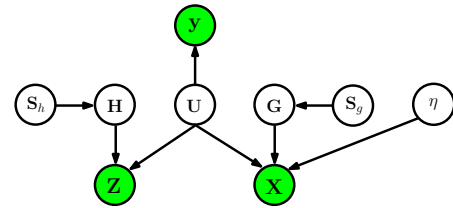


Fig. 1. The probabilistic graphical model of SHML, where  $\mathbf{X}$  is the continuous view,  $\mathbf{Z}$  is the ordinal view, and  $\mathbf{y}$  are the labels.

where  $\delta(a) = 1$  if  $a$  is true and  $\delta(a) = 0$  otherwise, and  $[b_0, \dots, b_3]$  are set to  $[-\infty, -1, 1, \infty]$ .

**Labels.** For ordinal labels  $\mathbf{y}$ , we use a Gaussian process ordinal regression model<sup>15</sup> based the latent representation  $\mathbf{U}$ ,

$$p(\mathbf{y}|\mathbf{U}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}) \prod_{i=1}^n \sum_{r=0}^2 \delta(y_i = r) \delta(b_r \leq f_i < b_{r+1})$$

where  $[b_0, \dots, b_3]$  are set to  $[-\infty, -1, 1, \infty]$ , and  $K_{ij} = k(\mathbf{u}_i, \mathbf{u}_j)$  is the cross-covariance between  $\mathbf{u}_i$  and  $\mathbf{u}_j$ . We can choose  $k$  from a rich family of kernel functions such as linear, polynomial, and Gaussian kernels to model relationships between the labels  $\mathbf{y}$  and the latent features  $\mathbf{U}$ .

Note that the labels  $\mathbf{y}$  are linked to the data  $\mathbf{X}$  and  $\mathbf{Z}$  via the latent features  $\mathbf{U}$  and the projection matrices  $\mathbf{H}$  and  $\mathbf{G}$ . Due to the sparsity in  $\mathbf{H}$  and  $\mathbf{G}$ , only a few groups of variables in  $\mathbf{X}$  and  $\mathbf{Z}$  are selected to predict  $\mathbf{y}$ . Note that each of group is linked to a feature in  $\mathbf{U}$ .

**Sparse Priors.** Because we want to identify a few critical interactions between different data sources, we use spike and slab prior distributions<sup>12</sup> to sparsify the projection matrices  $\mathbf{G}$  and  $\mathbf{H}$ . Specifically, we use a  $p \times k$  matrix  $\mathbf{S}_g$  to represent the selection of elements in  $\mathbf{G}$ : if  $s_{ij} = 1$ ,  $g_{ij}$  is selected and follows a Gaussian prior distribution with variance  $\sigma_1^2$ ; if  $s_{ij} = 0$ ,  $g_{ij}$  is not selected and forced to almost zero (*i.e.*, sampled from a Gaussian with a very small variance  $\sigma_2^2$ ). Specifically, we have the following prior over  $\mathbf{G}$ :

$$p(\mathbf{G}|\mathbf{S}_g, \mathbf{\Pi}_g) = \prod_{i=1}^p \prod_{j=1}^k \pi_g^{ij s_{ij}^{ij}} (1 - \pi_g^{ij})^{1-s_{ij}^{ij}} (s_{ij}^{ij} \mathcal{N}(g_{ij}|0, \sigma_1^2) + (1 - s_{ij}^{ij}) \mathcal{N}(g_{ij}|0, \sigma_2^2))$$

where  $\pi_g^{ij}$  in  $\mathbf{\Pi}_g$  is the probability of  $s_{ij}^{ij} = 1$ , and  $\sigma_1^2 \gg \sigma_2^2$  (in our experiment, we set  $\sigma_1^2 = 1$  and  $\sigma_2^2 = 10^{-6}$ ). Without any prior preference over the selecting probabilities, we assign uniform priors,  $p(\mathbf{\Pi}_g) = 1$ . Similarly,  $\mathbf{H}$  is sampled from

$$p(\mathbf{H}|\mathbf{S}_h, \mathbf{\Pi}_h) = \prod_{i=1}^q \prod_{j=1}^k \pi_h^{ij s_{ij}^{ij}} (1 - \pi_h^{ij})^{1-s_{ij}^{ij}} (s_{ij}^{ij} \mathcal{N}(h_{ij}|0, \sigma_1^2) + (1 - s_{ij}^{ij}) \mathcal{N}(h_{ij}|0, \sigma_2^2))$$

where  $\mathbf{S}_h$  are binary selection variables and  $\pi_h^{ij}$  in  $\mathbf{\Pi}_h$  is the probability of  $s_{ij}^{ij} = 1$ . Again, we assign uninformative uniform priors over  $\mathbf{\Pi}_h$ :  $p(\mathbf{\Pi}_h) = 1$ .

Finally, the joint distribution of our model, SHML, is simply the product of all the prior distributions and the conditional density distributions.

### 3. Algorithm

#### 3.1. Estimating latent variables

Given the model specified in the previous section, now we present an efficient, principled method to estimate the latent features  $\mathbf{U}$ , the projection matrices  $\mathbf{H}$  and  $\mathbf{G}$ , the selection indicators  $\mathbf{S}_g$  and  $\mathbf{S}_h$ , the selection probabilities  $\mathbf{\Pi}_g$  and  $\mathbf{\Pi}_h$ , the variance  $\eta$ , the auxiliary variables  $\mathbf{C}$  for generating ordinal data  $\mathbf{Z}$ , and the auxiliary variables  $\mathbf{f}$  for generating the labels  $\mathbf{y}$ . In a Bayesian framework, this estimation task amounts to computing their posterior distributions. However, computing the exact posteriors turns out to be infeasible since we cannot calculate the normalization constant of the exact posterior distribution. Thus, we resort

to a variational Bayesian Expectation Maximization (VB-EM) approach. More specifically, in the E step, we approximate the posterior distributions of  $\mathbf{H}, \mathbf{G}, \mathbf{S}_g, \mathbf{S}_h, \mathbf{\Pi}_g, \mathbf{\Pi}_h, \eta, \mathbf{C}$  and  $\mathbf{f}$  by a factorized distribution  $Q(\mathbf{H})Q(\mathbf{G})Q(\mathbf{S}_g)Q(\mathbf{S}_h)Q(\mathbf{\Pi}_g)Q(\mathbf{\Pi}_h)Q(\eta)Q(\mathbf{C})Q(\mathbf{f})$ ; and in the M step, based on the approximate distributions, we optimize the latent features  $\mathbf{U}$ .

To obtain the variational approximation, we minimize the Kullback-Leibler (KL) divergence between the approximate and the exact posteriors. To this end, we use coordinate descent; we update an approximate distribution, say,  $Q(\mathbf{H})$ , while fixing the other approximate distributions, and iteratively refine all the approximate distributions. The detailed updates are given in the following paragraphs.

### 3.1.1. Updating variational distributions for continuous data

For the continuous data  $\mathbf{X}$ , the approximate distributions of the projection matrix  $\mathbf{G}$ , the noise variance  $\eta$ , the selection indicators  $\mathbf{S}_g$  and the selection probabilities  $\mathbf{\Pi}_g$  are

$$Q(\mathbf{G}) = \prod_{i=1}^p \mathcal{N}(\mathbf{g}_i; \boldsymbol{\lambda}_i, \boldsymbol{\Omega}_i) \quad Q(\eta) = \text{Gamma}(\eta | \tilde{r}_1, \tilde{r}_2), \quad (1)$$

$$Q(\mathbf{S}_g) = \prod_{i=1}^p \prod_{j=1}^k \beta_{ij}^{s_{ij}^{ij}} (1 - \beta_{ij})^{1-s_{ij}^{ij}} \quad Q(\mathbf{\Pi}_g) = \prod_{i=1}^p \prod_{j=1}^k \text{Beta}(\pi_g^{ij} | \tilde{l}_1^{ij}, \tilde{l}_2^{ij}). \quad (2)$$

The mean and covariance of  $\mathbf{g}_i$  are calculated as  $\boldsymbol{\Omega}_i = (\langle \eta \rangle \mathbf{U} \mathbf{U}^\top + \frac{1}{\sigma_1^2} \text{diag}(\langle \mathbf{s}_g^i \rangle) + \frac{1}{\sigma_2^2} \text{diag}(\mathbf{1} - \langle \mathbf{s}_g^i \rangle))^{-1}$  and  $\boldsymbol{\lambda}_i = \boldsymbol{\Omega}_i (\langle \eta \rangle \mathbf{U} \tilde{\mathbf{x}}_i)$ , where  $\langle \cdot \rangle$  means expectation over a distribution,  $\tilde{\mathbf{x}}_i$  and  $\mathbf{s}_g^i$  are the transpose of the  $i$ -th rows of  $\mathbf{X}$  and  $\mathbf{S}_g$ ,  $\langle \mathbf{s}_g^i \rangle = [\beta_{i1}, \dots, \beta_{ik}]^\top$ , and  $\langle g_{ij}^2 \rangle$  is the  $j$ -th diagonal element in  $\boldsymbol{\Omega}_i$ . The parameters of the Gamma distribution  $Q(\eta)$  are updated as  $\tilde{r}_1 = r_1 + \frac{np}{2}$  and  $\tilde{r}_2 = r_2 + \frac{1}{2} \text{tr}(\mathbf{X} \mathbf{X}^\top) - \text{tr}(\langle \mathbf{G} \rangle \mathbf{U} \mathbf{X}^\top) + \frac{1}{2} \text{tr}(\mathbf{U} \mathbf{U}^\top \langle \mathbf{G}^\top \mathbf{G} \rangle)$ . The parameter  $\beta_{ij}$  in  $Q(s_{ij}^{ij})$  is calculated as  $\beta_{ij} = 1 / (1 + \exp(\langle \log(1 - \pi_g^{ij}) \rangle - \langle \log(\pi_g^{ij}) \rangle + \frac{1}{2} \log(\frac{\sigma_1^2}{\sigma_2^2}) + \frac{1}{2} \langle g_{ij}^2 \rangle (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})))$ . The parameters of the Beta distribution  $Q(\pi_g^{ij})$  is given by  $\tilde{l}_1^{ij} = \beta_{ij} + 1$  and  $\tilde{l}_2^{ij} = 2 - \beta_{ij}$ .

The moments required in the above distributions are calculated as  $\langle \eta \rangle = \frac{\tilde{r}_1}{\tilde{r}_2}$ ,  $\langle \mathbf{G} \rangle = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_p]^\top$ ,  $\langle \mathbf{G}^\top \mathbf{G} \rangle = \sum_{i=1}^p \boldsymbol{\Omega}_i + \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^\top$ ,  $\langle \log(\pi_g^{ij}) \rangle = \psi(\tilde{l}_1^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij})$  and  $\langle \log(1 - \pi_g^{ij}) \rangle = \psi(\tilde{l}_2^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij})$ , where  $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$ .

### 3.1.2. Updating variational distributions for ordinal data

For the ordinal data  $\mathbf{Z}$ , we update the approximate distributions of the projection matrix  $\mathbf{H}$ , the auxiliary variables  $\mathbf{C}$ , the sparse selection indicators  $\mathbf{S}_h$  and the selection probabilities  $\mathbf{\Pi}_h$ . Specifically, the variational distributions of  $\mathbf{C}$ ,  $\mathbf{H}$ ,  $\mathbf{S}_h$  and  $\mathbf{\Pi}_h$  are

$$Q(\mathbf{C}) \propto \prod_{i=1}^q \prod_{j=1}^k \delta(b_{z_{ij}} \leq c_{ij} < b_{z_{ij}+1}) \mathcal{N}(c_{ij} | \bar{c}_{ij}, 1) \quad Q(\mathbf{H}) = \prod_{i=1}^q \mathcal{N}(\mathbf{h}_i; \boldsymbol{\gamma}_i, \boldsymbol{\Lambda}_i), \quad (3)$$

$$Q(\mathbf{S}_h) = \prod_{i=1}^q \prod_{j=1}^k \alpha_{ij}^{s_{ij}^{ij}} (1 - \alpha_{ij})^{1-s_{ij}^{ij}} \quad Q(\mathbf{\Pi}_h) = \prod_{i=1}^q \prod_{j=1}^k \text{Beta}(\pi_h^{ij} | \tilde{d}_1^{ij}, \tilde{d}_2^{ij}), \quad (4)$$

where  $\bar{c}_{ij} = \boldsymbol{\gamma}_i^\top \mathbf{u}_j$ ,  $\boldsymbol{\Lambda}_i = (\mathbf{U} \mathbf{U}^\top + \frac{1}{\sigma_1^2} \text{diag}(\langle \mathbf{s}_h^i \rangle) + \frac{1}{\sigma_2^2} \text{diag}(\mathbf{1} - \langle \mathbf{s}_h^i \rangle))^{-1}$ ,  $\boldsymbol{\gamma}_i = \boldsymbol{\Lambda}_i (\mathbf{U} \tilde{\mathbf{c}}_i)$  where  $\tilde{\mathbf{c}}_i$  is the transpose of the  $i$ -th row of  $\mathbf{C}$ ,  $\alpha_{ij} = 1 / (1 + \exp(\langle \log(1 - \pi_h^{ij}) \rangle - \langle \log(\pi_h^{ij}) \rangle + \frac{1}{2} \log(\frac{\sigma_1^2}{\sigma_2^2}) +$



$\frac{1}{2}\langle h_{ij}^2 \rangle (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))$ ,  $\tilde{d}_1^{ij} = \alpha_{ij} + 1$ ,  $\tilde{d}_2^{ij} = 2 - \alpha_{ij}$ ,  $\langle \mathbf{s}_h^i \rangle = [\alpha_{i1}, \dots, \alpha_{ik}]^\top$ , and  $\langle h_{ij}^2 \rangle$  is the  $j$ -th diagonal element in  $\mathbf{\Lambda}_i$ .

The required moments for updating the above distributions can be calculated as  $\langle \log(\pi_h^{ij}) \rangle = \psi(\tilde{d}_1^{ij}) - \psi(\tilde{d}_1^{ij} + \tilde{d}_2^{ij})$ ,  $\langle \log(1 - \pi_h^{ij}) \rangle = \psi(\tilde{d}_2^{ij}) - \psi(\tilde{d}_1^{ij} + \tilde{d}_2^{ij})$ ,  $\langle \tilde{c}_i \rangle = [\langle c_{i1} \rangle, \dots, \langle c_{in} \rangle]^\top$  and  $\langle c_{ij} \rangle = \bar{c}_{ij} - (\mathcal{N}(b_{z_{ij}+1}|\bar{c}_{ij}, 1) - \mathcal{N}(b_{z_{ij}}|\bar{c}_{ij}, 1)) / (\Phi(b_{z_{ij}+1} - \bar{c}_{ij}) - \Phi(b_{z_{ij}} - \bar{c}_{ij}))$ , where  $\Phi(\cdot)$  is the cumulative distribution function of a standard Gaussian distribution. Note that in Equation (3),  $Q(\mathbf{C})$  is the product of truncated Gaussian distributions and the truncation is controlled by the observed ordinal data  $\mathbf{Z}$ .

### 3.1.3. Updating variational distributions for labels

We update the variational distribution of the auxiliary variables  $\mathbf{f}$  as follows:

$$Q(\mathbf{f}) \propto \prod_{i=1}^n \delta(b_{y_i} \leq f_i < b_{y_i+1}) \mathcal{N}(f_i | \bar{f}_i, \sigma_{f_i}^2) \quad (5)$$

where  $\bar{f}_i = \mathbf{K}_{i,-i} \mathbf{K}_{-i,-i}^{-1} \langle \mathbf{f}_{-i} \rangle$  and  $\sigma_{f_i}^2 = \mathbf{K}_{i,i} - \mathbf{K}_{i,-i} \mathbf{K}_{-i,-i}^{-1} \mathbf{K}_{-i,i}$ .  $\mathbf{K}_{i,-i}$  is the covariance between  $\mathbf{u}_i$  and  $\mathbf{U}_{-i}$ ,  $\mathbf{K}_{-i,-i}$  is the covariance on  $\mathbf{U}_{-i}$  ( $\mathbf{U}_{-i} = [\mathbf{u}_1, \dots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \dots, \mathbf{u}_n]$ ),  $\langle \mathbf{f}_{-i} \rangle = [\langle f_1 \rangle, \dots, \langle f_{i-1} \rangle, \langle f_{i+1} \rangle, \dots, \langle f_n \rangle]^\top$ , and each  $\langle f_i \rangle$  is  $\langle f_i \rangle = \bar{f}_i - \sigma_{f_i}^2 \cdot (\mathcal{N}(b_{y_i+1}|\bar{f}_i, \sigma_{f_i}^2) - \mathcal{N}(b_{y_i}|\bar{f}_i, \sigma_{f_i}^2)) / (\Phi(\frac{b_{y_i+1} - \bar{f}_i}{\sigma_{f_i}}) - \Phi(\frac{b_{y_i} - \bar{f}_i}{\sigma_{f_i}}))$ . Note that  $Q(\mathbf{f})$  is also the product of truncated Gaussian distributions and the truncated region is decided by the ordinal label  $\mathbf{y}$ . In this way, the supervised information from  $\mathbf{y}$  is incorporated into estimation of  $\mathbf{f}$  and then estimation of the other quantities by the recursive updates.

### 3.1.4. Optimizing the latent representation $\mathbf{U}$

After the expectations of the other variables are calculated, we optimize  $\mathbf{U}$  by maximizing the following variational lower bound

$$\begin{aligned} F(\mathbf{U}) = & -\frac{1}{2} \text{tr}(\mathbf{U} \mathbf{U}^\top) + \langle \eta \rangle \text{tr}(\mathbf{X}^\top \langle \mathbf{G} \rangle \mathbf{U}) - \frac{1}{2} \text{tr}(\langle \mathbf{H}^\top \mathbf{H} \rangle \mathbf{U} \mathbf{U}^\top) - \frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} \text{tr}(\langle \mathbf{f} \mathbf{f}^\top \rangle \mathbf{K}^{-1}) \\ & - \frac{\langle \eta \rangle}{2} \text{tr}(\langle \mathbf{G}^\top \mathbf{G} \rangle \mathbf{U} \mathbf{U}^\top) + \text{tr}(\langle \mathbf{C} \rangle^\top \langle \mathbf{H} \rangle \mathbf{U}) + \text{constant}, \end{aligned} \quad (6)$$

where  $\langle \mathbf{H} \rangle = [\mathbf{h}_1, \dots, \mathbf{h}_q]^\top$ ,  $\langle \mathbf{H}^\top \mathbf{H} \rangle = \sum_{i=1}^p \mathbf{\Lambda}_i + \gamma_i \gamma_i^\top \langle \mathbf{f} \mathbf{f}^\top \rangle = \langle \mathbf{f} \rangle \langle \mathbf{f} \rangle^\top - \text{diag}(\langle \mathbf{f} \rangle^2) + \text{diag}(\langle \mathbf{f}^2 \rangle)$ ,  $\langle f_i^2 \rangle = \langle f_i \rangle^2 + \sigma_{f_i}^2 + \sigma_{f_i}^2 \cdot ((b_{y_i} - \langle f_i \rangle) \mathcal{N}(b_{y_i}|\langle f_i \rangle, \sigma_{f_i}^2)) / (\Phi(\frac{b_{y_i+1} - \langle f_i \rangle}{\sigma_{f_i}}) - \Phi(\frac{b_{y_i} - \langle f_i \rangle}{\sigma_{f_i}})) - \sigma_{f_i}^2 \cdot ((b_{y_i+1} - \langle f_i \rangle) \mathcal{N}(b_{y_i+1}|\langle f_i \rangle, \sigma_{f_i}^2)) / (\Phi(\frac{b_{y_i+1} - \langle f_i \rangle}{\sigma_{f_i}}) - \Phi(\frac{b_{y_i} - \langle f_i \rangle}{\sigma_{f_i}}))$ , and the constant means a value independent of  $\mathbf{U}$  so that it is irrelevant for optimizing  $\mathbf{U}$ . Note that we can optimize the dimension  $k$  by maximizing the full variational lower bound of our model, which involves other quantities as well, such as  $\langle \mathbf{H} \rangle$  and  $\langle \mathbf{G} \rangle$ . To save space, we do not present the long equation for the full lower bound (which can be easily derived based on what we have presented). We use the L-BFGS algorithm to maximize the cost function  $F$  over  $\mathbf{U}$ . The gradient of  $\mathbf{U}$  is given by

$$\frac{\partial F}{\partial \mathbf{U}} = \langle \eta \rangle \langle \mathbf{G} \rangle^\top \mathbf{X} + \langle \mathbf{H} \rangle^\top \langle \mathbf{C} \rangle - (\mathbf{I} + \langle \eta \rangle \langle \mathbf{G}^\top \mathbf{G} \rangle + \langle \mathbf{H}^\top \mathbf{H} \rangle) \mathbf{U} - \frac{1}{2} (\mathbf{K}^{-1} - \frac{1}{2} \mathbf{K}^{-1} \langle \mathbf{f} \mathbf{f}^\top \rangle \mathbf{K}^{-1}) \frac{\partial \mathbf{K}}{\partial \mathbf{U}}. \quad (7)$$

Note that  $\frac{\partial \mathbf{K}}{\partial \mathbf{U}}$  depends on the form of the kernel function  $k(\mathbf{u}_i, \mathbf{u}_j)$ .

**Computational complexity.** Based on the previous equations, we can show that the total computational complexity of our algorithm is  $O(\max(n^3, (p+q)nk^2))$ —it is either cubic in the number of samples  $n$  or linear in the number of the features.

### 3.2. Predicting disease status

Let us denote the training data as  $\mathcal{D}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{Z}_{\text{train}}, \mathbf{y}_{\text{train}}\}$  and the test data as  $\mathcal{D}_{\text{test}} = \{\mathbf{X}_{\text{test}}, \mathbf{Z}_{\text{test}}\}$ . To obtain the latent representation  $\mathbf{U}_{\text{train}}$  and  $\mathbf{U}_{\text{test}}$  for prediction, we carry out variational EM simultaneously on  $\mathcal{D}_{\text{train}}$  and  $\mathcal{D}_{\text{test}}$ . The benefit is that the variational EM learning procedure can utilize both the training and test data. Note that there are no updates for ordinal label part on  $\mathbf{D}_{\text{test}}$  and the terms regarding ordinal labels should also be removed from Equation (6) and (7). After both  $\mathbf{U}_{\text{test}}$  and  $\mathbf{U}_{\text{train}}$  are obtained from the M-step, we predict the labels for test data as follows:

$$\mathbf{f}_{\text{test}} = \mathbf{K}(\mathbf{U}_{\text{test}}, \mathbf{U}_{\text{train}}) \mathbf{K}^{-1}(\mathbf{U}_{\text{train}}, \mathbf{U}_{\text{train}}) \langle \mathbf{f}_{\text{train}} \rangle \quad y_{\text{test}}^i = \sum_{r=0}^{R-1} r \cdot \delta(b_r \leq f_{\text{test}}^i < b_{r+1}),$$

where  $y_{\text{test}}^i$  is the prediction for  $i$ -th test sample.

## 4. Experiments

### 4.1. Simulation Study

We first design a simulation study to examine SHML in terms of (i) estimation accuracy on finding associations between the two views and (ii) prediction accuracy on the ordinal labels.

**Simulation data.** To generate the ground truth, we set  $n = 200$  (200 instances),  $p = q = 40$ , and  $k = 5$ . We designed  $\mathbf{G}$ , the  $40 \times 5$  projection matrix for the continuous data  $\mathbf{X}$ , to be a block diagonal matrix; each column of  $\mathbf{G}$  had 8 elements being ones and the rest of them were zeros, ensuring each row with only one nonzero element. We designed  $\mathbf{H}$ , the  $40 \times 5$  projection matrix for the ordinal data  $\mathbf{Z}$ , to be a block diagonal matrix; each of the first four columns of  $\mathbf{H}$  had 10 elements being ones and the rest of them were zeros, and the fifth column contained only zeros. We randomly generated the latent representations  $\mathbf{U} \in \mathbb{R}^{k \times n}$  with each column  $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . To generate  $\mathbf{Z}$ , we first sampled the auxiliary variables  $\mathbf{C}$  with each column  $\mathbf{c}_i \sim \mathcal{N}(\mathbf{H}\mathbf{u}_i, \mathbf{I})$ , and then decided the value of each element  $z_{ij}$  by the region  $c_{ij}$  fell in—in other words,  $z_{ij} = \sum_{r=0}^{R-1} r \delta(b_r \leq c_{ij} < b_{r+1})$ . Similarly, to generate  $\mathbf{y}$ , we sampled the auxiliary variables  $\mathbf{f}$  from  $\mathcal{N}(\mathbf{0}, \mathbf{U}^\top \mathbf{U} + \mathbf{I})$  and then each  $y_i$  was generated by  $p(y_i | f_i) = \delta(y_i = 0) \delta(f_i \leq 0) + \delta(y_i = 1) \delta(f_i > 0)$ .

**Comparative methods.** We compared SHML with several state-of-the-art methods including (1) CCA,<sup>4</sup> which finds the projection directions that maximize the correlation between two views, (2) sparse CCA,<sup>6,18</sup> where sparse priors are put on the CCA directions, and (3) multiple-response regression with lasso (MRLasso)<sup>19</sup> where each column of the second view ( $\mathbf{Z}$ ) is regarded as the output of the first view ( $\mathbf{X}$ ). We did not include results from the sparse probabilistic projection approach<sup>20</sup> because it performed unstably in our experiments. Regarding the software implementation, we used the built-in Matlab routine for CCA and the code by<sup>18</sup> for sparse CCA. We implemented MRLasso based on the Glmnet package ([cran.r-project.org/web/packages/glmnet/index.html](http://cran.r-project.org/web/packages/glmnet/index.html)).

To test prediction accuracy, we compared our method with the following ordinal or multinomial regression methods: (1) lasso for multinomial regression,<sup>7</sup> (2) elastic net for multinomial regression,<sup>8</sup> (3) sparse ordinal regression with the spike and slab prior, (4) CCA + lasso, for which we first ran CCA to obtain the latent features  $\mathbf{H}$  and then applied lasso to predict  $\mathbf{y}$ , (5) CCA + elastic net, for which we first ran CCA to obtain the projection matrices and then applied elastic net on the projected data, (6) Gaussian Process Ordinal Regression (GPOR),<sup>15</sup> and (7) Laplacian Support Vector Machine (LapSVM),<sup>21</sup> a semi-supervised SVM classification method. We used the published code for lasso, elastic net, GPOR and LapSVM. For all the methods, we used 10-fold cross validation on the training data for each run to choose the kernel form (Gaussian or linear or Polynomials) and its parameters (the kernel width or polynomial orders) for SHML, GPOR, and LapSVM.

Because alternative methods cannot learn the dimension automatically for simple comparison, we provided the dimension of the latent representation to all the methods we tested in our simulations. We partitioned the data into 10 subsets and used 9 of them for training and 1 subset for testing; we repeated the procedure 10 times to generate the averaged test results.

**Results.** To estimate linkage (*i.e.*, interactions) between  $\mathbf{X}$  and  $\mathbf{Z}$ , we calculated the cross covariance matrix  $\mathbf{GH}^\top$ . We then computed the precision and the recall based on the ground truth. The precision-recall curves are shown in Figure 2. Clearly, our method successfully recovered almost all the links and significantly outperformed all the competing methods. This improvement may come from i) the use of the spike and slab priors, which not only remove irrelevant elements in the projection matrices but also avoid over-penalizing the active association structures (the Laplace prior used in sparse CCA does over penalize the relevant ones) and ii) more importantly, the supervision from the labels  $\mathbf{y}$ , which is probably the biggest difference between ours and the other methods for the association study. The prediction accuracies on unknown  $\mathbf{y}$  and their standard errors are shown in Figure 3a and the AUC and their standard errors are shown in Figure 3b. Our proposed SHML model achieves significant improvement over all the other methods. It reduces the prediction error of elastic net (which ranks the second best) by 25%, and reduces the error of LapSVM by 48%.

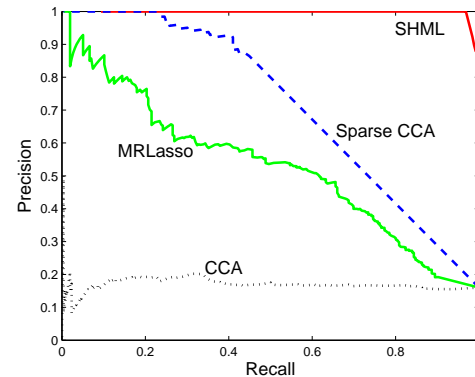


Fig. 2. The precision-recall curves for association discovery.

## 4.2. AD Study

We conducted joint association analysis and AD diagnosis based on the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI 1) dataset. The ADNI study is a longitudinal multisite observational study of elderly individuals with normal cognition, mild cognitive impairment, or AD. Specifically, we used SHML to study the associations of genotypes and brain atrophy measured by MRI and to predict the disease status (normal vs MCI vs AD). Note that the labels are ordinal since the three states represent increasing severity levels of AD.

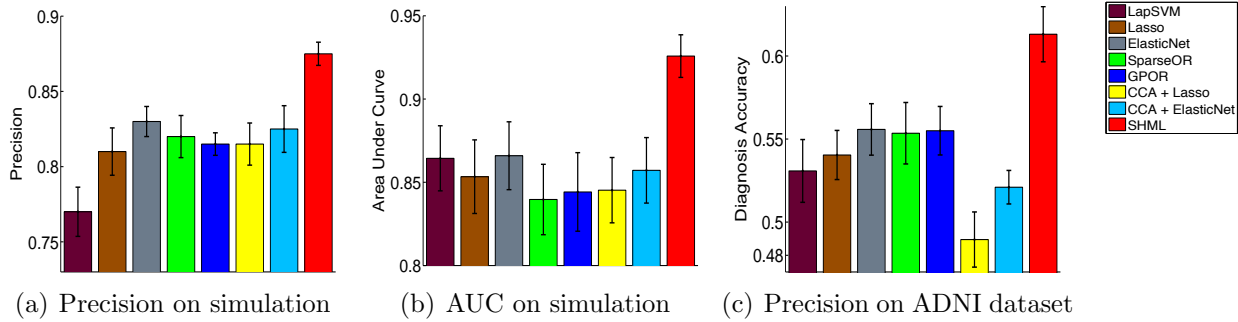


Fig. 3. The prediction results on simulated and real datasets. The results are averaged over 10 runs. The error bars represent standard errors. For the real ADNI dataset, we predict the ordinal disease status, Normal, MCI and AD.

Genetic and phenotypic data used in this study were obtained from the ADNI database (<http://www.loni.ucla.edu/ADNI>). Genomic DNA samples of 818 ADNI 1 subjects were analyzed on the Human610-Quad BeadChip according to the manufacturer’s protocols. After quality control, a list of 512,788 SNPs was used in an initial GWAS analysis associating them with the disease trait (AD vs. normal subjects). As a result, the top 1000 SNPs were pre-selected for analysis in this study. For structural MRI, we used image analysis results from UCSF based on the Freesurfer software package (<http://surfer.nmr.mgh.harvard.edu>); the resulting imaging data includes volumetric, cortical thickness and surface area measurements for a variety cortical and subcortical regions. After removing missing data, the final dataset consists of 618 subjects (183 normal, 308 MCI and 134 AD), and 924 SNPs and 328 MRI features measuring the brain atrophies for each subject at baseline.

We compared SHML with the alternative methods on accuracy of predicting whether a subject is in the normal or MCI or AD condition. We randomly split the dataset into 556 training and 62 test samples 10 times and ran all the competing methods on each partition. We used the 10-fold cross validation for each run to tune free parameters on the training data. In SHML, in order to determine  $k$ , the dimension of  $\mathbf{U}$ , we computed the variational lower bound as an approximation to the model marginal likelihood with various  $k$  values  $\{10, 20, 40, 60\}$ . We chose the value with the largest approximate evidence, which led to  $k = 20$  (see Figure 4). Our experiments confirmed that, with  $k = 20$ , SHML achieved highest prediction accuracy, demonstrating the benefit of evidence maximization.

The accuracies for predicting unknown labels  $\mathbf{y}$  and their standard errors are shown in Figure 3c. Our method achieved the highest prediction accuracy, higher than that of the second best method, GP ordinal Regression, by 10% and than that of the worst method, CCA+lasso, by 22%.

We also examined the strongest associations discovered by SHML based on the whole dataset. First of all, the ranking of MRI features in terms of their prediction power of different disease stages (normal, MCI and AD) demonstrates that most of the top ranked

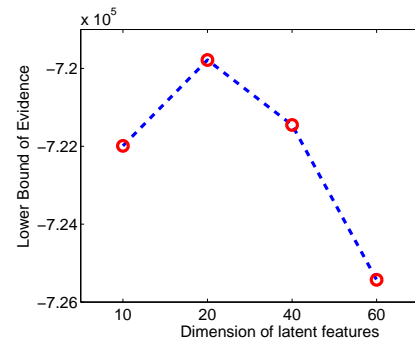


Fig. 4. The variational lower bound of the marginal likelihood (i.e., evidence).

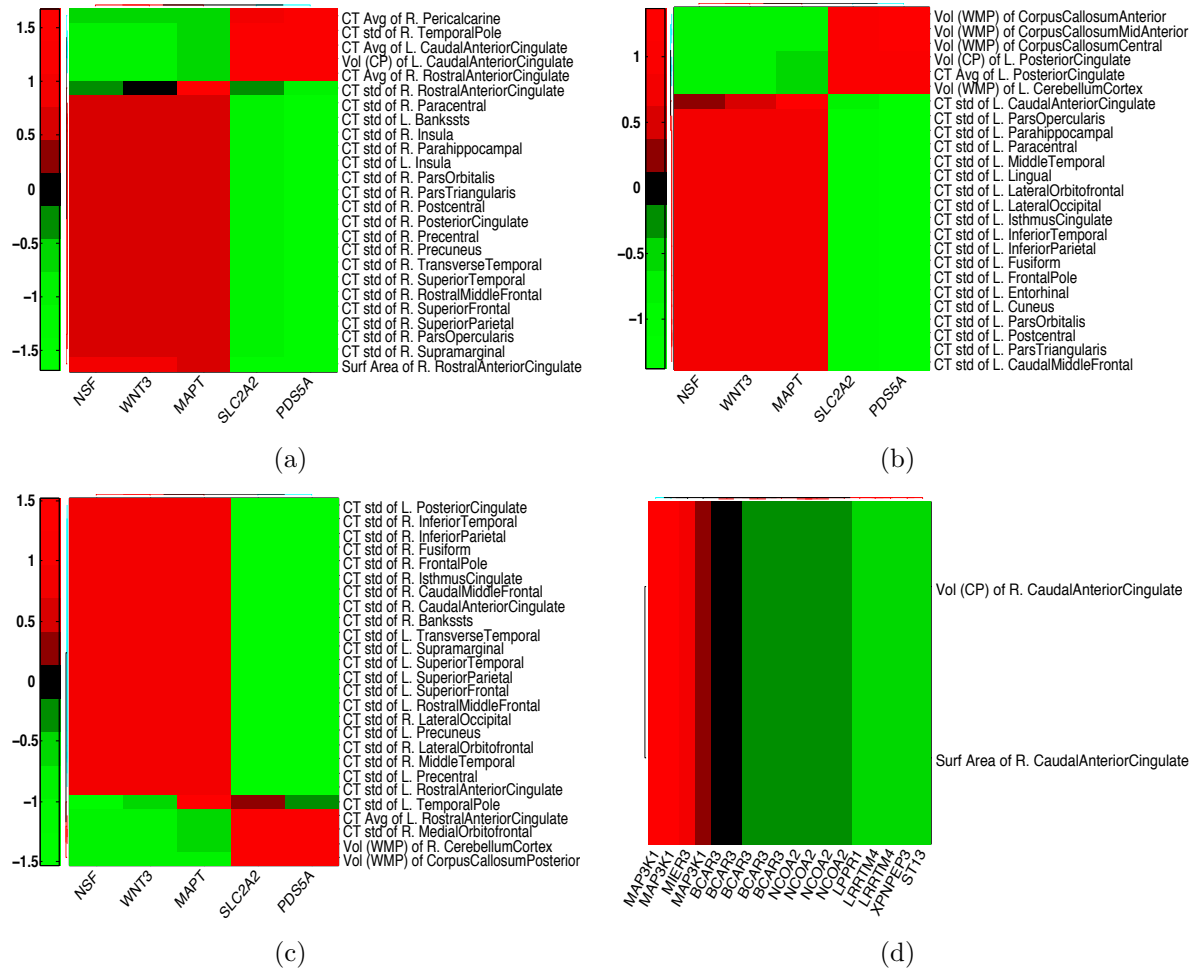


Fig. 5. The estimated associations between MRI features and SNPs. In each sub-figure, the MRI features are listed on the right and the SNP names are given at the bottom.

features are the cortical thickness measurements, followed by the volume of white matter, volume of gray matter in cortical regions, and the cortical surface area measurements. These results are consistent with the literature for demonstrating that the cortical thickness measurement is potentially a more sensitive measurement of the brain atrophy for Alzheimer's dementia.<sup>22,23</sup> Particularly, thickness measurements of frontal lobe, middle temporal lobe, and precuneus were found to be most predictive compared with other brain regions. These findings are consistent with their atrophy pattern and prediction power of AD found in the literature<sup>23–27</sup>. We also found that measurements of the same structure on the left and right hemisphere have similar weights (See Table 1); this is again consistent with the related literature—no asymmetrical relationship has been found for the brain regions involved in AD.<sup>28</sup>

Table 1. The weights of the average cortical thickness of ROI on the left and right hemispheres.

ROI	weight	
	left	right
Superior Frontal	1.37	1.35
Middle Temporal	1.33	1.37
Precuneus	1.33	1.36
Inferior Parietal	1.29	1.34
Inferior Temporal	1.32	1.29
Caudal Middle Frontal	1.32	1.31
Rostral Middle Frontal	1.31	1.30

Secondly, the analysis of associating genotypes to AD also generated interesting results. Similar to the MRI features, SNPs that are in the vicinity of each other are selected together due to the *group-selection* characteristics of our algorithm. The top ranked SNPs are associated with a few genes including PSMC1P12 (proteasome 26S subunit, ATPase), NCOA2 (The nuclear receptor coactivator 2), and WDR52 (WD repeat domain 52). These genes have been associated with diseases such as breast neoplasms, carcinoma, and endometrial neoplasms.<sup>29</sup>

At last, biclustering of the genotype-MRI association, as shown in Figure 5, revealed interesting patterns in terms of the relationship between genetic variations and brain atrophy in association with AD. For example, the highest ranked association was found between genes such as MAP3K1 (mitogen-activated protein kinase kinase kinase 1) and MIER3 (mesoderm induction early response 1, family member 3) with the caudate anterior cingulate cortex. MAP3K1 and MIER3 genes are associated with biological process such as apoptosis, cell cycle, chromatin binding and DNA binding (<https://portal.genego.com/>), and cingulate cortex has been shown to be severely affected by AD<sup>30</sup>. The strong association discovered in this work might indicate potential genetic effect in the atrophy pattern observed in this cingulate sub-region. Additionally, SNPs in MAPT (microtubule-associated protein tau) gene were also found to have association with brain atrophy in a variety of cortical regions including frontal, cingulate and temperate lobes. The hyperphosphorylation of tau protein, which is a product of MAPT, can result in the self-assembly of tangles that are involved in the pathogenesis of AD. Therefore, the genetic variation of MAPT has been associated with increased risk of AD<sup>31–35</sup>. The association between MATP gene and brain atrophies found in this analysis is consistent with the gray matter loss observed in MATP genetic variant carrier in recent studies.<sup>36</sup>

In summary, SHML discovered the synergistic predictive relationships between brain atrophy, genetic variations and the disease status, and achieved higher prediction accuracy than the alternative methods.

## 5. Conclusions

We have presented, SHML, a new Bayesian supervised multiview learning algorithm for AD study. By integrating association discovery with disease diagnosis, it improves performance for both tasks. Although we have focused on the AD study in this paper, we expect that SHML can be applied to a wide range of applications in biomedical research—for example, eQTL analysis supervised by additional labeling information. As to the future work, we plan to incorporate additional biological or side information into our model to improve its quality. In particular, linkage disequilibrium structures encode important correlation information between SNPs. Our current model uses *independent*, uniform priors over the selection probabilities of SNPs, which ignore the correlation between SNPs (note that the posterior distribution of the model does capture some correlation between genetic variations based on the data likelihood). To overcome this limitation, we plan to use graph Laplacian matrices to encode linkage disequilibrium structures and use these matrices in our prior distributions. We have explored a similar strategy to incorporate biological pathway constraints for biomarker selection and obtained improved performance over the models that do not use the pathway information.<sup>37</sup> We expect a similar improvement can be obtained by incorporating LD structures into SHML.

## Acknowledgments

This work was supported by NSF IIS-0916443, NSF IIS-1054903, and the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. Data used in the work were obtained from the ADNI database. ADNI funding information is available at [http://adni.loni.ucla.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_DSP\\_Policy.pdf](http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_DSP_Policy.pdf)

## References

1. D. Harold *et al.*, *Nat. Genet.* **41**, 1088 (2009).
2. M. Vounou *et al.*, *Neuroimage* **60**, 700 (2012).
3. H. Harold, *Biometrika* **28**, 321 (1936).
4. F. Bach and M. Jordan, *A probabilistic interpretation of canonical correlation analysis*, tech. rep., UC Berkeley (2005).
5. E. Parkhomenko, D. Tritchler and J. Beyene, *BMC Proc.* **1 Suppl 1**, p. S119 (2007).
6. M. Daniela and R. Tibshirani, *Stat Appl Genet Mol Biol.* **383** (2009).
7. R. Tibshirani, *Journal of the Royal Statistical Society, Series B* **58**, 267 (1994).
8. H. Zou and T. Hastie, *Journal of the Royal Statistical Society, Series B* **67**, 301 (2005).
9. D. MacKay, *Neural Computation* **4**, 415 (1991).
10. P. Yu, R. A. Dean *et al.*, *J. Alzheimers Dis.* **32**, 373 (2012).
11. L. Shen, Y. Qi *et al.*, *Med Image Comput Comput Assist Interv.* **13**, 611 (2010).
12. E. George and R. McCulloch, *Statistica Sinica* **7**, 339 (1997).
13. I. Goodfellow *et al.*, Large-scale feature learning with spike-and-slab sparse coding, in *ICML*, 2012
14. S. Mohamed *et al.*, Bayesian and L1 approaches for sparse unsupervised learning, in *ICML*, 2012.
15. W. Chu and Z. Ghahramani, *Journal of Machine Learning Research* **6**, 1019 (2005).
16. J. Zhou *et al.*, Modeling disease progression via fused sparse group lasso, in *KDD'12*, 2012.
17. L. Yuan *et al.*, Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data, in *KDD'12*, 2012.
18. L. Sun *et al.*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 194 (2011).
19. S. Kim, K. Sohn and E. Xing, *Bioinformatics* **25**, 204 (2009).
20. C. Archambeau and F. Bach, Sparse probabilistic projections, in *NIPS'09*, 2009.
21. S. Melacci and B. Mikhail, *Journal of Machine Learning Research* **12**, 1149 (2011).
22. J. Lerch, J. Pruessner, A. Zijdenbos *et al.*, *Neurobiol Aging* **1**, 23 (2008).
23. S. Teipel *et al.*, *Medical Clinics of North America* **97**, 399 (2013).
24. J. Whitwell, S. Przybelski, S. Weigand *et al.*, *Brain* **130**, 1777 (2007).
25. S. Risacher, A. Saykin, J. West, H. Firpi and B. McDonald, *Curr. Alzheimer Res.* **6**, 347 (2009).
26. S. Galluzzi, C. Geroldi *et al.*, *J. Neurol.* **14**, 2004 (2010).
27. J. Whitwell, H. Wiste *et al.*, *Arch. Neurol.* **69**, 614 (May 2012).
28. O. Y. Kusbeci *et al.*, *Dement Geriatr Cogn Disord* **28**, 1 (2009).
29. P. J. Stephens, P. S. Tarpey *et al.*, *Nature* **486**, 400 (Jun 2012).
30. B. F. Jones *et al.*, *Cereb. Cortex* **16**, 1701 (Dec 2006).
31. M. J. Bullido *et al.*, *Neurosci. Lett.* **278**, 49 (Jan 2000).
32. H. Tanahashi, T. Asada and T. Tabira, *Neuroreport* **15**, 175 (Jan 2004).
33. T. Feulner, S. Laws *et al.*, *Mol. Psychiatry* **15**, 756 (Jan 2010).
34. E. Di Maria, S. Cammarata *et al.*, *J. Alzheimers Dis.* **19**, 909 (2010).
35. L. Samaranch, S. Cervantes *et al.*, *J. Alzheimers Dis.* **22**, 1065 (2010).
36. J. Goñi *et al.*, *J. Alzheimers Dis.* **33**, 1009 (2013).
37. S. Zhe *et al.*, *Bioinformatics* **29**, 1987 (2013).

## TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

GRACIELA GONZALEZ

*Department of Biomedical Informatics, Arizona State University  
Scottsdale, AZ 85259, USA  
Email: [ggonzalez@asu.edu](mailto:ggonzalez@asu.edu)*

KEVIN BRETONNEL COHEN

*U. Colorado School of Medicine  
Aurora, CO  
Email: [kevin.cohen@gmail.com](mailto:kevin.cohen@gmail.com)*

ROBERT LEAMAN

*National Center for Biotechnology  
Information Bethesda, MD 20894, USA  
Email: [robert.leaman@nih.gov](mailto:robert.leaman@nih.gov)*

CASEY S. GREENE

*Department of Genetics, Geisel School of  
Medicine at Dartmouth  
Hanover, NH 03755, USA  
Email: [Casey.S.Greene@dartmouth.edu](mailto:Casey.S.Greene@dartmouth.edu)*

NIGAM SHAH

*Center for Biomedical Informatics Research  
Stanford, CA 94305  
Email: [nigam@stanford.edu](mailto:nigam@stanford.edu)*

MARICEL G. KANN

*Department of Biological Science  
University of Maryland, Baltimore County  
Baltimore, MD 21250, USA  
Email: [mkann@umbc.edu](mailto:mkann@umbc.edu)*

JIEPING YE

*Computer Science and Engineering,  
Arizona State University, Tempe, AZ 85287  
Email: [jieping.ye@asu.edu](mailto:jieping.ye@asu.edu)*

Text and data mining methods constantly advance and are applied in different fields. In order for them to impact the biomedical discovery process, it is necessary to thoroughly engage scientists at both ends, and conduct thorough empirical evaluations as to their ability to suggest novel hypotheses and address the most crucial questions. The PSB 2014 Session on Text and Data Mining for Biomedical Discovery presents eight papers that advance the field in this mutually reinforcing fashion. Work presented in this session includes data mining and analysis techniques that are applicable to a broad spectrum of problems, including the analysis and visualization of mass spectrometry based proteomics data and longitudinal data, as well as gene function, protein function and protein fold prediction. Text mining approaches selected for presentation include a method for predicting genes involved in disease or in drug response, a method for extracting events relevant to biological pathways, and an approach that mixes text and data mining techniques to predict important milestones in the female reproductive lifespan.



## 1. Introduction

This session seeks to bring together researchers with a strong text or data mining background who are collaborating with bench scientists for the deployment of integrative approaches in translational bioinformatics. It serves as a unique forum to discuss novel approaches to text and data mining methods that respond to specific scientific questions, enabling predictions that integrate a variety of data sources and can potentially impact scientific discovery.

Successes in the application of computational approaches that solve biological problems have led to the broad application of these methods to an ever-growing set of specific problem areas. Consequently it is no longer possible to enumerate the biological questions targeted by computational approaches. These questions include, but are not limited to, the problems addressed by papers in this session. Broadly though, we can discuss trends in the field.

While data mining approaches have previously been applied to biological questions in ways that assume the functions of genes are constant, advances in underlying computational platforms and methodology are now allowing computational biologists to begin to address problems in a context specific manner. This means that instead of asking about the overall function of a gene, we are now identifying the role of a gene in a given environment, cell lineage, or individual. We anticipate that approaches that embrace rather than ignore such underlying biological complexities will provide the next generation of advances in personalized medicine.

## 2. Challenges

The biomedical domain presents specific challenges to text and data mining given the diversity, complexity and volume of the information being mined. The submissions to this session allowed us a unique glimpse at these challenges, which can perhaps be summarized as the constant call to fully incorporate the richness of the available resources and tackle the analysis of data of ever-growing complexity.

Thus, an overarching challenge for biomedical text mining is to incorporate the many knowledge resources that are available to us into the natural language processing pipeline. In the biomedical domain, unlike the general text mining domain, we have access to large numbers of extensive, well-curated ontologies and knowledge bases. However, we have, in general, failed to take advantage of them for tasks like coreference resolution, semantic typing of possible subjects and objects of predicates in information extraction, and the like.

Biomedical ontologies provide an explicit characterization of a given domain of interest and can enhance biomedical discovery significantly when used in a pragmatic manner. Using existing ontologies (from the UMLS and BioPortal) as sources of terms in building lexicons, for figuring out what concept subsumes what other concept, and as a way of normalizing alternative names to one identifier, would likely increase the quality of data-mining efforts. For example, using ontologies as described enabled the use of unstructured clinical notes for generating practice-based evidence on the safety of a highly effective, generic drug for peripheral vascular disease (PubMed 23717437).

Among the papers in this session, there are several examples where important advances to biomedical discovery are based on precisely this expansion on the use of knowledge and

literature resources. For example, Funk et al predict pharmacogenomic genes on a genome-wide scale using Gene Ontology annotations and simple features mined from the biomedical literature. Ravikumar et al, present a rule-based literature mining system to extract pathway information from text to assist human curators.

Today, the data being generated is massive, complex, and increasingly diverse due to recent technological innovations. However, the impacts of this data revolution on our lives are being hampered by the limited amount of data that has been analyzed. This necessitates data mining tools and methods that can match the scale of the data and support timely decision-making through integration of multiple heterogeneous data sources. We see in this session numerous contributions to methods and approaches, better outlined in the next section.

Finally, another area in which the field has fallen short and that the papers in this session can only begin to address, is that of making text mining applications that are easily adaptable by end users. Many researchers have developed systems that can be adapted by other text mining specialists, but applications that can be tuned by bench scientists are mostly lacking.

### **3. Overview of Contributions**

Funk et al. describe a method for predicting genes involved in disease or in drug response based on combining heterogeneous data, including curated Gene Ontology annotations, text-mined Gene Ontology annotations, and surface linguistic features. These feature types are combined and passed as input to a classifier.

Ravikumar et. al. develop a system to extract events relevant to biological pathways from the literature by combining named entity recognition and normalization with pattern templates to detect event mentions and the role of each entity. Notably, the system resolves both entity and event anaphora with discourse analysis. The authors evaluate their system against PharmGKB pathway annotations, and manually examine a subset of the results.

Malinowski et al report on development and performance of data-mining techniques to identify the age at menarche (AM) and age at menopause (AAM), which are important milestones in the reproductive lifespan; and are often recorded in free-text notes. The authors demonstrate the ability to discriminate age at naturally-occurring menopause (ANM) from medically-induced menopause. Their ultimate goal is to apply the methods to data from the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study, in an attempt to support clinical studies that incorporate these female reproductive milestones.

Han describes an application of a dimensionality reduction technique, called derivative component analysis (DCA) for the analysis and visualization of mass spectrometry based proteomics data. As an implicit feature selection algorithm, DCA enables to extract true signals by capturing subtle data characteristics and removing built-in data noises for input proteomics profiles.

Zupan and Zitnik develop a general matrix factorization-based data integration approach for gene function prediction that fuses heterogeneous data sources, such as gene expression data, known protein annotation, interaction and literature data. The fusion is achieved by

simultaneous matrix tri-factorization that shares matrix factors between data sources. The proposed approach is applicable for any number of data sources, which can be expressed in a matrix form.

Liu et al. describe a method for analyzing longitudinal data. Functional regression is a popular approach for longitudinal data analysis, as it is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. The proposed approach empowers basic functional regression models to simultaneously identify features with significant predictive power across time points, enforce smoothness of functional coefficients, and achieve interpretable estimations of functional coefficients using a novel sparsity-inducing penalty.

Clark and Radivojac develop a novel machine learning algorithm for protein function and fold prediction. In particular, their method introduces a kernel function on time series data that can be obtained from protein sequences and structures. The proposed kernel showed high performance in the task of classifying proteins in SCOP classes. Accurate functional classification of proteins is critical for understanding the molecular mechanisms involved in all biological process across species, which translates into advances in biomedical research. Furthermore, this methodology is applicable to problems beyond computational biology.

Vembu and Morris demonstrate *LMGraph*, two step approach to construct binary predictors from gene networks and features. The first step extracts informative features from the network. The second step combines these network-extracted features with node features to construct predictors. The authors demonstrate that this two-step approach outperforms related algorithms, suggesting that such combined approaches could offer benefits to other methods.

# VECTOR QUANTIZATION KERNELS FOR THE CLASSIFICATION OF PROTEIN SEQUENCES AND STRUCTURES

WYATT T. CLARK AND PREDRAG RADIVOJAC\*

*Department of Computer Science and Informatics, Indiana University  
Bloomington, Indiana 47405, U.S.A.*

*\*E-mail: predrag@indiana.edu*

We propose a new kernel-based method for the classification of protein sequences and structures. We first represent each protein as a set of time series data using several structural, physicochemical, and predicted properties such as a sequence of consecutive dihedral angles, hydrophobicity indices, or predictions of disordered regions. A kernel function is then computed for pairs of proteins, exploiting the principles of vector quantization and subsequently used with support vector machines for protein classification. Although our method requires a significant pre-processing step, it is fast in the training and prediction stages owing to the linear complexity of kernel computation with the length of protein sequences. We evaluate our approach on two protein classification tasks involving the prediction of SCOP structural classes and catalytic activity according to the Gene Ontology. We provide evidence that the method is competitive when compared to string kernels, and useful for a range of protein classification tasks. Furthermore, the applicability of our approach extends beyond computational biology to any classification of time series data.

*Keywords:* Protein classification, protein structure, protein function, kernels, vector quantization, support vector machines.

## 1. Introduction

The wealth and diversity of experimental data in the life sciences has strongly influenced the development of classification methods for biological macromolecules. Over the past couple of decades the scope and sophistication of these methods has significantly increased, leading to the adoption of classification schemes that are designed to integrate diverse types of biological data (e.g. sequence, structure, interaction networks, text), enable principled incorporation of domain knowledge, and rigorously deal with data of varying degrees of quality.<sup>1,2</sup>

Among the various classification strategies, kernel-based methods<sup>3,4</sup> have recently been introduced in a range of contexts such as the prediction of remote homology,<sup>5</sup> protein structure<sup>6</sup> and function,<sup>7,8</sup> protein-protein interactions,<sup>9</sup> gene-disease associations,<sup>10</sup> the activity of chemical compounds,<sup>11</sup> etc. Although some kernel methods have been developed to predict properties of individual residues,<sup>12</sup> most of these approaches have been used at the level of entire proteins. For example, several string kernels were introduced to provide inferences regarding remote homology from amino acid sequences.<sup>5,13–15</sup> Similarly, graph kernels have gained significant attention owing to the fact that a variety of biological data can be modeled through graphs.<sup>16</sup> A number of approaches have also considered integrating kernels built on different types of data.<sup>17,18</sup>

Kernels can be roughly described as symmetric positive semi-definite similarity functions that operate on pairs of objects from some input space.<sup>19</sup> Their mathematical properties guarantee the existence of a Hilbert space, potentially of infinite dimensionality, such that the value of the kernel function can be computed as the inner product of the images of

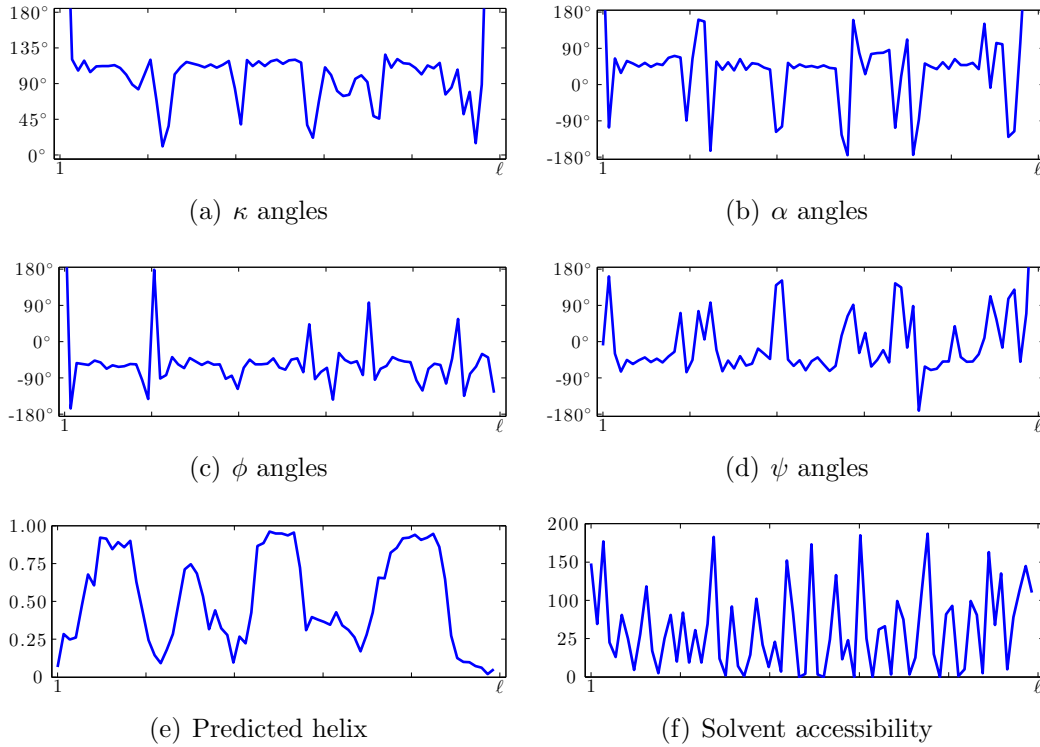


Fig. 1. Time series representation of various protein properties for the  $\ell = 73$  amino acid long DNA helicase RuvA subunit d1ixra1 from *Thermus thermophilus* (PDB ID 1ixr).

the input objects. When coupled with learning algorithms such as support vector machines, kernel functions also guarantee a globally optimal solution to the optimization problem.<sup>19</sup> Although most kernel-based approaches are in practice formed by vectorizing input objects, thereby not fully exploiting their theoretical potential, they still enable a practitioner to incorporate domain knowledge into modeling the relationship between objects, rather than simply encoding properties of the objects into a potentially high-dimensional vector space and providing them to a standard classifier.

In this work we focus on kernel-based strategies and develop novel methodology for the nonalignment-based classification of proteins into distinct categories. In contrast to most previously implemented kernel approaches, we represent a protein's sequence or structure, if available, as a set of time series properties (we consider a time series to be an ordered sequence of real-valued numbers<sup>20</sup>). One such time series representation of a DNA helicase subunit from *Thermus thermophilus* is shown in Figure 1, where six different types of properties have been generated based on the protein's sequence and structure. Given the time series data, we utilize ideas from vector quantization (VQ), initially developed for lossy signal compression,<sup>21</sup> to define a kernel function between pairs of protein sequences that we use for classification. We extensively evaluated methods on two distinct and relevant problems: (i) the classification of protein structures into structural classes and (ii) the prediction of protein function from amino acid sequence. Our experiments provide evidence that the new kernels are a viable approach in various practical scenarios.

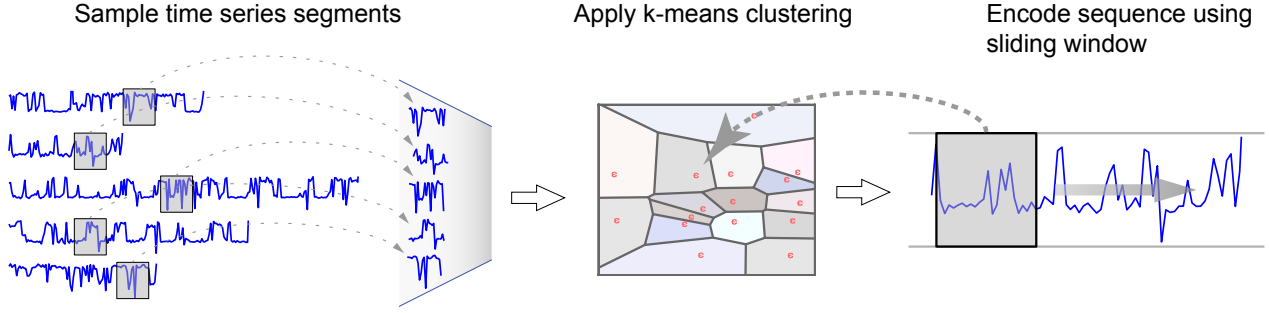


Fig. 2. A schematic representation of using VQ to encode a sequence represented as a property vector. First individual time series property vectors are broken up into  $n$  length segments as shown on the left. These sub-sampled vectors from a database of sequences are then used to create a clustering in  $n$ -dimensional space as shown in the center. Finally, an original property vector is encoded using the derived set of centroids by counting the number of overlapping sub-segments which are the closest to each centroid.

## 2. Methods

Let  $\mathcal{S} = \{s_1, s_2, s_3, \dots\}$  be a universe of protein sequences, where each  $s \in \mathcal{S}$  is a string of symbols from an alphabet of amino acids  $\mathcal{A} = \{A, C, D, \dots, Y\}$ . Let also  $\mathcal{S}_L \subset \mathcal{S}$  be a set of labeled sequences, e.g. those with known structural class or function, that is provided as training data. The objective is to use an inductive supervised framework to probabilistically annotate the remaining sequences, i.e. those from the unlabeled set  $\mathcal{S}_U = \mathcal{S} - \mathcal{S}_L$ .

To map protein sequences into a real-valued vector representation, let  $s = s_1 s_2 \dots s_\ell$  be a length- $\ell$  protein sequence in  $\mathcal{S}$  and  $\mathbf{p} = (p_1, p_2, \dots, p_\ell)$  some property vector defined by any particular mapping from  $\mathcal{A}^\ell$  to  $\mathbb{R}^\ell$ . For example,  $\mathbf{p}$  may be provided as a vector of hydrophobicity indices corresponding to amino acids in  $s$ . Alternatively, it can be represented as predicted helical propensities as outputted by some predictor of secondary structure. For those sequences with available structures,  $\mathbf{p}$  may correspond to a sequence of dihedral angles calculated from the protein structure model of  $s$ .

Consider now a single property vector  $\mathbf{p}$ , such as a hydrophobicity profile, corresponding to a particular sequence  $s \in \mathcal{S}$ . We decompose  $\mathbf{p}$  into  $n$ -dimensional overlapping sub-vectors  $\mathbf{p}_{[1,n]}, \mathbf{p}_{[2,n+1]}, \dots, \mathbf{p}_{[\ell-n+1,\ell]}$ , where  $\mathbf{p}_{[i,i+j]} = (p_i, p_{i+1}, \dots, p_{i+j})$ , and  $n \ll \ell$  is a small integer. For example,  $\mathbf{p}_{[1,n]}$  corresponds to the first  $n$  elements of  $\mathbf{p}$ . For a property vector (amino acid sequence) of length  $\ell$ , there are  $\ell - n + 1$  length- $n$  sub-vectors.

As described in Figure 2, given a set of length- $n$  property sub-vectors  $\mathcal{P}$  derived from the sequence universe  $\mathcal{S}$ , we generate a partition of  $\mathbb{R}^n$  into  $m$  regions  $\mathcal{R} = \{R_1, R_2, \dots, R_m\}$ . These regions are represented by a set of  $n$ -dimensional vectors, or centroids,  $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_m\}$ . Each region  $R_i$  represents a Voronoi region such that

$$R_i = \{\mathbf{x} : d(\mathbf{x}, \mathbf{c}_i) \leq d(\mathbf{x}, \mathbf{c}_j), i \neq j\},$$

where  $d(\mathbf{x}, \mathbf{c})$  is the Euclidean distance between vector  $\mathbf{x}$  and centroid  $\mathbf{c}$ . We determine  $\mathcal{C}$  using k-means clustering, where the initial set of clusters is generated by the splitting method.<sup>22</sup> We chose to use k-means clustering as opposed to simply creating a lattice in  $n$ -dimensional space,<sup>23</sup> because sampled property vectors do not fill the space evenly, but instead cluster

around evolutionarily conserved or sterically preferred regions.

### 2.1. Property kernels

Variable length property vectors can be transformed into vectors of length  $m$  using the partition of  $\mathbb{R}^n$  defined by  $\mathcal{R}$ . Specifically, a property vector  $\mathbf{p}$  is mapped into a vector of length  $m$  as

$$\mathbf{x} = (\varphi_1(\mathbf{p}), \varphi_2(\mathbf{p}), \dots, \varphi_m(\mathbf{p})),$$

where  $\varphi_i(\mathbf{p})$  is the number of  $n$ -dimensional vectors  $\mathbf{p}_{[i]}$  in  $\mathbf{p}$  that belong to region  $R_i$ . Given two property vectors  $\mathbf{p}$  and  $\mathbf{q}$  and their respective count vectors  $\mathbf{x}$  and  $\mathbf{y}$ , a *vector quantization property kernel* function is defined as

$$k(\mathbf{p}, \mathbf{q}) = \mathbf{x}^T \mathbf{y},$$

where  $T$  is the transpose operator. Note that in this notation each count vector is assumed to be a column vector, i.e.  $(a, b, c) = [a \ b \ c]^T$ , as in Ref. 24. Since the function  $k(\mathbf{p}, \mathbf{q})$  is defined as an inner product between two count vectors, it is a kernel function.<sup>19</sup>

Given a set of property kernels  $\{k_i(\mathbf{x}, \mathbf{y})\}$ , we construct the composite property kernel as a linear combination

$$k(\mathbf{x}, \mathbf{y}) = \sum_i k_i(\mathbf{x}, \mathbf{y}),$$

where before and after combining, each kernel is normalized using

$$k(\mathbf{x}, \mathbf{y}) \leftarrow \frac{k(\mathbf{x}, \mathbf{y})}{\sqrt{k(\mathbf{x}, \mathbf{x})k(\mathbf{y}, \mathbf{y})}}.$$

It is important to mention that both the inner product kernel formulation and the composite kernel based on a linear combination were selected for their simplicity. Functions such as the Jaccard similarity coefficient and the Gaussian kernel (which introduces a parameter into kernel selection) sometimes provide performance improvements to an inner product definition. Similarly, kernels can be combined using a product or hyperkernel formulations; however, recent evidence suggests that more sophisticated schemes typically result in only minor improvements over linear combinations.<sup>25</sup>

### 2.2. Computational complexity

The computation of a count vector can be accomplished in  $O(\ell mn)$  time if each  $n$ -dimensional vector from  $\mathbf{p}$  is compared with all centroids in  $\mathcal{C}$ . Approximation algorithms are available through a decision tree-like organization of the centroids. In such a case, only  $\log m$  distance calculations are needed resulting in  $O(\ell n \log m)$  time;<sup>22</sup> however, there is no guarantee that the closest centroid will be found. The memory requirements include  $O(mn)$  space for storing  $\mathcal{C}$ .

### 2.3. Spectrum kernel

We compare the property kernels to a string kernel approach, as described in Ref. 5, for a wide range of word sizes ( $n \in \{1 \dots 10\}$ ). For a given word size  $n$ , a sparse  $20^n$ -length vector was created for a protein sequence, where each dimension represented the number of times a potential substring of length  $n$  that could be generated using the 20 amino acid alphabet occurred. An  $\ell$ -length sequence,  $\mathbf{s}$ , contains  $\ell - n + 1$  such overlapping strings.

## 3. Data and experiments

In the first experiment, prediction was performed as a one-versus-all classification at the SCOP class level for single domain proteins categorized as  $\alpha$ ,  $\beta$ ,  $\alpha + \beta$ , or  $\alpha/\beta$ .<sup>26</sup> We utilized Astral 1.75A (40%) to ensure that redundancy in the data set did not lead to inflated assessment of performance. Table 1 summarizes the positive and negative data points used for each category of SCOP.

In the second experiment we attempted to distinguish enzymes, or those proteins annotated with the term “catalytic activity” and its subtypes, from all other proteins. Gene Ontology<sup>27</sup> (GO) annotations were obtained from the April 2012 release of Swiss-Prot<sup>28</sup> in conjunction with the May 4, 2012 version of GO. Only annotations supported by evidence codes EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC were used. This resulted in a total of 24,882 proteins with experimentally verified annotations, 9,506 of which were annotated with the term “catalytic activity” and 18,936 of which represented putatively negative data points.

We tested a range of combinations of values for  $n$  (window size), and  $m$  (number of centroids) for each property. For values of  $n$ , we tested  $n \in \{2^i : i = 1 \dots 5\}$ . Similarly, for the number of centroids,  $m$ , we tested  $m \in \{2^i : i = 4, 6, 8, 10, 12\}$ . For each property type and all values of  $m$  and  $n$ , we performed k-means clustering using  $10^6$  randomly sampled vectors from all sequences in  $\mathcal{S}$ . SVM<sup>light</sup> with the default value for the capacity parameter was used as a prediction engine in all experiments.<sup>29</sup> In each experiment, the total costs of misclassification for positive and negative examples were equal.

Table 1. Summary of data used for SCOP classification documenting the number of positives and negatives used for the classification of protein structures as  $\alpha$ ,  $\beta$ ,  $\alpha + \beta$ , or  $\alpha/\beta$ .

SCOP class	positives	negatives
all $\alpha$	1,901	7,486
all $\beta$	2,175	7,212
$\alpha + \beta$	2,665	6,722
$\alpha/\beta$	2,646	6,741

### 3.1. Mapping proteins into property vectors

Several structure-based properties were generated by converting the atomic 3D coordinates into backbone angles. The usefulness of representing a structure in this manner is that backbone angles are invariant to the translation and rotation of the original 3D coordinates. Four types of backbone angles were utilized:  $\alpha$ ,  $\kappa$ ,  $\phi$ , and  $\psi$ . All angles were obtained using DSSP.<sup>30</sup> In addition to generating dihedral angles, DSSP also outputs solvent accessibility values which we used as the fifth structure-based property.

We also generated seven sequence-based properties for both the task of categorizing structures and predicting function. These features were generated in order to represent biologically



Table 2. Optimal performance, according to  $AUC$ , of each property-based feature when predicting SCOP folds. For each property feature the combination of  $m$  and  $n$  values that obtained the highest  $AUC$  for an individual SCOP category are reported. The last block of columns shows the weighted  $AUC$ ,  $AUC^w$ , obtained across all SCOP categories. Results when combining all sequence- and structure-based features are shown in the bottom section of the table. Because different combinations of  $m$  and  $n$  were used when combining properties, these values are not shown for the Sequence + Structure and Sequence + Structure + String kernel combination of features.

Property\Category	All $\alpha$			All $\beta$			$\alpha/\beta$			$\alpha + \beta$			$AUC^w$	
	$m$	$n$	$AUC$	$m$	$n$	$AUC$	$m$	$n$	$AUC$	$m$	$n$	$AUC$	$m$	$n$
$\alpha$ angles	4,096	8	0.994	4,096	8	0.982	4,096	16	0.967	4,096	32	0.829	-	-
$\kappa$ angles	4,096	16	0.995	4,096	16	0.986	4,096	16	0.978	4,096	32	0.903	-	-
$\phi$ angles	4,096	16	0.990	4,096	16	0.975	4,096	16	0.964	4,096	32	0.809	-	-
$\psi$ angles	4,096	8	0.991	4,096	16	0.981	4,096	16	0.970	4,096	32	0.850	-	-
Solvent accessibility	4,096	8	0.960	4,096	8	0.951	4,096	32	0.914	4,096	32	0.710	-	-
B-factor predictions	256	8	0.790	256	8	0.709	64	8	0.809	16	4	0.587	-	-
Helix predictions	16	4	0.868	16	8	0.871	64	16	0.842	64	32	0.637	-	-
Hydrophobicity indices	256	4	0.711	1,024	4	0.728	64	2	0.834	64	2	0.585	-	-
Loop predictions	256	16	0.872	64	8	0.848	16	32	0.821	1,024	32	0.607	-	-
PDB disorder predictions	1,024	8	0.842	64	4	0.849	64	4	0.819	256	32	0.591	-	-
Sheet predictions	64	2	0.904	64	8	0.853	256	16	0.836	64	32	0.641	-	-
VSL2B disorder predictions	16	2	0.699	64	32	0.628	64	8	0.812	64	2	0.589	-	-
String kernel	-	2	0.863	-	3	0.878	-	4	0.860	-	5	0.634	-	-
Sequence	64	16	0.915	64	16	0.876	64	16	0.880	64	16	0.621	64	16
Structure	4,096	32	0.992	4,096	32	0.983	4,096	32	0.979	4,096	32	0.903	4,096	32
Sequence + Structure	-	-	0.989	-	-	0.967	-	-	0.970	-	-	0.851	-	-
Sequence + Structure + String kernel	-	-	0.989	-	-	0.967	-	-	0.970	-	-	0.851	-	-

relevant properties associated with a region of a protein sequence: (i) hydrophobicity, calculated using the Kyte-Doolittle scale<sup>31</sup> in a sliding window of length  $w = 11$ ; (ii) flexibility, calculated as predicted B-factors using our previous model;<sup>32</sup> secondary structure predictions of (iii) helix, (iv) sheet and (v) loop propensities using our in-house predictor; and intrinsic disorder, (vi) using the previously published VSL2B model<sup>33</sup> as well as (vii) predictions from the same in-house predictor used for secondary structures.

### 3.2. Performance evaluation

We performed 10-fold cross-validation in all experiments. For each binary classification task we calculated the area under the Receiver Operating Characteristic (ROC) curve ( $AUC$ ). While we evaluated each feature type for a combination of window size and number of clusters on each classification task separately, we also desired to obtain a single value that could be used to benchmark each combination of parameters on all evaluated SCOP classes. To do this we used a weighted average of  $AUC$  values across multiple one-versus-all classification tasks, where the weight for each task was calculated using the ratio of structures in the given category and the total number of structures. We refer to this performance measure as  $AUC^w$ .

We also calculated the signal-to-noise ratio ( $SNR$ ) obtained when encoding and decoding property-based representations of proteins using vector quantization. Given an original property vector  $\mathbf{p}$  and the reconstructed version of this vector  $\hat{\mathbf{p}}$ , the signal-to-noise ratio was calculated using the logarithmic decibel scale as

$$SNR_{dB}(\mathbf{p}, \hat{\mathbf{p}}) = \log_{10} \frac{\sum_{i=1}^{\ell} p_i^2}{\sum_{i=1}^{\ell} (p_i - \hat{p}_i)^2}.$$

On this scale one decibel signifies that the noise (or sum of squared differences between the original and reconstructed signals) represents  $1/10$ -th of the signal.

## 4. Results

### 4.1. Prediction of structural categories

Table 2 shows the performance of each property kernel when predicting SCOP classes. Among structure-based properties we found that  $\kappa$  angles had the best performance, both for individual SCOP classes and in terms of its  $AUC^w$  (0.961). Solvent accessibility values performed the worst out of structure-based properties, obtaining the lowest  $AUC$  for all SCOP classes as well as lowest  $AUC^w$  (0.868). All structure-based properties outperformed sequence-based properties.

Among sequence-based properties, predicted secondary structures performed the best, especially predictions that a residue is in a helix ( $AUC^w = 0.788$ ), a sheet ( $AUC^w = 0.787$ ) or a loop ( $AUC^w = 0.771$ ). Calculated hydrophobicity and VSL2B-based predictions of disorder propensity performed the worst ( $AUC^w = 0.707$  and  $AUC^w = 0.682$ , respectively). Interestingly, the predictor of disordered residues developed from PDB performed considerably better than VSL2B ( $AUC^w = 0.764$  vs.  $AUC^w = 0.682$ ), an outcome that may be due to the differences in training samples between the two models.

In order to test the predictive ability of integrating multiple properties, we implemented a linear combination of individual property kernels (Table 2, Figure 3). To reduce the computational complexity of this task we only combined properties utilizing  $m$  and  $n$  values that achieved the highest  $AUC^w$  for each property type.

We observed an improvement of about three percentage points when combining sequence-based properties, achieving an  $AUC^w$  of 0.813 compared to the best performance of an individual sequence based property of 0.788 (helix predictions). The combined kernel for structure-based properties saw no improvement over the best performing individual model ( $AUC^w = 0.961$  for both the combined kernel and  $\kappa$  angles), and actually exhibited lower performance when both sequence and structure-based properties were combined ( $AUC^w = 0.939$ ).

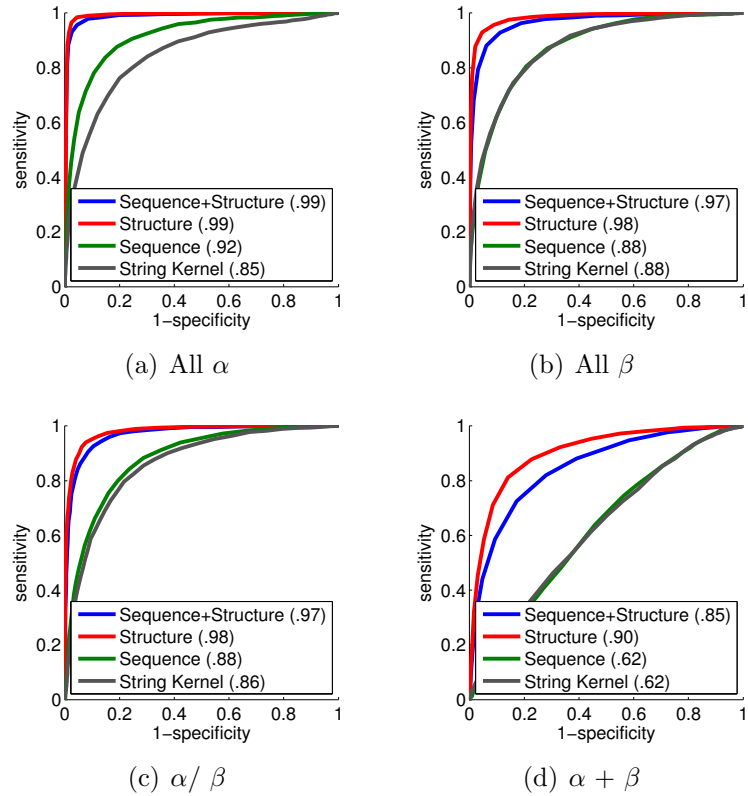


Fig. 3. ROC curves showing classification performance on SCOP classes.

#### 4.2. Prediction of protein function

The performance of each property kernel when predicting whether a protein is annotated with the catalytic activity term is shown in the first column of Table 3. Here we found that disorder-based predictions (VSL2B disorder:  $AUC = 0.742$ ; PDB disorder:  $AUC = 0.718$ ) and predicted B-factors ( $AUC = 0.722$ ) performed the best, whereas, contrary to their performance in distinguishing between SCOP categories, predicted secondary structures performed the worst (helix:  $AUC = 0.687$ ; sheet:  $AUC = 0.681$ ; loop:  $AUC = 0.698$ ).

Table 3 also shows the aggregate performance of each property when predicting and testing on six subclasses of catalytic activity. As with predicting catalytic activity in general, predicted B-factors also performed well in predicting catalytic activity subclasses ( $AUC^w = 0.659$ ). Predicted secondary structures had mixed performance, with predicted loops obtaining a comparatively high  $AUC^w$  of 0.647, whereas helix and sheet predictions only achieved an  $AUC^w$  of 0.611 and 0.621, respectively.

We also generated a reduced redundancy data set of proteins with GO annotations in which the maximum pairwise sequence identity outputted by BLAST between any two sequences was

Table 3. Performance, according to  $AUC$  and  $AUC^w$ , of each property-based feature when predicting catalytic activity and catalytic activity subclass, respectively. For each property feature the combination of  $m$  and  $n$  values that obtained the highest  $AUC$  are reported.

Property\Category	Catalytic activity			Catalytic subclass		
	$m$	$n$	$AUC$	$m$	$n$	$AUC^w$
B-factors	256	32	0.722	–	–	0.659
Helix	256	8	0.687	–	–	0.611
Hydrophobicity	4,096	2	0.701	–	–	0.653
Loop	256	32	0.698	–	–	0.647
PDB disorder	256	32	0.718	–	–	0.650
Sheet	256	4	0.681	–	–	0.621
VSL2B disorder	256	16	0.742	–	–	0.620

Table 4. Performance, according to  $AUC$  and  $AUC^w$ , of each the string kernel and combination of properties when predicting catalytic activity and catalytic activity subclass respectively. Results are shown for the full (redundant) data set and the non-redundant 40% data set (NR40).

Property\Category	Catalytic activity						Catalytic subclass					
	Full data set			NR40			Full data set			NR40		
	$m$	$n$	$AUC$	$m$	$n$	$AUC$	$m$	$n$	$AUC^w$	$m$	$n$	$AUC^w$
String kernel	-	5	0.857	-	5	0.733	-	5	0.930	-	5	0.649
VQ kernel	256	16	0.776	256	16	0.775	4,096	32	0.767	4,096	32	0.583
VQ + String kernel	-	-	0.781	-	-	0.775	-	-	0.767	-	-	0.585

at most 40%. This non-redundant data set (NR40) was generated to estimate the performance of each property when, for a given query protein, there is no sequence that is both annotated and of a reasonable level of sequence similarity. As shown by Figure 4 and Table 4, the performance of the property kernels was unaffected by the reduction in sequence identities between pairs of proteins, whereas string kernel performance was reduced.

### 4.3. String kernel performance

The string kernel did not show superior performance to any of the property kernels (both based on sequence and structure data) when predicting SCOP categories, only obtaining an  $AUC^w$  of 0.794 compared to an  $AUC^w$  of 0.961 obtained by the combined structure kernel and  $AUC^w$  of 0.813 obtained by the combined sequence kernel.

The performance of the string kernel in the task of function prediction was influenced by data set redundancy. When using redundant data, we found that the string kernel outperformed sequence-based properties in both the task of predicting catalytic activity and its subclass ( $AUC^w$  of 0.857 and 0.930), respectively (Figure 4(a)). However, when the redundancy in the protein function data was removed, the relative performance between the string kernel and the vector quantization kernel has reversed. As shown by Figure 4(b) the combined sequence-based property kernel achieved an  $AUC$  of 0.775 compared to 0.733 for the string kernel approach. Interestingly, this trend did not hold for the subclasses of catalytic activity, potentially due to the reduced data set sizes used to train individual models.

#### 4.4. Optimal parameter values

We found that structure-based properties consistently preferred large numbers of centroids, obtaining maximum  $AUC$  at  $m = 4096$  for all structure-based properties and all classification tasks. Optimal window sizes were 8 or 16 amino acids for most SCOP classes. Sequence-based properties were less consistent in the best-performing values of  $m$  and  $n$ , covering a range of values for each feature and SCOP class.

There was very little variation in preferred values of  $m$  when predicting catalytic activity with all features aside from predicted hydrophobicity obtaining maximum  $AUC$  values at  $m = 256$ . There was much more variation in preferred window sizes with hydrophobicity obtaining smallest optimal window size of 2, and B-factor, loop and PDB disorder predictions preferring longer window sizes ( $n = 32$ ). Sequence based properties were much more consistent in the preferred values of  $m$  and  $n$  when predicting catalytic activity subclass, almost always favoring large values of  $m$  (4,096).

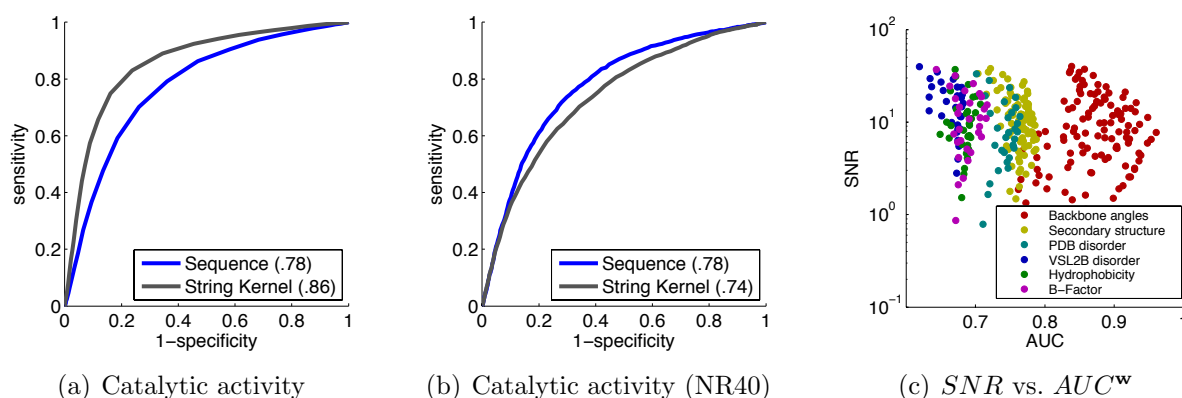


Fig. 4. Figure 4(a) shows ROC curves obtained when predicting catalytic activity using sequence properties (blue curve) and the string kernel (grey curve). AUC values are shown in parentheses in the figure legend. Figure 4(b) shows ROC curves obtained when predicting catalytic activity on the 40% non redundant data set of proteins (NR40) using sequence properties (blue) and the string kernel (grey). Figure 4(c) shows obtained  $SNR$  values plotted as a function of  $AUC^w$  values for the prediction of SCOP class for each feature type.

#### 4.5. Comparing AUC and SNR

Figure 4(c) shows a scatterplot of  $SNR$  and  $AUC^w$  values. Although, as a class, dihedral angles obtained higher values of  $AUC^w$ , these values were only weakly correlated with higher  $SNR$  values ( $\rho = 0.07$ ). For all other groups of properties in Figure 4(c) we observed a negative correlation between  $AUC^w$  and  $SNR$ .

### 5. Discussion

This paper introduced vector quantization (VQ) kernels and investigated their usefulness in different protein classification tasks. Several results show that the proposed kernel holds potential both as a standalone approach in protein classification and, more importantly, as a method that can be integrated into other strategies. The VQ kernel performed particularly

well in classification of SCOP classes, and as such could be readily exploited to automate the process of assigning new protein structures to structural classes. Such a method, similar to the FragBag approach,<sup>34</sup> is likely to be significantly faster than structure alignments that are commonly used for this purpose. Comparatively lower performance was observed in experiments relying on sequence-based properties only. Unsurprisingly, in these experiments, the property kernels outperformed string kernels when applied to non-redundant proteins, while they exhibited inferior performance to string kernels when high sequence identities were allowed.

The usefulness and biological significance of representing a protein sequence in a time series form has been long known. To the best of our knowledge, the use of a hydrophobicity plot (also referred to as hydropathy profile) was introduced by Rose who suggested that the local maxima and minima in the hydropathy profile typically correspond to the hydrophobic core and turns, respectively, in a protein's structure.<sup>35</sup> This idea quickly evolved into a tool for analysis of general properties of proteins, such as globular conformations<sup>36</sup> or membrane-spanning domains.<sup>31</sup> Advanced methods, such as the alignment of hydrophobic profiles<sup>37</sup> and Fast Fourier Transform (FFT) kernel<sup>17</sup> approach, have been proposed more recently, both in the context of recognizing membrane proteins.

The FFT kernel method is most related to the VQ kernels introduced here. In this method, Lanckriet and colleagues<sup>17</sup> first apply a low-pass filter to the original hydropathy profiles, pad the shorter profile with zeros (if the profiles are of different lengths), and subsequently calculate the kernel value between two FFT-derived spectra using a Gaussian kernel function with a free parameter  $\sigma$ . While this method provided solid performance in the task of predicting membrane proteins, we believe the kernel method introduced here offers better interpretability of results (through the selection and analysis of centroids) and more room for further refinements. For example, the simple inner product function between the count vectors  $k(\mathbf{p}, \mathbf{q}) = \mathbf{x}^T \mathbf{y}$  can be augmented by a positive semi-definite matrix  $\mathbf{Q}$  into a more general form  $\mathbf{x}^T \mathbf{Q} \mathbf{y}$ , perhaps by defining  $\mathbf{Q}$  through a non-singular matrix of similarities between centroids ( $\mathbf{S}$ ) and using  $\mathbf{Q} = \mathbf{S}^T \mathbf{S}$ . In addition, the centroid selection can be combined with motif discovery in time series data.<sup>20</sup> In terms of time complexity, the FFT kernel can be computed in  $O(\ell \log \ell)$  time compared to  $O(\ell \log m)$  time for the VQ kernel, where  $\ell$  is the length of the protein and  $m$  the number of clusters. The VQ kernel may also hold promise to more easily integrate multiple types of properties and exploit their correlation via a joint clustering or some form of "matrix quantization".

In summary, the VQ kernel introduced in this work is a robust methodology that can easily be extended to any type of data that is or can be transformed into a time-series.

## Acknowledgments

This work was funded by the National Science Foundation grant DBI-0644017.

## References

1. P. Radivojac *et al.*, *Nat Methods* **10**, 221 (2013).
2. Y. Moreau and L.-C. Tranchevent, *Nat Rev Genet* **13**, 523 (2012).

3. B. Schölkopf, K. Tsuda and J.-P. Vert (eds.), *Kernel methods in computational biology* (The MIT Press, 2004).
4. W. S. Noble, Support vector machine applications in computational biology, in *Kernel methods in computational biology*, eds. B. Schölkopf, K. Tsuda and J.-P. Vert (The MIT Press, 2004) pp. 71–92.
5. C. Leslie *et al.*, *Pac Symp Biocomput* **575**, 564 (2002).
6. J. Qiu *et al.*, *Bioinformatics* **23**, 1090 (2007).
7. J.-P. Vert, *Bioinformatics* **18**, S276 (2002).
8. K. M. Borgwardt *et al.*, *Bioinformatics* **21**, i47 (2005).
9. A. Ben-Hur and W. S. Noble, *Bioinformatics* **21**, i38 (2005).
10. T. De Bie *et al.*, *Bioinformatics* **23**, i125 (2007).
11. L. Ralaivola *et al.*, *Neural Netw* **18**, 1093 (2005).
12. V. Vacic *et al.*, *J Comput Biol* **17**, 55 (2010).
13. T. Jaakkola *et al.*, Using the Fisher kernel method to detect remote protein homologies, in *Proc Int Conf Intell Syst Mol Biol, ISMB*, 1999.
14. T. Jaakkola *et al.*, *J Comput Biol* **7**, 95 (2000).
15. R. Kuang *et al.*, Profile-based string kernels for remote homology detection and motif extraction, in *Proc IEEE Computat Syst Bioinform Conf, CSB*, 2004.
16. S. V. N. Vishwanathan *et al.*, *J Mach Learn Res* **11**, 1201 (2010).
17. G. R. G. Lanckriet *et al.*, *Bioinformatics* **20**, 2626 (2004).
18. A. Sokolov and A. Ben-Hur, *J Bioinform Comput Biol* **8**, 357 (2010).
19. J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis* (Cambridge University Press, 2004).
20. P. Patel *et al.*, Mining motifs in massive time series databases, in *Proc IEEE Int Conf Data Mining, ICDM*, 2002.
21. Y. Linde *et al.*, *IEEE Trans Commun* **28**, 84 (1980).
22. A. Gersho and R. M. Gray, *Vector quantization and signal compression* (Kluwer Academic Publishers, 1992).
23. T. Tuytelaars and C. Schmid, Vector quantizing feature space with a regular lattice, in *Proc IEEE Int Conf Computer Vision, ICCV*, 2007.
24. G. Strang, *Introduction to linear algebra* (Wellsley-Cambridge Press, 2003).
25. M. Gönen and E. Alpaydin, *J Mach Learn Res* **12**, 2211 (2011).
26. A. G. Murzin *et al.*, *J Mol Biol* **247**, 536 (1995).
27. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).
28. A. Bairoch *et al.*, *Nucleic Acids Res* **33**, D154 (2005).
29. T. Joachims, *Learning to classify text using support vector machines: methods, theory, and algorithms* (Kluwer Academic Publishers, 2002).
30. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
31. J. Kyte and R. F. Doolittle, *J Mol Biol* **157**, 105 (1982).
32. P. Radivojac *et al.*, *Protein Sci* **13**, 71 (2004).
33. K. Peng *et al.*, *BMC Bioinformatics* **7**, 208 (2006).
34. I. Budowski-Tal *et al.*, *Proc Natl Acad Sci U S A* **107**, 3481 (2010).
35. G. D. Rose, *Nature* **272**, 586 (1978).
36. G. D. Rose and S. Roy, *Proc Natl Acad Sci U S A* **77**, 4643 (1980).
37. J. S. Lolkema and D.-J. Slotboom, *FEMS Microbiol Rev* **22**, 305 (1998).

# COMBINING HETEROGENOUS DATA FOR PREDICTION OF DISEASE RELATED AND PHARMACOGENES

CHRISTOPHER S. FUNK\*, LAWRENCE E. HUNTER, and K. BRETONNEL COHEN

*Computational Bioscience Program, University of Colorado School of Medicine,  
Aurora, CO 80045, USA*

*\*E-mail: Christopher.Funk@ucdenver.edu, Larry.Hunter@ucdenver.edu, Kevin.Cohen@gmail.com*

Identifying genetic variants that affect drug response or play a role in disease is an important task for clinicians and researchers. Before individual variants can be explored efficiently for effect on drug response or disease relationships, specific candidate genes must be identified. While many methods rank candidate genes through the use of sequence features and network topology, only a few exploit the information contained in the biomedical literature. In this work, we train and test a classifier on known pharmacogenes from PharmGKB and present a classifier that predicts pharmacogenes on a genome-wide scale using only Gene Ontology annotations and simple features mined from the biomedical literature. Performance of  $F=0.86$ ,  $AUC=0.860$  is achieved. The top 10 predicted genes are analyzed. Additionally, a set of enriched pharmacogenic Gene Ontology concepts is produced.

## 1. Introduction

One of the most important problems in the genomic era is identifying variants in genes that affect response to pharmaceutical drugs. Variability in drug response poses problems for both clinicians and patients.<sup>1</sup> Variants in disease pathogenesis can also play a major factor in drug efficacy.<sup>2,3</sup> However, before variants within genes can be examined efficiently for their effect on drug response, genes interacting with drugs or causal disease genes must be identified. Both of these tasks are open research questions.

Databases such as DrugBank<sup>4</sup> and The Therapeutic Target DB<sup>5</sup> contain information about gene-drug interactions, but only The Pharmacogenomics Knowledgebase (PharmGKB)<sup>6</sup> contains information about how variation in human genetics leads to variation in drug response and drug pathways. Gene-disease variants and relationships are contained in Online Mendelian Inheritance in Man (OMIM),<sup>7</sup> the genetic association database,<sup>8</sup> and the GWAS catalog.<sup>9</sup> Curated databases are important resources, but they all suffer from the same problem: they are incomplete.<sup>10</sup> One approach to this problem is the development of computational methods to aid in database curation. We explore here a method that takes advantage of the large amount of information in the biomedical literature that is waiting to be exploited.

Having a classifier that is able to predict as-yet-uncurated pharmacogenes would allow researchers to focus on identifying the variability within the genes that could affect drug response or disease, and thus, shorten the time until information about these variants is useful in a clinical setting. (We use the term “pharmacogene” to refer to any gene such that a variant has been seen to affect drug response or is implicated in a disease.) Computational methods have been developed to predict the potential relevance of a gene to a query drug.<sup>11</sup> Other computational methods have been developed to identify genetic causes underlying disorders through gene prioritization, but many of these are designed to work on small sets of disease-specific genes.<sup>12–17</sup> The method which is closest to the one that we present here is described in



Costa *et al.*<sup>18</sup> they create separate classifiers to predict morbidity-associated and druggable genes on a genome-wide scale. A majority of these methods use sequence-based features, network topology, and other features from curated databases; only a few use information from literature.<sup>12,16,17</sup>

In the work presented here, the goal is to predict pharmacogenes at genome-wide scale using a combination of features from curated databases and features mined from the biomedical literature. We evaluate a number of hypotheses:

- (1) There is a set of GO concepts that are enriched when comparing the functions of important pharmacogenes and the rest of the human genome and by examining this set of enriched GO concepts, a classifier can be created to provide hypotheses regarding further genes in which variants could be of importance.
- (2) Text-mined features will increase performance when combined with features from curated databases.

## 2. Methods

### 2.1. *Pharmacogenes*

By *pharmacogene*, we mean any gene such that a variant of that gene has been seen to affect drug response or such that variants have been implicated in disease. PharmGKB contains over 26,000 genes, with only a few having annotations that signify their importance in disease or drug response. For the experiments reported here, only those genes in which a variant exists in the PharmGKB relationship database, specifically gene-disease or gene-drug relationships, are considered to be gold-standard pharmacogenes. By this definition, 1,124 genes meet the criteria for classification as pharmacogenes and are positively labeled training instances; these make up <5% of all genes in PharmGKB. PharmGKB is constantly being updated, so a snapshot of PharmGKB on May 2, 2013 was taken and is used as the gold standard.

### 2.2. *Background genes*

The rest of the 25,110 genes in PharmGKB, which do not contain disease or drug relationships, are considered to be background genes and will be used as negatively labeled training instances. We acknowledge the fact that PharmGKB is incomplete and that a missing annotation is not indicative of a gene not being involved in disease or drug relationships, but the fact that they have not been discovered or curated yet. (This is an obvious motivation for the work reported here.) Two data sets were created from the background genes. One consists of all 25,110 genes. This is referred to as the unbalanced set. The second consists of 1,124 background genes that have similar numbers of publications as the known pharmacogenes. This is referred to as the balanced set. That is, the two sets differ in whether or not they result in a balanced set of positive and negative exemplars.

### 2.3. *Functional annotations from curated databases*

Links within PharmGKB were used to obtain Entrez Gene (EG) identifiers for both pharmacogenes and background genes. To extract all Gene Ontology (GO)<sup>19</sup> annotated functions

associated with these genes, the NIH’s gene2go file was used. Only curated evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, and ISS) were used, in order to ensure high-quality annotations. This dataset will be referred to as the curated dataset. It contains many EGID to GO ID mappings obtained solely from curated GO annotations.

## 2.4. Functional annotations from biomedical literature

Entrez Gene IDs and the NIH’s gene2pubmed file were used to relate genes to documents of which they are the primary subject. By using the gene2pubmed file, we assume that all information retrieved from the article is associated with the gene that is the primary subject. Note that this is not always true and could introduce noise.

The 26,234 genes are mapped to 379,978 unique PubMed/MEDLINE articles. From these ~380,000 articles, two different textual datasets were created, one consisting only of abstracts and the other containing full text. The abstract dataset consists of all abstracts from all articles. For ~26,000 articles, we were only able to download XML or plain text, because PMC articles are available in any format, with some, such as PDF, not being suitable for natural language processing. The ~26,000 full-text articles constitute our full-text dataset. All full-text documents come from the PubMed Open Access Subset.

To extract gene functions (GO concepts) from these corpora, ConceptMapper, a dictionary-based concept recognizer,<sup>20</sup> was used with parameters tuned for each branch of the Gene Ontology (Molecular Function, Biological Process, and Cellular Component), as seen in Funk *et al.* (under review). Descriptive statistics of the documents and the functional annotations retrieved from them and from the curated database are shown in Table 1.

Table 1. **Summary of gene-document and gene-annotation associations** The number of genes within each dataset along with the mean number of biomedical literature documents associated with each **set of genes** and mean number of GO annotations per gene. (+) denotes that this set of genes is the positive labeled set while (–) denotes the negative training sets. The row labelled “Total Numbers” gives the count, not means, of documents and GO annotations.

	Mean # Docs			Mean # GO Annotations		
	# Genes	Abstracts	Full-text	GOA curated	NLP abstracts	NLP full-text
All genes	26,234	35.5	3.1	8.8	80.1	122.0
Known pharmacogenes (+)	1,124	215.2	15.5	16.3	227.5	220.7
All background genes (–)	25,110	26.7	2.5	8.2	72.8	128.7
Small background genes (–)	1,124	211.1	17.1	20.4	310.0	298.9
Total Numbers	26,234	379,978	25,987	112,356	1,891,566	1,951,982

## 2.5. Enrichment of Gene Ontology concepts

FatiGO<sup>21</sup> was used to test whether there are functional concepts that are enriched when pharmacogenes are compared to background genes. FatiGO is a tool that uses Fisher’s exact test to extract over- or under-represented GO concepts from two lists of genes and provides a list of enriched GO concepts and their respective p-values as output. The p-values are corrected for multiple testing as described in Ge *et al.*<sup>22</sup> The gene lists and all three sets of annotations—curated, and text-mined—were provided to FatiGO as custom annotations. Fisher’s exact test was conducted between GO concepts annotated to pharmacogenes and

those annotated to background genes for all three sets of Gene Ontology concepts (curated, mined from abstracts, and mined from full text).

## 2.6. Binary Classification

All classifiers were implemented in the Weka toolkit, version 3.6.9. Three different base-lines were used: OneR, a one node decision tree; Naive Bayes; and randomly assigning class labels. Against these, we compared three systems: Random Forests and two different Support Vector Machine implementations. Random Forests provide fast decision-tree training. Support Vector Machines (SVM) are currently the most popular classifier. The built-in classifiers for OneR (`weka.classifiers.rules.OneR`), Naive Bayes (`weka.classifiers.bayes.NaiveBayes`), Random Forest (`weka.classifiers.trees.RandomForest`), and Support Vector Machine (`weka.classifiers.functions.SMO`) were used with default parameters. LibSVM (`weka.classifiers.functions.LibSVM`) was used with all but one default parameter. By default LibSVM maximizes accuracy; with the unbalanced dataset, this is not optimal, so weights of 90.0 and 10.0 were assigned to the pharmacogene and background classes, respectively. When using LibSVM with the balanced dataset, equal weights were given to both classes. All numbers reported are from five-fold cross-validation.

Table 2. **Machine learning features per dataset** A breakdown of the number and type of features used.

Dataset	# Genes	# Features	Type
GOA curated	12,704	39,329	Curated GO annotations from the GOA database.
NLP abstract	23,849	39,329	GO annotations recognized from MEDLINE abstracts.
NLP full-text	15,168	39,329	GO annotations recognized from full-text journal articles.
Abstract GO + Bigrams	23,849	858,472	GO annotations and bigrams from MEDLINE abstracts.
Full-text GO + Bigrams	15,168	906,935	GO annotations and bigrams from full-text journal articles.
Combined GO + Bigrams	23,867	1,189,175	Curated and NLP GO annotations and all bigrams.
Abstract GO + Collocations	23,849	346,878	GO annotations and collocations from MEDLINE abstracts.
Full-text GO + Collocations	15,168	54,951	GO annotations and collocations from full-text journal articles.
Combined GO + Collocations	23,867	349,243	Curated and NLP GO annotations and all collocations.

## 2.7. Features derived from natural language processing

Additional features were extracted from the abstract and full-text document collections using natural language processing. (This is in addition to the automatically extracted Gene Ontology annotations, which are also produced by natural language processing.) These features were word bigrams and collocations. Collocations, or sets of words that co-occur more often than expected, have not been commonly used in text classification, but provide a better reflection of the semantics of a text than bigrams. Both bigrams and collocations were extracted using the Natural Language Tool Kit (NLTK).<sup>23</sup> Any bigram or collocation where one of the tokens only contained punctuation was removed. Additionally, only those features that appear in three or more documents were retained. Six different NLP-derived feature sets were created by combining the three datasets (abstract, full-text, curated + abstract + full-text) along with the two different types of surface linguistic features (bigrams and collocations); these feature sets were tested and trained on both the balanced and unbalanced datasets.

## 2.8. *Machine learning input*

A breakdown of the kind and number of features used in each dataset can be seen in Table 2.

## 2.9. *Evaluation metrics*

The performance of our classifier was assessed by estimating precision (P), recall (R), and F-measure (F). The area under the receiving operator characteristic curve (AROC) is reported, as it allows for comparison against other classifiers, but with a word of caution interpreting the unbalanced dataset: inflated AROCs have been seen when working with skewed class distributions.<sup>24</sup> All scores were determined by taking the average of 5-fold cross-validation for all datasets.

# 3. Results and Discussion

## 3.1. *Enriched Gene Ontology concepts*

To assess the viability of a machine learner separating background and pharmacogenes, we first determine whether functional differences between the pharmacogenes and background genes exist. At least one curated or text-mined functional annotation was retrieved for 23,647 out of 26,236 total genes (90% of all genes in PharmGKB). The details of obtaining the annotations are given in Sections 2.3 and 2.4. The gene sets and their annotations were passed to FatiGO, a web tool that extracts over- and under-represented GO concepts from two lists of genes, and a list of enriched GO concepts and probabilities was returned as output. Examining the output from FatiGO, we found that, depending on the dataset, between 800-4000 GO concepts were enriched, consistent with our hypothesis that there are enriched pharmacogenetic functions. The top 10 enriched GO concepts for Molecular Function and Biological Process can be seen in Tables 3 and 4, respectively. These lists were obtained by comparing the annotations from all pharmacogenes to all background genes. To ensure that bias was not introduced solely because there is a large difference in the number of genes and the number of annotations between the two sets, another comparison was done between all pharmacogenes and the set of 1,124 background genes with equal representation in the biomedical literature. The enriched GO concepts returned are similar the concepts returned when comparing against all background genes, and therefore we can conclude that no bias is introduced. Because 800-4000 statistically enriched GO concepts were returned for each dataset, we can conclude that there are functional differences between the set of pharmacogenes and background genes.

Many of the enriched GO concepts can be categorized as playing a role in pharmacodynamics (PD) or pharmacokinetics (PK). Pharmacodynamics is the study of the activity of a drug in the body, e.g. its binding and effect on the body. Examples of PD concepts are “integral to plasma membrane” (GO:0005887), “drug binding” (GO:0008144), and “positive regulation of protein phosphatase type 2B activity” (GO:0032514)—they are either associated with receptors that drugs bind to, or refer to the possible effect that a drug has on the body. Pharmacokinetics is the study of drug absorption, distribution, metabolism, and excretion. Examples of PK concepts are “xenobiotic metabolic process” (GO:0006805), “small molecule metabolic process” (GO:0044281), and “active transmembrane transporter activity”

(GO:0022804)—they refer to metabolism of a molecule or are involved in the metabolism or transportation of a molecule.

Table 3. **Top 10 enriched GO concepts from the Molecular Function hierarchy.** The enriched GO concepts from the Molecular Function branch of Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

GOA curated		
Concept ID	Concept name	Adj. P-value
GO:0005515	protein binding	$< 1.0 \times 10^{-8}$
GO:0019899	enzyme binding	$< 1.0 \times 10^{-8}$
GO:0042803	protein homodimerization activity	$< 1.0 \times 10^{-8}$
GO:0046982	protein heterodimerization activity	$< 1.0 \times 10^{-8}$
GO:0004497	monooxygenase activity	$< 1.0 \times 10^{-8}$
GO:0005245	voltage-gated calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0020037	heme binding	$< 1.0 \times 10^{-8}$
GO:0004713	protein tyrosine kinase activity	$< 1.0 \times 10^{-8}$
GO:0004674	protein serine/threonine kinase activity	$< 1.0 \times 10^{-8}$
GO:0003677	DNA binding	$< 1.0 \times 10^{-8}$

NLP abstracts		
Concept ID	Concept name	Adj. P-value
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005322	low-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005321	high-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005320	apopliprotein	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0005041	low-density lipoprotein receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0004697	protein kinase C activity	$< 1.0 \times 10^{-8}$
GO:0045289	luciferin monooxygenase activity	$< 1.0 \times 10^{-8}$

NLP full-text		
Concept ID	Concept name	Adj. P-value
GO:0042031	angiotensin-converting enzyme inhibitor activity	$< 1.0 \times 10^{-8}$
GO:0005262	calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0004872	receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016791	phosphatase activity	$< 1.0 \times 10^{-8}$
GO:0008083	growth factor activity	$< 1.0 \times 10^{-8}$
GO:0004601	peroxidase activity	$< 1.0 \times 10^{-8}$

There are interesting differences when examining the top enriched concepts between the different datasets (curated, abstracts, and full text). Impressionistically, curated annotations seem to be more specific, while NLP annotations appear to be more general (especially evident when examining Biological Processes, Table 4). This may be the case because there are limitations to the depth in GO that concept recognizers can identify; a large gap exists between how near-terminal concepts are stated in the ontology and their expression in free text.

### 3.2. Classification of pharmacogenes

Having established that the functions of pharmacogenes are different from background genes, the next step is to test the ability of machine learning to differentiate between them. Our goal is to predict at genome-wide scale pharmacogenes that

are not currently known in PharmGKB to have drug or disease relationships. We approach the problem as binary classification, where the classifier separates pharmacogenes from the rest of the genes.

### 3.3. Classification using Gene Ontology concepts

To see how well known pharmacogenes can be classified through their functional annotation similarity, five classifiers were created using the manually curated and text-mined functional annotations on both the unbalanced and balanced datasets. Baselines for comparison against are a one-node decision tree (OneR), Naive Bayes, and randomly assigning class labels. Performance of all classifiers and baselines can be seen in Table 5. A breakdown of features used

for each dataset can be seen in Table 2 and a summary of functional annotations is seen in Table 1.

The results are shown in Table 5. A clear effect of balance versus imbalance in the data is evident. F-measure increases between 0.29 and 0.53 when using a balanced training set. Examining performance across unbalanced training sets, we notice that Naive Bayes produces the highest recall (0.68) but the lowest precision (0.17), whereas Random Forest produces highest precision (0.69) but lowest recall (0.11). The same trends do not hold for the balanced training sets. On both training sets, it is the SVM-based classifiers that balance precision and recall and produce the highest F-measures. The highest F-measures of 0.81 and 0.78, are produced by LibSVM and SMO, respectively, on the balanced NLP abstract annotations. Naive Bayes and Random Forrest perform poorly in comparison to the SVM classifiers, but better than a single-node decision tree or random assignment; OneR performs slightly better than random assignment.

For a majority of the classifiers, GO annotations from literature produce the best performance—surprisingly, text-mined annotations seem to be better features than those from curated datasets. This could be explained by the difference in number of annotations, there are 15 times more text-mined annotations than curated ones (Table 1). Another explanation could be that more information is encoded in text-mined annotations than just gene function.

From this set of experiments, we can conclude that using only Gene Ontology concepts, we are able to classify pharmacogenes on the balanced training set but it remains unclear, because of poor performance, whether it is sufficient to use only GO concepts with an unbalanced training set. We can also conclude that LibSVM should be used for the next set of experiments

Table 4. **Top 10 enriched GO concepts from the Biological Process hierarchy.** The enriched GO concepts from the Biological Process branch of the Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

GOA curated		
Concept ID	Concept name	Adj. P-value
GO:0044281	small molecule metabolic process	$< 1.0 \times 10^{-8}$
GO:0007596	blood coagulation	$< 1.0 \times 10^{-8}$
GO:0030168	platelet activation	$< 1.0 \times 10^{-8}$
GO:0006805	xenobiotic metabolic process	$< 1.0 \times 10^{-8}$
GO:0048011	neurotrophin TRK receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007268	synaptic transmission	$< 1.0 \times 10^{-8}$
GO:0008543	fibroblast growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007173	epidermal growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0045087	innate immune response	$< 1.0 \times 10^{-8}$
GO:0055085	transmembrane transport	$< 1.0 \times 10^{-8}$
NLP abstracts		
Concept ID	Concept name	Adj. P-value
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0006629	lipid metabolic process	$< 1.0 \times 10^{-8}$
GO:0007610	behavior	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0014823	response to activity	$< 1.0 \times 10^{-8}$
GO:0006280	mutagenesis	$< 1.0 \times 10^{-8}$
GO:0042638	exogen	$< 1.0 \times 10^{-8}$
NLP full-text		
Concept ID	Concept name	Adj. P-value
GO:0009626	plant-type hypersensitive response	$< 1.0 \times 10^{-8}$
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0016311	dephosphorylation	$< 1.0 \times 10^{-8}$
GO:0032514	positive regulation of protein phosphatase type 2B activity	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0042592	homeostatic process	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0050817	coagulation	$< 1.0 \times 10^{-8}$

because it is best performing and was the fastest to train (training time not shown).

### 3.4. Classification using GO concepts and literature features

To test the hypothesis that features derived from surface linguistic features can increase performance over conceptual features alone, we trained classifiers with two additional feature types: bigrams and collocations. Bigrams consist of every sequence of two adjacent words in a document and are commonly used in text classification. Collocations are a subset of bigrams, containing words that co-occur more frequently than expected. They are a better representation of the semantics of a text than bigrams alone. The methods for extracting these features are described above in Section 2.7. Adding bigrams and collocations introduces up to 30x more features than functional annotations alone (Table 2).

The performance of LibSVM with GO annotations and bigrams/collocations on both training sets can be seen in Table 6. Baselines are the same.

Table 5. **Classification using Gene Ontology concepts**

Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

Classifier	GOA curated P/R/F	NLP abstracts P/R/F	NLP full-text P/R/F
<b>Unbalanced Training</b>			
Random	0.05/0.50/0.09	0.07/0.50/0.12	0.05/0.50/0.09
OneR	0.57/0.01/0.03	0.56/0.17/0.25	0.80/0.10/0.18
Naive Bayes	0.17/0.60/0.26	0.17/0.68/0.27	0.17/0.59/0.26
Random Forest	0.53/0.17/0.25	0.69/0.12/0.21	0.58/0.11/0.18
SMO	0.43/0.31/0.36	0.39/0.41/0.40	0.37/0.34/0.35
LibSVM	0.29/0.55/ <b>0.38</b>	0.41/0.58/ <b>0.48</b>	0.37/0.52/ <b>0.42</b>
<b>Balanced Training</b>			
Random	0.50/0.50/0.50	0.50/0.50/0.50	0/50/0.50/0.50
OneR	0.71/0.41/0.52	0.68/0.51/0.59	0.73/0.48/0.56
Naive Bayes	0.65/0.72/0.68	0.75/0.70/0.72	0.67/0.70/0.68
Random Forest	0.63/0.71/0.67	0.72/0.77/0.74	0.67/0.73/0.69
SMO	0.64/0.66/0.65	0.79/0.77/0.78	0.70/0.73/0.72
LibSVM	0.71/0.71/ <b>0.71</b>	0.83/0.80/ <b>0.81</b>	0.76/0.79/ <b>0.78</b>

On the unbalanced training set, the maximum F-measure seen is 0.57, obtained by using text-mined functional annotations and bigrams extracted from abstracts. By using bigrams in addition to GO annotations, precision is increased by 0.17 while recall is decreased by 0.02, resulting in an increase in F-measure of 0.09 (Table 5 versus Table 6). On the balanced training set, the maximum F-measure seen is 0.81, also obtained by using text-mined functional annotations and bigrams from abstracts. With the addition of bigrams, both precision and recall are increased by 0.06 and 0.03, respectively, resulting in an increase in F-measure of 0.06 (comparing Table 5 to Table 6).

#### 3.4.1. Comparison with other methods

As mentioned in the introduction, there are very few methods against which our method can be compared. Most gene-disease or gene prioritization methods are designed to work on small sets of disease-specific genes,<sup>12–14</sup> while our method predicts pharmacogenes on a genome-wide scale. One method, Garten *et al.*,<sup>25</sup> utilizes text mining to extract drug-gene relationships from the biomedical literature, also using PharmGKB as a gold standard, with an AUC of 0.701. The closest methods to ours do not predict pharmacogenes as defined here, but only predict disease genes. CIPHER<sup>26</sup> predicts human disease genes with precision of  $\sim 0.10$  using protein-protein interaction networks and gene-phenotype associations. PROSPECTR<sup>27</sup> uses

Table 6. **Classification with GO concepts and natural language processing** Five-fold cross-validation performance of LibSVM when combining Gene Ontology concepts and literature-based features. Both the balanced and unbalanced training results are shown. The highest F-measure and AROC are bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

Classifier	Abstract GO + Bigrams		Full-Text GO + Bigrams		Combined GO + Bigrams	
	P/R/F	AUC	P/R/F	AUC	P/R/F	AUC
<b>Unbalanced Training</b>						
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.58/0.56/ <b>0.57</b>	<b>0.771</b>	0.50/0.46/0.48	0.711	0.50/0.54/0.52	0.756
<b>Balanced Training</b>						
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.75/0.59/0.66	0.696	0.71/0.53/0.61	0.663	0.79/0.50/0.61	0.685
LibSVM	0.89/0.83/ <b>0.86</b>	<b>0.860</b>	0.79/0.82/0.80	0.807	0.86/0.83/0.85	0.848
Classifier	Abstract GO + Collocations		Full-Text GO + Collocations		Combined GO + Collocations	
	P/R/F	AUC	P/R/F	AUC	P/R/F	AUC
<b>Unbalanced Training</b>						
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.54/0.56/ <b>0.55</b>	<b>0.767</b>	0.41/0.52/0.46	0.730	0.47/0.56/0.51	0.763
<b>Balanced Training</b>						
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.78/0.46/0.58	0.664	0.67/0.64/0.66	0.675	0.75/0.59/0.66	0.698
LibSVM	0.87/0.82/ <b>0.85</b>	<b>0.850</b>	0.77/0.80/0.78	0.786	0.85/0.81/0.83	0.833

23 sequence-based features and predicts disease genes from OMIM with precision = 0.62 and recall = 0.70 with an AUC of 0.70. The most directly comparable method, presented in Costa *et al.*,<sup>18</sup> utilizes topological features of gene interaction networks to predict both morbidity genes (P=0.66, R=0.65, AUC=0.72) and druggable genes (P=0.75, R=0.78, AUC=0.82). While the majority of other methods utilize sequence-based features, protein interactions, and other genomic networks, our method requires only Gene Ontology annotations and simple bigrams/collocations extracted from biomedical literature. Precision and recall for our classifier trained on the unbalanced dataset with GO annotations and bigrams from abstracts are slightly lower than both PROSPECTR and the method presented in Costa *et al.*, our AUC (0.771) is higher than all but the predicted druggable genes from Costa *et al.* Performance on the balanced training set using GO concepts and bigrams extracted from abstracts (F=0.86, AUC=0.860) are higher than any of the methods presented here.

### 3.4.2. Limitations

There are two major limitations of our work. The first is that we grouped together all pharmacogenes, while it may have been more useful to differentiate between disease-associated and drug-response-associated variant. The other limitation is that we don't provide a ranking, but rather just a binary classification.

### 3.5. Prediction of pharmacogenes

Now that classifiers have been created and evaluated, we can analyze the predicted pharmacogenes. 141 genes were predicted to be pharmacogenes by all six unbalanced datasets seen in Table 6. Predictions from unbalanced models were analyzed because the models produced through balanced training were unknowingly weighted for recall. For example, the balanced model trained on abstract GO and bigrams produces a recall of 0.99 and precision of 0.10



Table 7. **Top 10 predicted pharmacogenes** Top 10 pharmacogenes predicted by all combined classifiers and ranked by functional similarity to the known pharmacogenes. All information from PharmGKB and OMIM is presented along with the class that was predicted by Costa *et al.*<sup>18</sup> (Morbid: mutations that cause human diseases, Druggable: protein-coding genes whose modulation by small molecules elicits phenotypic effects).

EG ID	Symbol	PharmGKB Annotations	OMIM Phenotype	Costa <i>et al.</i> <sup>18</sup> predicted
2903	<i>GRIN2A</i>	None	Epilepsy with neurodevelopment defects	Druggable
7361	<i>UGT1A</i>	None	None	Not tested
2897	<i>GRIK1</i>	None	None	Druggable
1128	<i>CHRM1</i>	None	None	Druggable
1131	<i>CHRM3</i>	Member of Proton Pump Inhibitor Pathway	Eagle-Barrett syndrome	Druggable
3115	<i>HLA-DPB1</i>	None	Beryllium disease	Morbid/Druggable
6571	<i>SLC18A2</i>	Member of Nicotine, Selective Serotonin Reuptake Inhibitor, and Sympathetic Nerve Pathway	None	Morbid/Druggable
477	<i>ATP1A2</i>	None	Alternating hemiplegia of childhood, Migraine (familial basilar and familial hemiplegic)	Morbid/Druggable
3643	<i>INSR</i>	Member of Anti-diabetic Drug Potassium Channel Inhibitors and Anti-diabetic Drug Repaglinide Pathways	Diabetes mellitus, Hyperinsulinemic hypoglycemia, Leprechaunism, Rabson-Mendenhall syndrome	Morbid/Druggable
2905	<i>GRIN2C</i>	None	None	Druggable

when the classifier is applied to all genes in PharmGKB; this is not informative and further work and error analysis will be conducted to examine why this is.

The top 10 predicted genes, ranked by functional similarity (as calculated by ToppGene) to the known pharmacogenes, along with all known information from PharmGKB and Online Mendelian Inheritance in Man (OMIM),<sup>7</sup> and if/what the gene was predicted to be by Costa *et al.* can be seen in Table 7. We first notice that there are no gene-disease or gene-drug relationships in PharmGKB for these predicted genes, but a few of them participate in curated pathways. We expand our search to see if other databases have drug or disease information about them. OMIM provides insight into genetic variation and phenotypes; half of the predicted genes have a variant that plays a role in a mutant phenotype. We also looked up our predicted genes in the results from a previous study on predicting morbid and druggable genes, and 90% (9 out of 10) of our predicted pharmacogenes were also predicted to be morbid (variations cause hereditary human diseases) or druggable.<sup>18</sup>

To assess the hypothesized pharmacogenes further, PubMed and STITCH<sup>28</sup> were used to find any known drug or disease associations not in PharmGKB or OMIM. The top-ranked gene, *GRIN2A*, seems to play a part in schizophrenia and autism spectrum disorders<sup>29</sup> along with binding to memantine, a class of Alzheimer's medication blocking glutamate receptors. Interestingly, *UGT1A* is unable to be found in STITCH or OMIM, but an article from May 2013 introduces a specific polymorphism that suggests that it is an important determinant of acetaminophen glucuronidation and could affect an individual's risk for acetaminophen-induced liver injury.<sup>30</sup> It is also known to be linked to irinotecan toxicity. We also find genetic variations in *GRIK1* have been linked to schizophrenia<sup>31</sup> and down syndrome.<sup>32</sup> Even only examining the top three predicted pharmacogenes, there is evidence in other databases and literature that suggests these should be further examined by the PharmGKB curators for possible annotation.

## 4. Conclusions

One of the surprising findings of this study was that features extracted from abstracts performed better than features extracted from full text. Since full text was available for a smaller number of genes, the comparison may not be appropriate. Pursuing this remains for further research.

The collocation features performed almost as well as the bigrams, despite the fact that we took a poor approach to extracting them, since we did collocation recognition on the document level, rather than on the level of the document collection as a whole. With a better approach to collocation extraction, performance of the collocation features might have been much higher.

The fact that features derived from text-mined functional annotations outperformed manually curated annotations was a surprise. In this work, we did not evaluate the correctness of text-mined functional annotations. Therefore, the performance of the text-mined functional annotation features is the only indication of how well the actual Gene Ontology concept recognition worked. Based on the fact that they performed higher than the manually curated Gene Ontology concepts, it appears that the performance of the ConceptMapper approach was at minimum good enough for this task.

In this paper we identified a set of functions enriched in known pharmacogenes. This list could be used to rank genes predicted by our classifier, but also has usefulness beyond the work presented here. The list could prove useful in literature-based discovery by providing linkages to identify gene-drug or gene-disease relationships from disparate literature sources.

We also present a classifier that is able to predict pharmacogenes at a genome wide scale ( $F=0.86$ ,  $AUC=0.860$ ). The top 10 hypothesized pharmacogenes predicted by our classifier are presented; 50% contain allelic variations in OMIM and 90% were previously predicted but remain unannotated in PharmGKB. Additionally, using other sources at least the top three genes predicted are known to bind a drug or to be associated with a disease. Other methods attempting similar problems, utilize sequence based features and genomic networks; only a few incorporate literature features. Our method, on the other hand, uses mainly features mined from the biomedical literature along with functional annotations from databases. Because our method offers comparable performance to others utilizing sequence and network based features, this work illustrates the importance of incorporating curated databases with information available in the biomedical literature for biomedical discovery.

## Acknowledgments

This work was supported by NIH grants 5R01 LM009254-07, 5R01 LM008111-08, and 2T15LM009451 to LEH.

## References

1. W. E. Evans and M. V. Relling, *Science* **286**, 487 (1999).
2. J. Poirier, M.-C. Delisle, R. Quirion, I. Aubert, M. Farlow, D. Lahiri, S. Hui, P. Bertrand, J. Nalbantoglu and B. M. Gilfix, *Proceedings of the National Academy of Sciences* **92**, 12260 (1995).
3. J. A. Kuivenhoven, J. W. Jukema, A. H. Zwinderman, P. de Knijff, R. McPherson, A. V. Bruschke, K. I. Lie and J. J. Kastelein, *New England Journal of Medicine* **338**, 86 (1998).

4. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic acids research* **34**, D668 (2006).
5. X. Chen, Z. L. Ji and Y. Z. Chen, *Nucleic acids research* **30**, 412 (2002).
6. M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman and T. E. Klein, *Nucleic acids research* **30**, 163 (2002).
7. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic acids research* **33**, D514 (2005).
8. K. G. Becker, K. C. Barnes, T. J. Bright and S. A. Wang, *Nature genetics* **36**, 431 (2004).
9. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proceedings of the National Academy of Sciences* **106**, 9362 (2009).
10. W. A. B. Jr., K. B. Cohen, L. Fox, G. K. Acquah-Mensah and L. Hunter, *Bioinformatics* **23**, i41 (2007).
11. N. T. Hansen, S. Brunak and R. Altman, *Clinical Pharmacology & Therapeutics* **86**, 183 (2009).
12. S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan *et al.*, *Nature biotechnology* **24**, 537 (2006).
13. O. Vanunu, O. Magger, E. Rupp, T. Shlomi and R. Sharan, *PLoS computational biology* **6**, p. e1000641 (2010).
14. J. E. Hutz, A. T. Kraja, H. L. McLeod and M. A. Province, *Genetic epidemiology* **32**, 779 (2008).
15. J. Chen, B. J. Aronow and A. G. Jegga, *BMC bioinformatics* **10**, p. 73 (2009).
16. L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts and Y. Moreau, *Nucleic acids research* **36**, W377 (2008).
17. G. Gonzalez, J. C. Uribe, L. Tari, C. Brophy and C. Baral, in *Incorporating Interactions, Connectivity, Confidence, and Context Measures. in Pacific Symposium in Biocomputing. 2007. Maui, 2007*.
18. P. R. Costa, M. L. Acencio and N. Lemke, *BMC genomics* **11**, p. S9 (2010).
19. T. G. O. Consortium, *Genome Research* **11**, 1425 (2001).
20. M. Tanenblatt, A. Coden and I. Sominsky, in *International Conference on Language Resources and Evaluation*, 2010.
21. F. Al-Shahrour, R. Díaz-Uriarte and J. Dopazo, *Bioinformatics* **20**, 578 (2004).
22. H. Ge, A. J. Walhout and M. Vidal, *TRENDS in Genetics* **19**, 551 (2003).
23. S. Bird, in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006.
24. U. Kaymak, A. Ben-David and R. Potharst, *Engineering Applications of Artificial Intelligence* **25**, 1082 (2012).
25. Y. Garten, N. P. Tatonetti and R. B. Altman, in *Pac Symp Biocomput*, 2010.
26. X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Molecular Systems Biology* **4** (2008).
27. E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, *BMC bioinformatics* **6**, p. 55 (2005).
28. M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen and P. Bork, *Nucleic acids research* **36**, D684 (2008).
29. J. Tarabeux, O. Kebir, J. Gauthier, F. Hamdan, L. Xiong, A. Piton, D. Spiegelman, E. Henrion, B. Millet, F. Fathalli *et al.*, *Translational psychiatry* **1**, p. e55 (2011).
30. M. Freytsis, X. Wang, I. Peter, C. Guillemette, S. Hazarika, S. X. Duan, D. J. Greenblatt, W. M. Lee *et al.*, *Journal of Pharmacology and Experimental Therapeutics* **345**, 297 (2013).
31. Y. Hirata, C. C. Zai, R. P. Souza, J. A. Lieberman, H. Y. Meltzer and J. L. Kennedy, *Human Psychopharmacology: Clinical and Experimental* **27**, 345 (2012).
32. D. Ghosh, S. Gochhait, D. Banerjee, A. Chatterjee, S. Sinha and K. Nandagopal, *Genetic Testing and Molecular Biomarkers* **16**, 1226 (2012).

# A NOVEL PROFILE BIOMARKER DIAGNOSIS FOR MASS SPECTRAL PROTEOMICS

HENRY HAN<sup>†1,2</sup>

<sup>1</sup>*Department of Computer and Information Science, Fordham University, New York NY 10023 USA* <sup>2</sup>*Quantitative Proteomics Center, Columbia University, New York 10027 USA*  
 Email: xhan9@fordham.edu

Mass spectrometry based proteomics technologies have allowed for a great progress in identifying disease biomarkers for clinical diagnosis and prognosis. However, they face acute challenges from a data reproducibility standpoint, in that no two independent studies have been found to produce the same proteomic patterns. Such reproducibility issues cause the identified biomarker patterns to lose repeatability and prevent real clinical usage. In this work, we propose a profile biomarker approach to overcome this problem from a machine-learning viewpoint by developing a novel derivative component analysis (DCA). As an implicit feature selection algorithm, derivative component analysis enables the separation of true signals from red herrings by capturing subtle data behaviors and removing system noises from a proteomic profile. We further demonstrate its advantages in disease diagnosis by viewing input data as a profile biomarker. The results from our profile biomarker diagnosis suggest an effective solution to overcoming proteomics data's reproducibility problem, present an alternative method for biomarker discovery in proteomics, and provide a good candidate for clinical proteomic diagnosis.

## 1. Introduction

With the recent surge in proteomics, large volumes of mass spectral serum/plasma/urine proteomic data are available to conduct molecular diagnosis in complex diseases. As a promising way to revolutionize medicine, mass spectral proteomics demonstrates a great potential in identifying novel biomarker patterns from a proteome for diagnosis, prognosis, and other diverse clinical needs [1,2]. However, robust clinical diagnosis from mass spectral data remains an acute challenge in translational bioinformatics due to the special characteristics of proteomics data.

First, mass spectral proteomics data are high-dimensional data that can be represented as a matrix  $X \in \mathbb{R}^{n \times p}$  after preprocessing, where each row represents protein expression at a mass-to-charge ( $m/z$ ) ratio of peptides or proteins, usually called a feature from a machine learning perspective, and each column represents protein expression from a sample (observation) (e.g., a control or cancer subject) across all  $m/z$  ratios in an experiment. The number of rows is much greater than the number of columns,  $p \ll n$ , that is, #variables (peptides/proteins) is much greater than #samples. Usually  $n \sim O(10^4)$  and  $p \sim O(10^2)$ . While there are a large amount of  $m/z$  ratios (peptides or proteins), only a few number of variables (e.g., peaks) have meaningful contribution to data variations and disease diagnosis. Moreover, they are not noise-free data because preprocessing and normalization methods themselves cannot remove built-in system noise from mass spectrometry technology itself. In fact, it remains a challenge to separate true signals in a mass spectral profile from red herrings though different endeavors from machine learning.

Second, mass spectral proteomics data usually suffer from data reproducibility problems, which mean that no two independent studies have been found to produce same proteomic patterns [2,3]. As such, corresponding biomarker patterns identified, which consists a small set of meaningful peaks, from these data may lose repeatability due to the poor reproducibility and

difficulty in validating biomarker patterns identified from multiple data sources. In fact, there are almost no reproducible biomarker patterns reported for mass spectral proteomic data in the literature [2]. Although several methods are proposed to mitigate this problem from a quantification perspective [2,3], there is no method to tackle this problem from a machine learning viewpoint as of yet.

The non-reproducibility of proteomic source data and their biomarker patterns, which are usually obtained by peak-selection methods using different machine learning algorithms, is mainly due to mass spectrometry technology's exquisite sensitivity to any subtle change in the proteome caused by biological or technical factors [3]. In other words, tiny changes in the proteome may lead to a set of completely different mass spectral peak patterns. Thus, a desirable diagnosis from identified protein or peptide biomarkers may not be reusable for other "same data" generated using the identical patient and control samples under the same profiling technologies and protocols.

In this work, we propose a *de novo* profile biomarker approach to achieve clinical level diagnosis. Unlike traditional biomarker discoveries that collect a few meaningful peaks, our profile biomarker approach views input data as a "whole biomarker" by proposing a novel derivative component analysis (DCA), which evolves from our previous work [4-6], and combining it with state-of-the-art classifiers. It is noted that a profile biomarker has the same dimension as the input data but with less variance and storage. In our approach, we aim at the reproducibility of diagnosis performance instead of looking for specific peptides or proteins, i.e., we believe a profile biomarker would be more robust than traditional biomarkers, provided it could achieve clinical level diagnoses for different proteomic data. That is, the motivation of this study is to solve the data reproducibility problem in proteomics by developing a novel profile biomarker diagnosis.

Our profile biomarker approach relies on a novel feature selection algorithm: derivative component analysis (DCA) as proposed in this work. Traditional feature selection algorithms (e.g., *t-test*) are usually characterized by the explicit feature number decrease or dimension reduction of the input data. It is noted that a feature refers to a row of protein/peptide expression of all samples at an  $m/z$  ratio. However, as an implicit feature selection algorithm, DCA conducts feature selection implicitly, i.e., there is no feature number decrease after DCA. More importantly, DCA enables the retrieval of the true signals from input proteomic data by removing redundant information and built-in noises, which provide a robust information support for our profile-biomarker diagnosis. Considering similar diagnosis mechanisms for proteomic profiles, we use benchmark serum proteomic data to demonstrate our profile biomarker diagnosis in this study.

The paper is organized as follows. Section 2 discusses essential components in profile biomarker discovery and proposes DCA in addition to addressing the weaknesses of the traditional feature selection methods. Section 3 investigates DCA-based profile biomarker diagnosis by integrating it with state-of-the-art classifiers. We further demonstrate our approach's superiority by comparing it with other state-of-the-arts, besides addressing DCA-induced biomarker discovery. Finally we discuss the pros and cons of our profile biomarker diagnosis and conclude our paper.

## 2. Derivative Component Analysis (DCA)

Before we proceed, we need to answer the question: 'what essential components are needed to make a profile biomarker successful in proteomics?' We believe that essential components for a profile biomarker approach may rely on whether we can separate true signals from red herrings for

each proteomic profile. Traditional feature selection methods usually fail to capture true signals from mass spectral proteomic data set because of their built-in weaknesses. Although various feature selection methods are employed in proteomics to glean informative features for the sake of diagnosis [7], there is no study to address their weaknesses systematically.

We categorize feature selection into input-space and subspace methods. The former seeks a feature subset  $X' \in \mathbb{R}^{m \times p}$ ,  $m \ll n$ , in the same space  $\mathbb{R}^{n \times p}$  as input data  $X$  by conducting a hypothesis test (e.g., *t-test*), or wrapping a classifier to select features recursively; the latter conducts a dimension reduction by transforming data into a subspace  $S$  induced by a linear or nonlinear transformation  $f: X \rightarrow S$ , where  $S = \text{span}(s_1, s_2, \dots, s_k)$ ,  $s_k \in \mathbb{R}^k$ ,  $k \leq p \ll n$ , and seeking meaningful linear combinations of the features. For example, the subspace  $S$  will be spanned by all principal components when the transformation is induced by principal component analysis (PCA) [8]. In fact, almost all PCA, ICA, PLS, and NMF and their extensions such as nonnegative principal component analysis (NPCA), sparse NMF, and other related methods fall into this category [4-6,9]. However, the two types of methods have the following built-in limitations.

*The weakness of the input and subspace methods.* The input-space methods usually assume input data are clean or nearly clean, and lack de-noising schemes. The clean data assumption appears to be inappropriate for proteomic profiles because they usually contain nonlinear noise from technical or biological artifacts (e.g., built-in noise generated from profiling systems). The noise would enter feature selection as outliers and cause those peaks with less biological meaning to be selected, leading to an inaccurate or even poor decision function in classification and affecting the disease diagnosis and generalization.

On the other hand, subspace methods have difficulties capturing subtle data characteristics, because the subspace methods transform data into another subspace in order to seek meaningful feature combinations and the original spatial coordinates are lost in the transformation, which makes it almost impossible to track the mapping relationships between features and the specific data characteristics they interpret or contribute to. It is noted that subtle data characteristics refer to latent data behaviors interpreting transient data changes in a short time interval.

In contrast, global data characteristics refer to the holistic data behaviors interpreting long-time interval data changes, which happen more often than subtle data behaviors. The global data characteristics are easily extracted by general subspace methods like PCA, because there are more features contributing to holistic data behaviors than those contributing to subtle data behaviors. Furthermore, since most subspace methods treat all features uniformly regardless of which types of data behaviors they interpret, global characteristics are more likely to be selected than subtle data characteristics, because the former's features are more frequent than those of the latter in the feature domain.

As such, global data characteristics are usually over-extracted and subtle data characteristics may be totally missed or overshadowed after feature selection. The signals extracted from such feature selection are far from 'true signals' because the global data characteristics are over-expressed. The redundant global data characteristics would lead to a biased decision function for the following classifier (e.g., SVM) that favors the extracted global data characteristics, which may present a hurdle for clinical diagnosis, because the subtle characteristics are essential to achieve high-accuracy diagnosis for proteomics data, especially as different subtype tumor samples usually share similar or the same global data characteristics but different subtle data characteristics [5,6].

It is clear that the built-in weaknesses of the traditional feature selection methods prevent true

signal extraction and the possibility of profile biomarker diagnosis, because they lack de-noising and subtle data characteristics retrieval schemes. We sketch the key reasons for these weaknesses as follows before we present our derivative component analysis.

*The reasons for traditional feature selection's weaknesses.* The following are the major reasons why traditional feature selection methods are unable to extract subtle characteristics and remove systems noise effectively. 1) These methods are single resolution data analysis methods that view each feature as an indivisible information unit, which makes system noise removal almost impossible; 2) They treat all features uniformly regardless of their frequencies in the feature space, which makes subtle data characteristics extraction difficult due to lower frequencies in the feature domain. Mathematically, retrieving subtle data characteristics, which are represented by transient data behaviors, means to seek the derivative of the original data. However, this is theoretically quite difficult to complete in a single resolution mode.

*Derivative component analysis (DCA).* We propose a novel feature selection algorithm: derivative component analysis (DCA) to separate true signals from red herrings, that is, conduct de-noising for system noise and retrieve subtle data characteristics in a multi-resolution data analysis mode. As a multi-resolution feature selection algorithm, the proposed DCA no longer views a feature as an indivisible information element. Instead, all features are hierarchically decomposed into different components to discover data derivatives so as to capture the subtle data characteristics and conduct de-noising. The proposed derivative component analysis (DCA) mainly consists of the following three steps.

First, a discrete wavelet transform (DWT) [10] is applied to all features to decompose it hierarchically as a set of detail coefficient matrices  $cD_1, cD_2 \dots cD_J$  and an approximation matrix  $cA_J$  under a transform level  $J$ . It is worthwhile to point out that we view each  $m/z$  ratio as a corresponding time point in our context for the convenience of the DWT [10]. Since the DWT is calculated on a set of dyadic grid points hierarchically, the dimensionalities of the approximation and detail coefficient matrices shrink dyadically level by level.

It is noted that the approximation matrix and coarse level detail coefficient matrices (e.g.  $cD_J$ ) capture global data characteristics, because they contain contributions from those features contributing to data behaviors in 'long-time windows', and outlining the global tendency of the data. Similarly, the fine level detail coefficient matrices (e.g.,  $cD_1, cD_2$ ) capture subtle data characteristics, because they contain contributions from those features that disclose quick changes in 'short-time windows', and describe data derivatives locally. In fact, these fine level detail matrices are the components for reflecting the data derivatives in different short-time windows. As such, they can be called 'derivative components' for the functionality in describing data behaviors.

Furthermore, most system noises are transformed in these derivative components due to its heterogeneity with respect to the features contributing to the global tendency of data. Clearly, the DWT in the first step separates the global characteristics, subtle data characteristics, and noises in different resolutions.

Second, retrieve the most important subtle data characteristics and conduct de-noising by reconstructing these fine level detail coefficient matrices before or at a presetting cutoff level  $\tau$  (e.g.,  $\tau=3$ ). Such a construction is summarized in two steps: 1) Conduct principal component analysis (PCA) for the detail matrices  $cD_1, cD_2 \dots cD_\tau$ . 2) Reconstruct each detail coefficient matrix by using its first  $m$  principal components, in each principal component (PC) matrix. Usually,  $m = 1$ , i.e., we employ the first PC to reconstruct each detail coefficient matrix, which means we only retrieve the most important subtle data characteristics in the detail coefficient matrix

reconstruction. In fact, the first PC based reconstruction also achieves de-noising by suppressing the noises' contribution in the detail coefficient matrix reconstruction because the noises are usually least likely to appear in the 1<sup>st</sup> PC.

On the other hand, those coarse level detail coefficient matrices after the cutoff  $\tau$ :  $cD_{\tau+1}, cD_{\tau+2} \dots cD_J$  and approximation coefficient matrix  $cA_J$  are kept intact to retrieve global data characteristics. In fact, the parameter  $m$  can be also determined by using a variability explanation ratio  $\rho_m$  defined as follows, such that it is greater than a threshold  $\rho$  (e.g.,  $\rho = 60\%$ ), which is the variability explanation ratio interpreted by the first principal components of the detail coefficient matrices before or at the cutoff.

*Variability explanation ratio.* Given a data set with  $n$  variables and  $p$  observations, usually,  $p < n$ , the variability explanation ratio is the ratio  $\rho_m = \sum_{i=1}^m \sigma_i / \sum_{i=1}^p \sigma_i$  between the variance explained by the first  $m$  PCs and the total variances, where  $\sigma_j$  is the variance explained by the  $j^{th}$  PC, which is the  $j^{th}$  eigenvalue of the covariance matrix of the input proteomic data.

Such a selective reconstruction process extracts the most important subtle data characteristics and achieves de-noising by suppressing the noises' contribution to the fine detail coefficient matrix reconstruction. This is because only the 1<sup>st</sup> PC or few top PCs are employed to reconstruct each targeted fine level coefficient matrix  $cD_j$  and the other less important and noise-contained principal components are dropped in reconstruction.

Third, conduct the corresponding inverse DWT by using the current detail and approximation coefficient matrices to obtain meta-data  $X_*$ , which is a de-noised data set with <sup>subtle</sup> data characteristics extraction, because of the highlight of the most significant subtle data behaviors in the “derivative components” based reconstructions. The meta-data are just ‘true signals’ separated from red herrings that share the same dimensionality with the original data but with less memory storage because less important PCs are dropped in our reconstruction.

It is noted that, unlike traditional feature selection methods, DCA is an implicit feature selection method, where useful characteristics are selected implicitly without an obvious variable removal or dimension reduction. Algorithm 1 gives the details about DCA as follows, where we use  $X^T$  instead of  $X$  for the convenience of description, i.e., each row is a sample and each column is a feature.

**Algorithm 1 Derivative Component Analysis (DCA)**

1. **Input:**  $X^T = [x_1, x_2, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$ , DWT level  $J$ ; cutoff  $\tau$ ; wavelet  $\psi$ ; threshold  $\rho$ ;
2. **Output:** Meta-data  $X_*^T$
3. **Step 1.** Column-wise discrete wavelet transforms (DWT)
4. Conduct J-level DWT with wavelet  $\psi$  for each column of  $X^T$  to obtain
5.  $[cD_1, cD_2, \dots, cD_J; cA_J]$ ,  $cD_j \in \mathbb{R}^{p_j \times n}$ ,  $cA_J \in \mathbb{R}^{p_J \times n}$ , and  $p_j = \lceil p / 2^j \rceil$ ,  $j = 1, 2, \dots, J$ .
6. **Step 2.** Subtle data characteristics extraction and de-noising
7. for  $j = 1$  to  $J$
8.   if  $j \leq \tau$
9.     a) Do principal component analysis for each detail matrix  $cD_j$  to obtain its PC and score matrix
10.        $U = [u_1, u_2, \dots, u_{p_j}]$ ,  $u_i \in \mathbb{R}^n$  and  $S = [s_1, s_2, \dots, s_{p_j}]$ ,  $s_i \in \mathbb{R}^{p_j}$ ,  $i = 1, 2, \dots, p_j$ .
11.     b) Reconstruct matrix  $cD_j$  by employing first  $m$  principal components  $u_1, u_2, \dots, u_m$ , s.t.  $\rho_m \geq \rho$
12.        $cD_j \leftarrow cD_j \times (I \times I^T) / p_j + \sum_{i=1}^m u_i \times s_i^T$ ,  $I = [1, 1, \dots, 1]^T \in \mathbb{R}^{p_j}$



13. *end if*
14. *end for*
15. **Step 3.** Approximate the original data by the inverse discrete wavelet transform
16.  $X_s^T \leftarrow \text{inverseDWT}([cD_1, cD_2 \dots cD_J; cA_J])$  with the wavelet  $\psi$

Although an optimal DWT level can be obtained theoretically according to the maximum entropy principle [11], it is reasonable to adaptively select the DWT level  $J$  according to the 'nature' of input data, where large #samples corresponds to a relatively large  $J$  value, for the convenience of computation. As such, we select the DWT level as  $4 \leq J \leq \lfloor \log_2 p \rfloor$  considering the magnitude level of the samples number  $p$  in proteomics data to avoid too large or too small transform levels. Correspondingly, we empirically set the cutoff as  $1 < \tau \leq J/2$  to separate the fine and coarse level detail coefficient matrices for good performance.

Furthermore, we require the wavelet  $\psi$  in the DWT orthogonal and have compact supports such as *Daubechies* wavelets (e.g., 'db8'), for the sake of subtle data behavior capturing. Interestingly, we have found that the first PC of each fine level detail coefficient matrix usually has a quite high variability explanation ratio (e.g., >60%) for each fine level detail coefficient matrix  $cD_j$  ( $1 \leq j \leq \tau$ ). Thus, we relax the variability explanation ratio threshold by only using the first PC to reconstruct each  $cD_j$  matrix in order to catch subtle data characteristics along the maximum variance direction. In fact, we have found that using more PCs in the fine level detail coefficient matrix reconstruction does not demonstrate advantages in subtle data characteristics extraction and de-noising than using the first PC.

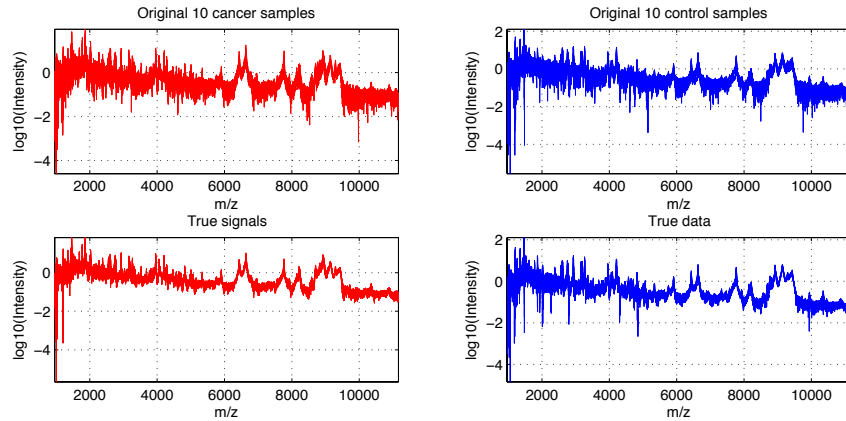


Fig 1. The true signals of 10 cancer and control samples across 16331  $m/z$  of the *Colorectal* data by DCA

Figure 1 shows the true signals (meta-data) of the 10 cancer and control samples, which are randomly selected from *Colorectal* data [12] with total 48 controls and 64 cancer samples across 16331  $m/z$  ratios, extracted by our DCA under the cutoff  $\tau=2$ , transform-level  $J=7$ , and wavelet 'db8'. Interestingly, the each type of samples in the extracted true signals appear to be smoother and more proximal to each other besides demonstrating less variations, because of the major subtle data characteristics extraction and system noise removal.

Such a case is demonstrated more clearly by Figure 2, where the 10 cancer and control samples and their true signals are highlighted between 1400 Da and 1500 Da. It is quite clear to observe that the same type samples are closer to each other spatially, and some small spikes are removed as the built-in noises in true signals. Obviously, from a classification viewpoint, these

true signals will contribute to high accuracy diagnoses than the original proteomic data, because the built-in noises and redundant global data characteristics would have a much lower chance to get involved in classification due to derivative component analysis. Instead, subtle data characteristics would have a greater chance of participating in the decision rule inference.

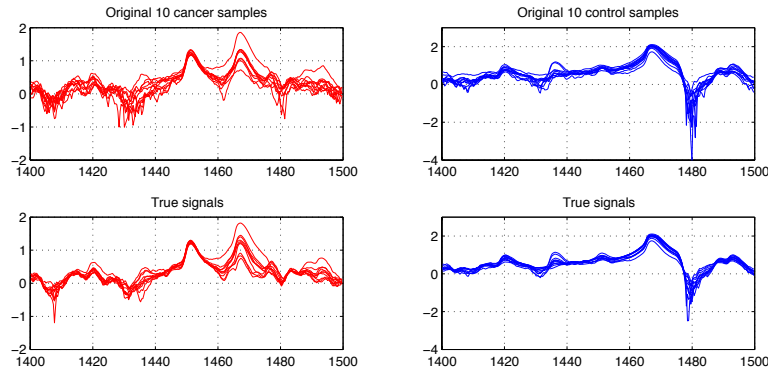


Fig 2. The true signals of 10 cancer and control samples of the *Colorectal* data between 1400-1500 Da.

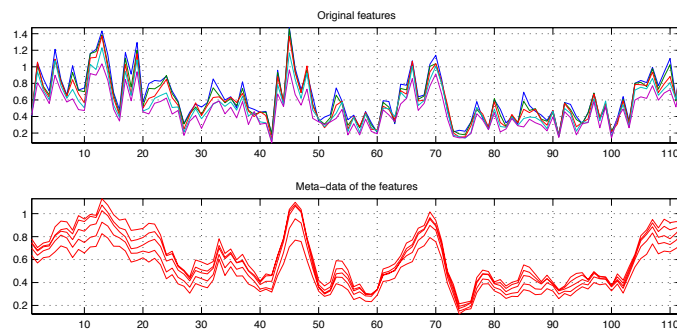


Fig 3. Random five features in *Colorectal* data and its meta-data across 112 samples (64 cancers + 48 controls).

Similarly, Figure 3 shows the meta-data of randomly picked five features from *Colorectal* data under the same parametric setting for DCA. Interestingly, the meta-data (meta-features) are *smoother* and have *values in a smaller range* than the original features for its subtle data characteristics extraction and de-noising. The meta-features are actually more distinguishable than their original features, which reflect the true expression level of the peptides/proteins at the  $m/z$  ratios better. In other words, DCA provides a ‘zoom’ mechanism to capture the original data’s subtle behaviors that are usually latent in general machine-learning methods.

### Profile Biomarker Diagnosis with DCA

Since DCA can separate true signals from red herrings by extracting subtle data characteristics and removing built-in noises, it is natural to combine DCA with start-of-the-art classifiers to conduct profile biomarker diagnosis, where input proteomics data are viewed as a profile biomarker. We chose support vector machines (SVM) for its efficiency and advantages in handling large-scale data, popularity in proteomics diagnosis and biomarker discovery [13]. As such, we propose novel derivative component analysis-based support vector machines (DCA-

SVM) in order to attain a profile biomarker disease diagnosis, which is actually equivalent to a binary or multi-class classification problem.

Given a binary type training samples  $X=[x_1, x_2, \dots, x_p]^T$  and their labels  $\{x_i, c_i\}_{i=1}^p$ ,  $c_i \in \{-1, 1\}$ , its corresponding meta-data  $Y=[y_1, y_2, \dots, y_p]^T$  are computed by using DCA. Then, a maximum-margin hyperplane  $O_h: w^T y + b = 0$  in  $\mathbb{R}^n$  is constructed to separate the '+1' ('cancer') and '-1' ('control') types of the samples in the meta-data  $Y$ , where  $w$  and  $b$  are the normal and offset vector of the hyperplane respectively. The hyperplane construction is equivalent to solving the following quadratic programming problem (standard SVM, i.e., C-SVM):

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^p \xi_i \\ \text{s.t.} \quad & c_i(w^T y_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, p \\ & \xi_i \geq 0 \end{aligned} \quad (1)$$

The C-SVM can be solved by seeking the solutions to the variables  $\alpha_i$  of a corresponding Lagrangian dual problem to get a decision function  $f(x') = \text{sign}(\sum_{i=1}^n \alpha_i c_i k(y_i \cdot y') + b)$  to determine the

class type of a testing sample  $x'$ , where  $y'$  and  $y_i$  are corresponding meta-samples computed from DCA for samples  $x'$  and  $x_i$ . The kernel function  $k(y_i, y)$  maps  $y_i$  and  $y'$  into a same-dimensional or high-dimensional feature space. In this work, we employ the 'linear' kernel for its simplicity and efficiency. Our multiclass DCA-SVM algorithm employs the 'one-against-one' to conduct multiclass phenotype diagnosis for its proved advantage over the 'one-against-all' and 'directed acyclic SVM' methods [14].

It is worthwhile to point out that our DCA-SVM has a different feature space due to true signal extraction from DCA. The standard SVM's feature-space usually contains noises from input proteomic data, and misses subtle data characteristics. Alternatively, the DCA-SVM's feature space contains 'de-noised' true signals with subtle data characteristics, which avoids the global data characteristics favored decision rule because subtle data characteristics are also invited in SVM hyperplane construction besides the global data characteristics. As such, the DCA-SVM can efficiently detect those samples with similar global characteristics but different subtle characteristics in disease diagnosis than the standard SVM.

### 3. Results

We demonstrate our profile biomarker diagnosis' superiority by using five benchmark serum proteomic data sets, which include *Cirrhosis*, *Colorectal*, *HCC*, *Ovarian-qaqc* and *ToxPath* data [12,15-17,19]. The benchmark data used in our experiments are heterogeneous data generated from different experiments via different profiling technologies such as MALDI-TOF and SELDI-TOF, and preprocessed by different methods. Table 1 describes the details of the five data sets.

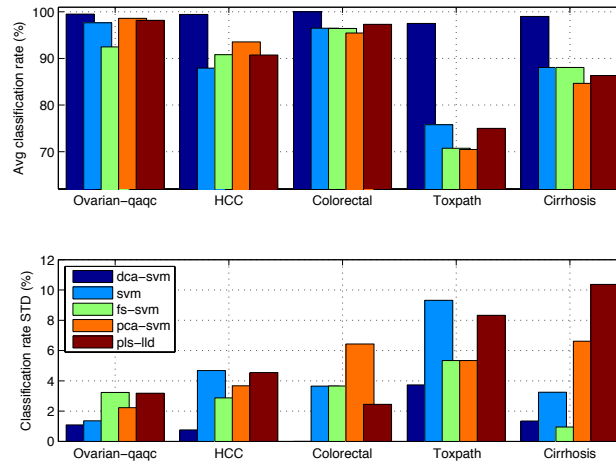
We compare the proposed DCA-SVM based profile-biomarker diagnosis with the following state-of-the-arts in this work. They include a partial least square (PLS) based linear logistic discriminant analysis (PLS-LLD) [18,20], standard SVM [13], a SVM combining with principal component analysis: PCA-SVM [5], and a SVM with input-space feature selection: *fs*-SVM, which employs *t-test* and *Anonal* (*one-way ANOVA*) to conduct feature selection for binary and multi-class data respectively. For each data, *fs*-SVM collects a meaningful feature set including all

features with  $p\text{-values} < 0.05$  using  $t\text{-test}$  or  $Anova1$  for phenotype diagnosis.

Table 1. Benchmark proteomic data

Data	#Feature	#Sample	Platform
<i>Cirrhosis</i>	23846	72 controls + 78 HCCs + 51 cirrhosis	MALDI-TOF
<i>Colorectal</i>	16331	48 controls + 64 cancers	MALDI-TOF
<i>HCC</i>	6107	181 controls + 176 cancers	SELDI-QqTOF
<i>Ovarian-qaqc</i>	15000	95 controls + 121 cancers	SELDI-TOF
<i>ToxPath</i>	7105	28 normals + 43 potential normals + 34 cardiotoxicities + 10 potential cardiotoxicities	SELDI-QqTOF

We employ the 'linear' kernel  $k(x, y) = (x \cdot y)$  in all SVM-related classifiers for its efficiency in omics data classification, rather than nonlinear kernels (e.g., Gaussian kernels), which usually lead to overfitting in diagnosis [4-6]. To avoid potential biases from presetting training/test data partition on diagnosis, we employ the  $k\text{-fold}$  ( $k=5$ ) cross-validation to evaluate the five classifiers' performances for all data sets. In addition to choosing the first ten PLS components in the PLS-LLD classifier, we uniformly set the DWT level  $J = 7$  under 'db8', cutoff  $\tau = 2$ ; and apply the first PC-based detail coefficient matrix reconstruction in DCA to retrieve true signals for all proteomic data sets.



**Fig 4** Comparing profile biomarker diagnosis' diagnostic accuracies and its standard deviations with those of others.

Before demonstrating our profile biomarker approach's advantages, we introduce several key diagnosis performance measures, namely, diagnostic accuracy, sensitivity, specificity and positive predication ratios, as follows. The diagnostic accuracy is the ratio of the correctly classified test samples over total test samples. The sensitivity, specificity, and positive predication ratio are defined as the ratios:  $\frac{TP}{TP+FN}$ ,  $\frac{TN}{TN+FP}$ , and  $\frac{TP}{FP+TP}$  respectively, where  $TP(TN)$  is the number of positive (negative) targets (a positive (negative) target is a proteomic sample with '+1' ('-1') label) correctly diagnosed and  $FP(FN)$  is the number of negative (positive) targets incorrectly

diagnosed by the classifier.

Figure 4 demonstrates rivaling clinical level performance from our profile biomarker diagnosis (DCA-SVM) by comparison with the other classifiers in average diagnosis accuracies and its standard deviations. It seems that our profile biomarker diagnosis achieves performance nearly clinical level and demonstrate strongly leading advantages over its peers in a stable manner. Alternatively, those comparison classifiers seem to show quite large level oscillations that may indicate they lack stability and good generalization capacities across different data sets, which exclude themselves as candidates for clinical proteomics diagnosis.

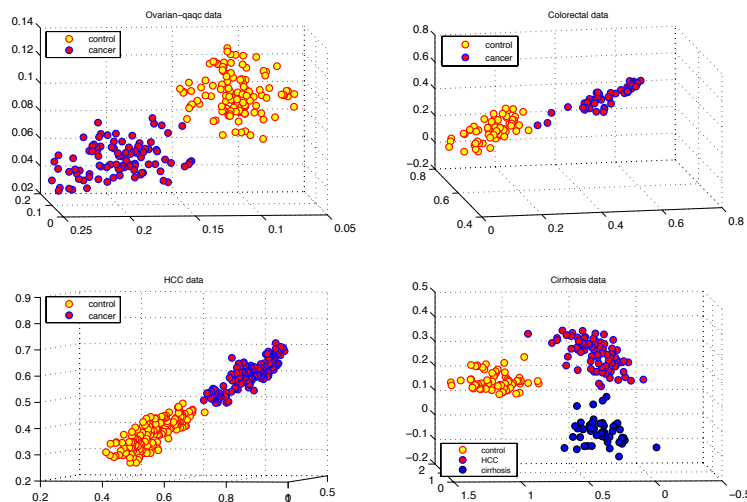
For example, our profile biomarker diagnosis achieves 99.52% (sensitivity 100%, specificity 99.17%), 100% (sensitivity 100%, specificity 100%), and 99.44% (sensitivity 98.00%, specificity 100%) diagnostic accuracies on the *Ovarian-qaqc*, *Colorectal* and *HCC* data respectively. It further reaches 97.50%, 99.01% diagnostic accuracies for *Toxpath* and *Cirrhosis* data respectively. However, the standard SVM classifier can only achieve 75.80% and 88.06% diagnosis for the same data sets respectively. Although some input-space or subspace methods may sometimes boost diagnosis for binary-type data set, we have found that they are unable to increase the SVM classifier's diagnosis and generation abilities significantly, especially for multiclass proteomic data. In fact, in contrast to the proposed profile biomarker diagnosis, all the comparison classifiers show high-level oscillations in diagnoses across different data sets. It is noteworthy that the high-level oscillations in diagnosis is further highlighted by corresponding large standard deviation values in diagnosis from those classifiers in Figure 4, where our DCA-SVM based profile biomarker diagnosis demonstrates its good stability and generalization for its smallest standard deviation values across all the data sets.

*Compare profile biomarker diagnosis with prior methods.* It is worthwhile to point out that our DCA-SVM based profile biomarker also demonstrates its superiority to its peers in terms of diagnostic accuracy, sensitivity, specificity and positive predication ratios. We further compare our profile biomarker diagnosis approach with the previous biomarker discovery diagnoses in the literature and have found that our method demonstrates good clinical level sensitivities in phenotype discriminations for different benchmark proteomic data. For example, Alexandrov et al 's work only achieved 97.5% diagnosis accuracy with sensitivity 98.4% and specificity 95.8% for *Colorecta* data by using a complicated method [12]. However, our profile biomarker diagnosis achieves 100% diagnosis accuracy with sensitivity 100% and specificity 100%. For *Ovarian-qaqc* data, our approach achieves a 99.53% clinical-level diagnosis accuracy with sensitivity 98.95% and specificity 100%, which is better than the original diagnosis level obtained in [17] and all the other peers. For *Cirrhosis* data, Resson *et al* partitioned this three-class data into two binary data sets and proposed a novel hybrid ant colony optimization based support vector machines (ACO-SVM), where ACO was used for biomarker discovery, to achieve 94% and 100% specificity to distinguish hepatocellular carcinoma (HCC) from cirrhosis [16]. There was no result available to distinguish normal, HCC, and cirrhosis in a multiclass diagnostic way. However, our proposed approach has achieved 99.01% diagnosis accuracy for this multi-class data set.

Can DCA be used to conduct biomarker discovery by collecting meaningful peaks if we relax the reproducibility concern? The answer is 'yes' because derivative component analysis can identify meaningful protein or peptide peaks from true signals. We simply apply *t-test* and *Anova* to identify the top-ranked features with the smallest *p-values*, i.e. we pick the three top-scored peaks as biomarkers for its statistical significance. Figure 5 illustrates the separation of four benchmark data sets with three top-ranked biomarkers (peaks). It is interesting to see that these

high-dimensional proteomic profiles can be separated almost completely with these biomarkers identified from true signals.

We can also obtain some meaningful biological depth by checking these biomarkers. For example, the SW plot in Figure 5 shows the separation of 176 controls and 181 cancers in the *HCC* data, by the top-ranked biomarkers (peaks) at 2534.2, 2584.3, and 6486.2  $m/z$  ratios, where each dot represents a sample (a patient with HCC or a healthy subject). It is also interesting to see that two biomarkers are from downstream  $m/z$  ratios, which were believed to be more sensitive to detect phenotype information than those from upstream  $m/z$  ratios [16,19]. Moreover, The separation can provide meaningful biological insight for pathological disease states. For example, we select three top-ranked biomarkers at 1668.99, 5907.73, 5907.13  $m/z$  ratios for the *Cirrhosis* dataset, which is a three-class high-resolution MALDI-TOF proteomic profile with 23846 features. The phenotype separations provided by the three biomarkers give very meaningful biological insights, i.e., the SE plot in Figure 5 shows the three clearly independent clusters, where Cirrhosis cluster with 51 samples (blue) have closer spatial distances to the HCC cluster 78 samples (red) than the normal cluster with 72 samples (yellow). Such spatial distances demonstrated by our biomarkers are actually consistent to their pathological distances: Cirrhosis is the middle stage to hepatocellular carcinoma (HCC) for a healthy subject.



**Fig 5** Separating disease phenotypes of four data sets by only using their three biomarkers with the smallest p-values.

#### 4. Conclusions and Discussion

In this study, we propose a profile biomarker diagnosis approach to overcome the data reproducibility issue in proteomics data and demonstrate its clinical level performances across different data. The profile biomarker diagnosis is based on the novel implicit feature selection algorithm: derivative component analysis and derivative component analysis based support vector machines proposed in this study. As an implicit feature selection algorithm, DCA is able to separate true signals from red herrings by extracting subtle data characteristics and removing system noise via calculating a same dimensional meta-data for input proteomic data. It is noted that the complexity of DCA is higher than that of PCA, because DCA calls the classic PCA in several fine level detail coefficient matrix reconstruction, in addition to the DWT and inverse

DWT. However, DCA demonstrates a promising way to overcome the data reproducibility issue in proteomics because the high-accuracy diagnosis results seem to be reproducible themselves for different data sets under our approach. In other words, our profile biomarker diagnosis presents itself as an ideal candidate to achieve clinical diagnosis in clinical proteomics. Furthermore, our work suggests a key issue in proteomic disease diagnosis, that is, subtle data characteristics gleaned and de-noising can be more important in proteomics data feature selection and following phenotype discrimination than dimension reduction. Moreover, the proposed derivative component analysis provides an alternative feature selection by implicitly extracting useful data characteristics while maintaining the data's original dimensionality.

Although we are quite optimistic to see that our profile biomarker diagnosis will be a potential candidate to achieve a clinical disease diagnosis in proteomics by conquering the reproducibility problem, rigorous proteomics clinical tests are needed urgently to explore such a potential and validate its clinical effectiveness. In our ongoing work, we are working with pathologists to investigate extending the profile biomarker diagnosis approach to TCGA and RNA-Seq data besides protein expression array analysis.

## 5. References

1. T. Rath et al, Serum Proteome Profiling Identifies Novel and Powerful Markers of Cystic Fibrosis Liver Disease, *PLoS ONE*, (2013)
2. J. Ioannidis et al, Improving Validation Practices in "Omics" Research, *Science* **334**, 1230, (2011)
3. R. Hüttenhain et al, Reproducible Quantification of Cancer-Associated Proteins in Body Fluids Using Targeted Proteomics, *Sci Transl Med* **4**, 142ra94, (2012)
4. X. Han, Nonnegative Principal component Analysis for Cancer Molecular Pattern Discovery, *IEEE/ACM Transaction of Computational Biology and Bioinformatics* **7** (3), p537-549, (2010)
5. X. Han, Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery, *BMC Bioinformatics*, **11**(Suppl 1): S1, (2010)
6. H. Han, and X. Li, Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery, *BMC Bioinformatics*, 12(S1):S7, (2011)
7. M. Hilario and A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, *briefings in bioinformatics*, **9**:2 101-119, (2008)
8. I. Jolliffe, *Principal component analysis*, Springer, New York, (2002)
9. J. Brunet, et al, Molecular pattern discovery using matrix factorization, *PNAS* **101**(12),4164–69, (2004)
10. S. Mallat, *A wavelet tour of signal processing*, Acad. Press, CA, USA, (1999)
11. T. Kapur and A. Keshavan, *Entropy optimization principles with applications*, Academic Press, (1992)
12. T. Alexandrov et al, Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation, *Bioinformatics*, Vol. 25(5):643-649, (2009)
13. V. Vapnik, *Statistical Learning Theory*. John Wiley, New York, (1998)
14. C. Hus and C. Lin, A Comparison of Methods for Multi-class Support Vector Machines, *IEEE Transactions on Neural Networks*, **13** (2):415-425, (2002)
15. H. Resson et al, Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics* **21**(21), 4039-4045, (2005)
16. H. Resson et al, Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* **23**(5), 619-626, (2007)
17. T. Conrads et al, High-resolution serum proteomic features for ovarian detection, *Endocrine-Related Cancer*, **11**, 163-178 (2004)
18. D. Nguyen, and D. Rocke, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18**:39–50, (2002)
19. E. Petricoin et al, Toxicoproteomics: serum proteomic pattern diagnostics for early detection of drug induced, *Toxicologic Pathology*, **32** (Suppl. 1):1–9, (2004)
20. D. Sampson et al A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches, *PLoS One*, (2011)

## TOWARDS PATHWAY CURATION THROUGH LITERATURE MINING – A CASE STUDY USING PHARMGKB

RAVIKUMAR K.E., KAVISHWAR B. WAGHOLIKAR, HONGFANG LIU

*Department of Health Sciences Research, College of Medicine, Mayo clinic, Rochester, MN, 55905*

*Email: {KomandurElayavilli.Ravikumar, Waghlikar.Kavishwar, Liu.Hongfang}@mayo.edu*

The creation of biological pathway knowledge bases is largely driven by manual effort to curate based on evidences from the scientific literature. It is highly challenging for the curators to keep up with the literature. Text mining applications have been developed in the last decade to assist human curators to speed up the curation pace where majority of them aim to identify the most relevant papers for curation with little attempt to directly extract the pathway information from text. In this paper, we describe a rule-based literature mining system to extract pathway information from text. We evaluated the system using curated pharmacokinetic (PK) and pharmacodynamic (PD) pathways in PharmGKB. The system achieved an F-measure of 63.11% and 34.99% for entity extraction and event extraction respectively against all PubMed abstracts cited in PharmGKB. It may be possible to improve the system performance by incorporating using statistical machine learning approaches. This study also helped us gain insights into the barriers towards automated event extraction from text for pathway curation.

### 1 Introduction

Genome-wide high throughput studies have led to an increased emphasis on understanding the biological interactions at the systems level rather than the individual molecular interactions. Biological pathway knowledge bases provide systems level interaction information, and are constructed by manual curation of the scientific literature. Due to extensive manual effort required, there is a significant delay in capturing the information in knowledge bases after the publication of scientific literature. Baumgartner et al 2007 (1) suggests that manual curation of biological databases is beyond human life span without significant assistance from text mining. Increase in the volumes of biomedical literature has witnessed simultaneous improvements in the ability to apply natural language processing (NLP) methods to full text articles and entire PubMed collection (2-4).

Despite a decade of research in biomedical text mining the effort to semi-automate the curation workflow of various biological databases and pathway databases in particular is still evasive (5). Some of the earlier systems targeted the acquisition of protein networks (binary relations) from literature are simply based on co-occurrence such as iHOP (6), Chillibot (7), or grammar-based rules such as Pathway Studio (8) and GeneWays (9). While extraction of such networks is useful, the networks cannot be easily mapped to pathways, which model information flow in biological cascades.

While most of the systems mentioned above extract binary relations there has been significant improvement in the state of the art by progressing the extraction from simple binary interactions to complex events, which form building blocks of a pathway. In the recent past the efforts to achieve automated biomedical text mining have been catalyzed by a series of BioCreative (10, 11) and BioNLP shared tasks (5, 12, 13). These competitions saw the emergence of systems (2, 3, 14, 15) that extract complex events where simple events are part of other events using both machine



learning and rule-based approaches. PathText (16) proposed an integrated approach to ease the manual effort involved in pathway curation task but still requires lot of manual effort. The most recent BioNLP shared task 2013 (5) organized a task dedicated to pathway curation. Only two systems, TEES (3) and NacTeM (17) participated in this task, which reported an F-measure of 52.84% and 51.10% respectively on the task. Schmidt et al 2012 (18) also explored text mining assisted pathway curation in a limited context of a specific pathway involving kinases.

While the recent studies indicate a step forward in the direction of pathway curation, they do not completely address all the issues necessary for pathway curation. We are not aware of any study that evaluates a text mining system for extracting biological pathways that uses a manually curated pathway database as the gold standard.

In this study we describe an event extraction that uses pattern templates (covering nearly 450 verbs describing biological events) to extract arguments and assign semantic roles for events described within a single sentence. In addition the system uses linguistic rules to connect information across sentences, which is a major distinguishing feature of the system from rest of the systems described above. Finally we investigate an important problem of great significance, the role our text mining system can play in assisting pathway curation through extraction of events and identify the challenges to our text mining system in extracting the event annotations in PharmGKB (19) pathway database.

## 2 Methods

Figure 1 shows the overall system architecture and the individual components of our text mining system.

### 2.1 Pre-processing and Named entity recognition

The pipeline starts with tokenization and sentence detection for a given document. The sentences are then assigned part of speech using Brill Tagger (20) trained on GENIA corpus (21). POS tagging is augmented by post-processing error correction rules. This is followed by shallow parsing using fnTBL chunker (22) trained on GENIA corpus (21). The shallow parsing is supplemented with detection of additional syntactic constructions related to noun phrases, which include co-ordination, appositives and verb groups.

The next component is named entity recognition (NER) component consisting of manually developed rules as outlined by Narayanaswamy et al 2003 (23) and dictionaries of words and morphological features like prefixes, suffixes and infixes for biomedical entities. The NER component classifies entities into 8 major categories namely Protein/Gene, protein sites, chemicals, drugs, organism, bodypart (include organ, tissue, cells and sub-cellular location), disease, quantitative parameter (e.g. conductance, voltage, binding constant, dissociation constant, IC50) and values (e.g. 20 nM, 30 pS, 10 ms). Based on the NER results we corrected the errors in POS tagging and shallow parsing module by having a feedback loop in order to improve the performance of event extraction.

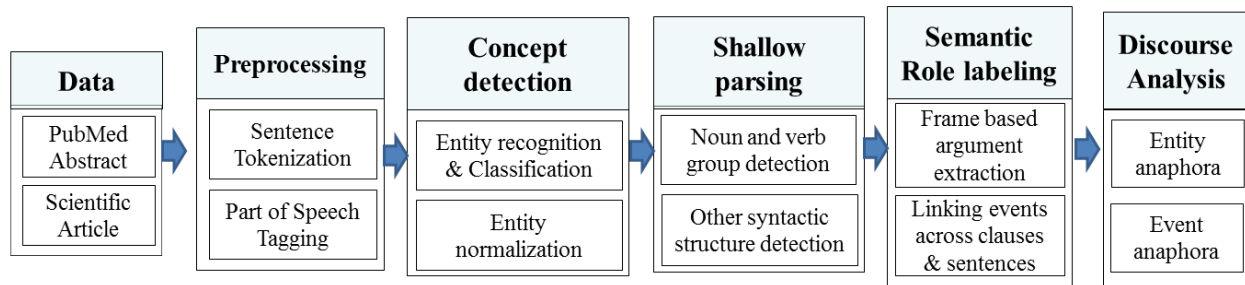


Figure1 – System Architecture

## 2.2 Event extraction

The event extraction module consists of two major sub components 1) detection of events within a clause or sentence based on pattern templates and 2) connecting events across sentences through discourse analysis.

### 2.2.1 Argument extraction based on verb frames

The system consists of rules for different classes of verbs or its nominal forms that extract and assign thematic roles to its arguments based on verb category and the semantic type of the arguments. The patterns for each verb were developed using a corpus of 300 abstracts related to electrophysiology sub-domain describing events about ion channel physiology. Currently there are 9 major classes and 50 sub-classes of verbs. The patterns consider the verbal forms such as activate, inhibit, transport and nominal forms such as activation and phosphorylation. These verb/nominal forms are marked as potential triggers and there are 450 such triggers identified across all categories. Table 1 lists the major category of event classes and the corresponding verbs for defining frames for argument extraction. Some example patterns are included below with example sentences can be found in Figure 2.

**Pattern 1:** <Agent> (PRP NP)\* REGULATE\_VERB <Theme> (PRP NP)\*

This template matches a clause with a verb and extends the clause on either side of the verb as long as each of the base noun phrases that it crosses is headed only by a preposition (shown in Figure 2A). Regulatory verbs (both positive, negative and neutral) such as “increased”, “stimulated”, “blocked”, and “prevented”, “regulated” have the above argument structure and are matched by this pattern.

**Pattern2:** < Nominal form NP> of <THEME> by <AGENT>

This pattern matches the sentence and extracts arguments (shown in Figure 2B). A similar pattern handles passive forms of the verb as shown in Figure 2C.

**Pattern 3:** <AGENT>, [Nominal form NP] of <THEME>

Pattern 3 handles nominal forms within appositive expressions like in “Gd3+, an *inhibitor of the flow -induced Ca2+ increase*, prevented the hyperpolarization” and extracts the arguments (“Gd3+” as agent and “flow -induced Ca2+ increase” as theme) for the trigger “inhibitor”.

### 2.2.2 Connecting events across clausal boundaries

We explored a few linguistic motivated approaches to connect or transfer arguments across clausal boundaries. Our strategy involve three steps: 1) fill empty semantic slots by transferring the arguments across events, 2) merge relevant frames and write parser to connect discourses, 3) resolve anaphoric expressions to find the right antecedent for both entities and events. Figure 2 shows the examples for frame based argument extraction output using BRAT annotation tool.

Table 1. Verb categories

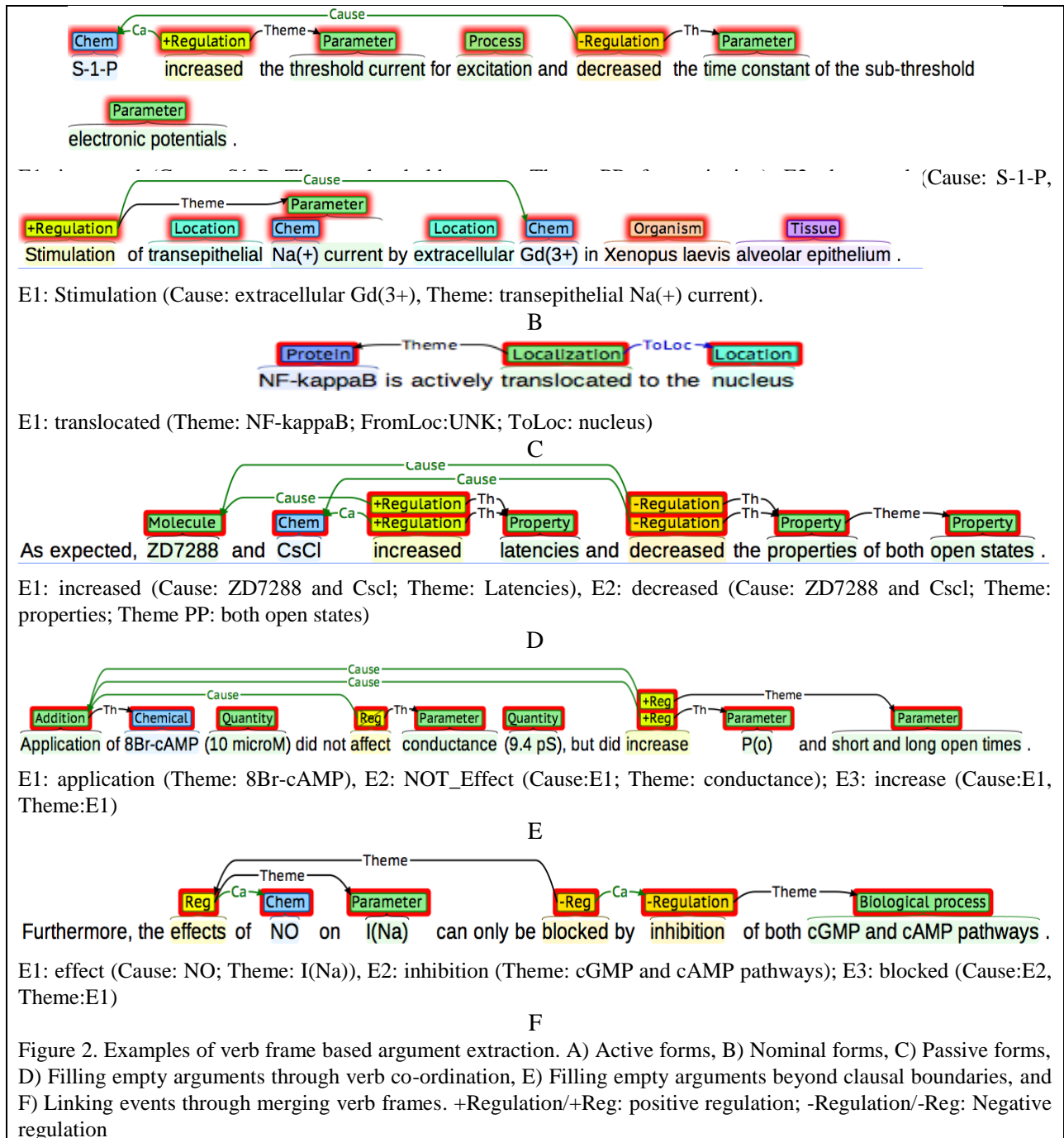
Category	Example verbs
Conversion	Phosphorylation, methylation, de-phosphorylation etc. and other PTMs
Localization	Transport, trans-located, movement
Gene expression	Expression, transcription, translation
Degradation	Degradation
Binding	Bind, binding, complex formation
Dissociation	Dissociate, bond break
<b>Regulation</b>	
Positive	Activation, induce, trigger
Negative	Inhibition, inactivation
Neutral	Modulate, regulate

Filling empty slots by transferring arguments across events - Quite often, syntactic arguments of verbs or its nominalized form, either the subject/object will be empty. Such situations demand mechanisms to fill the empty arguments by linking the current frame with another. Consider the example sentence shown in Figure 2D. While “ZD7288” and “CsCl” and “latencies” are extracted as the cause and theme respectively for the verb “increased”, the “properties of both open states” is extracted as the theme of the verb “decreased”. Our rule to allow transfer of arguments (either Cause or Theme) if the verbs are in co-ordination and belong to the same category (“Regulation” in this case) enable easy identification of “ZD7288 and CsCl” as the agent for the verb “decreased”.

The above co-ordination rule can handle even more complex co-ordination structures beyond clausal boundaries as shown in example in Figure 2E. Here the co-ordination between the verbs “did not affect” and “did increase” is identified, which triggers the argument transfer rule to help identify “8Br-cAMP” as the cause of the verb “increase”.

Linking sequential events by merging frames - We also link sequential events as conveyed in the text by merging the frames and connecting discourses. Consider the sentence shown in Figure 2F. From that sentence our verb frame based extraction module extracts the following outputs: EVENT1: effects (NO; I (Na)), and EVENT3: inhibition (UNK, both cGMP and cAMP pathways) for the events “effects” and “inhibition” respectively. If we carefully notice on either side of the verb “blocked” we have the nominal form of verb followed by a prepositional phrase. In such cases we connect both the events as EVENT2: (Event3, Event1).

We also have rules to extract lexical chains by handling discourse connectives such as “thereby”, via whereas etc., which are often used to connect two events in the text.



Anaphora resolution - We also have a simple anaphora resolution module to resolve both anaphoric entities and events. Our approach to anaphora resolution for entities is linguistic rules described in Kennedy and Boguraev 1996 (24). For demonstrative NPs such as “this kinase”, “these transcription factors” we consider features such as semantic type of the NPs, the distance

between the antecedent and the candidate anaphora and number (singular or plural form of NP) while deciding the right antecedent. For the anaphoric phrase “both sites” in the following snippet “Dephosphorylated hsp 90 is phosphorylated at both sites by casein kinase II ...”, the candidate antecedents that our method would consider are those phrases which refer to two objects of the type dictated by the head word “site” (protein sites). We look for antecedent phrases which are of the semantic type “protein sites”. In this case, the rule correctly identified the anaphor “serine 231 and serine 263” which appeared in a preceding sentence, “For the alpha protein, these sites correspond to serine 231 and serine 263.” Anaphora resolution plays critical role in recovering the actual arguments as shown in the following example.

Besides resolving anaphors at the entity level we also have rules to resolve event anaphora. Our strategy to resolve event anaphora is based on the identity of the verbs if they have the same root form post-lemmatization. For example, consider the following sentence, “This modulation may contribute to the migratory effect of **MIP1-alpha on microglia**”. The system extracted two outputs for the trigger “contribute” and “effect” as given below: **Event1:** contribute (this modulation, the migratory effect of MIP1-alpha on microglia); **Event2:** effect (MIP1-alpha, microglia). In the first event (Event1) the phrase “*this modulation*” is resolved to as referring to the modulation event, described in the prior sentence, “Thus, microglia in hippocampi from epileptic patients expresses high-conductance Ca<sup>2+</sup>-dependent K<sup>+</sup> channels that are modulated by the chemokine MIP1-alpha”. For the event “modulated” the system extracted the following output: “**Event3: modulated (the chemokine MIP1-alpha, high-conductance Ca<sup>2+</sup>-dependent K<sup>+</sup> channels)**”. After anaphora resolution the system finally gets the consolidated output as **Event1: (Event3, Event2)**.

### 3 Experiments

#### 3.1 Data set

We evaluated the performance of the system by extracting events from PubMed abstracts cited as literature evidence in PharmGKB, and comparing the system output with the manual annotations in the PharmGKB (19). PharmGKB pathway is a rich resource, which catalogs both the pharmacodynamics and pharmacokinetics pathways involving the interplay between the drugs, metabolites and genes through manual curation along with the citation to primary literature evidence namely the PubMed (25). PharmGKB pathway resource’s latest version (As on July 1<sup>st</sup> 2013) contains 99 pathways with citations to primary literature. Besides these it also contains other pathways assembled from other resources such as Reactome(26). In addition to events we evaluated the system for identifying all the participating molecules (genes/chemicals) involved in the pathways. We reported the performance as precision, recall and F-measure. For each event in a pathway we compared the individual fields (see Table3) namely From, To, and ControlledBy against the manual annotations. True positives were required to match all the four fields. For the manual evaluation we considered additional criteria during evaluation. By ignoring the gene

normalization we considered the extraction to be correct if the biology intuition tells that the identified gene mentioned in the text is synonymous to the one in the PharmGKB.

### 3.2 Post-processing the system output to compare against PharmGKB annotation

For the current study we retrieved all the PubMed IDs (1,036) cited as literature evidence in the 99 PharmGKB pathways and retrieved them from PubMed through Entrez batch search (27). We formatted our system's output to generate the annotations in the same format as that of PharmGKB event annotation. In order to further align our gene mentions with the PharmGKB annotation, we normalized the textual mentions of gene/protein to Gene symbols using GeNO (28). We remapped the entity annotations produced by our system with that of GeNO by comparing the output span indices of the two systems. Even if there were overlap in the indices we aligned both the annotations and assigned the Gene symbols identified by GeNO to the corresponding entity mentioned in the text. We mapped the Agent/Cause of the verb extracted by our system to the "Controlled By" field in PharmGKB while the Theme identified by our system is mapped to "From" field. If the theme of the verb did not undergo any transformation in its molecular state through post-translational modifications, metabolism etc. then the same theme is assigned to the "To" field as well. For example consider the sentence (PMID: 11287982)

Table 2. Sample PharmGKB annotation

From	To	Controlled By	Evidence
BCR-ABL	BCR-ABL	imatinib	11287972;12755554;13679030;16122278
imatinib	CGP	CYP1A2;CYP2C19;CYP2C9;CYP2D6;CYP3A4;CYP3A5	15828850;16122278

"**Imatinib** is a potent and selective *inhibitor* of the **protein tyrosine kinase Bcr-Abl, platelet-derived growth factor receptors** (PDGFRalpha and PDGFRbeta) and **KIT**". Imatinib, the agent of the verb "inhibitor" in the above sentence is mapped to the "ControlledBy" field and one of the theme "**Bcr-Abl**" is mapped to the "From" field. Since the verb inhibitor do not involve any transformation of the theme it is also assigned to the "To" field.

### 3.3 Evaluation

We performed two evaluations 1) automated evaluation on all the event descriptions in PharmGKB pathways 2) manual evaluation of event extraction for four selected pathways. The four pathways are Platelet aggregation inhibitor pathway, Warfarin pathway, Metformin pathway, and Aromatase inhibitor pathway. We assessed the utility of our system output in pathway curation. Besides events, we also evaluated the ability of the system to identify all the participating molecules (genes/chemicals) in the pathways. We used the standard metrics namely precision, recall and F-measure for evaluation. For each event in a pathway we compared the individual fields namely From, To, and ControlledBy against the manually curated one and if all the four fields are found to be correct we count them to be a true positive event. Otherwise we count them as both precision and recall error. We did not report the partial recall for the fields correctly identified by the system.

## 4 Results and discussion

### 4.1 Evaluation on complete PharmGKB data set

PharmGKB pathway annotation contains 894 events involving 1040 molecules (839 genes and 201 drugs) annotated from 99 PharmGKB pathways. We evaluated the ability of our system in identifying the molecules participating in events annotated in PharmGKB pathways as shown in Table 3. Out of the two classes of entities the performance of Gene named entity was extremely lower (F-measure: 56.96) as it involve normalizing the gene mentions in the text to Entrez gene symbol as per the requirements of PharmGKB annotations. However for identifying drugs and chemicals the F-measure was fairly high (82.68%) as it doesn't involve entity normalization.

Table 3. Evaluation of system's performance on entity identification on complete PharmGKB

Entity Type	Total Entity (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
Gene	839	632 (419)	66.30	49.94	56.96
Drug/Chemical	201	261 (191)	73.18	95.02	82.68

Table 4 lists the performance of our event extraction system on the 1036 abstracts cited as literature evidence in PharmGKB pathways. The 99 pathways in PharmGKB contain 894 events. Our system identified 952 events from the 1036 abstracts out of which only 323 were found to be correct leading to precision of 33.93%, recall of 36.13% and F-measure of 34.99%. However we observed that extra-sentential processing modules contributed to only 4.5% improvement to the final output. The likely reason may be that PharmGKB annotation of pathway events mostly involves only simple entities such as genes and proteins but not complex events such as biological processes.

Table 4. Evaluation of system's performance on event extraction from PharmGKB

Total Events (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
894	952 (323)	33.93	36.13	34.99

### 4.2 Manual evaluation of four hand-selected pathways

While we expected the recall to be lower we were surprised to observe lower precision, a feature atypical of rule-based systems. In order to better understand the reason behind the low precision we manually evaluated the performance on abstracts related to four hand-selected pathways, which has citations to 34 abstracts as literature evidence. The manual inspection of the system output on these 34 abstracts aimed to identify the reason behind the low recall and precision. We observed the following discrepancies between the extracted output and the gold standard annotation in the PharmGKB:

1) Certain annotations in PharmGKB are not actually present in either the abstract or in the full text article. For example in the Platelet Aggregation inhibitor pathway we have the following annotation in PharmGKB as given in Table 5 below.

We did not find any mention of the individual G-protein in the ControlledBy column either in the cited abstracts or in the full text articles. However, there is a general mention about the involvement of G-proteins from the G-12&13 families, which our system extracted correctly. Out of the total 24 annotations for this pathway in PharmGKB, there were 7 annotations, which do not have direct evidence in the literature considering both the abstract and full text article. Instead they were derived through biological inference. None of these annotations were identified by our system. While from a biologist perspective the annotation in the pathway database is correct we believe that the current state of the art of literature mining has not matured enough to extract such annotations. Inferencing by using the background knowledge from knowledge bases such as PRO, UniProt etc. alone can help resolve such uncertainties.

Table 5. PharmGKB annotation from platelet aggregation pathway

From	To	Controlled By	Evidence
ADCY3	ADCY3	GNA11,GNA12,GNA13,GNA15,GNAI1,GNAI2,GNAI3,GNAQ,GNB3,GNAS	15187029, 11997386

2) Another notable reason for lower recall is that the information in pathway database is synthesized from multiple abstracts while our system extracts information only from a single article.

3) Another observation clearly explains the reasons for the lower precision of the system. Our system extracted a few annotations with no corresponding entries in PharmGKB. On manual inspection we found that while those annotations are not wrong they do not confirm to the event definition of the PharmGKB database. For example from an abstract (PMID: 15187029) the system extracted two relations namely, *regulate (P2Y(12), PPI)* and *inhibit (P2Y(12), adenylate cyclase)* from the sentence “Furthermore, the Src family kinase inhibitor **PP1** selectively potentiates the contribution to the calcium response by **P2Y(12)**, although *inhibition of adenylate cyclase* by **P2Y(12)** is unaffected.” which are not annotated in PharmGKB. While both the relations extracted are correct from the biologist perspective it is not relevant in the context of PharmGKB annotation. The errors in gene normalization (both recall and precision) also contributed to the errors in event extraction as well. Table 6 lists the performance of our system on the selected 4 pathways through manual evaluation with and without ignoring the gene normalization.

Table 6. Evaluation of system’s performance on event extraction on handpicked PharmGKB dataset

Event Type	Total Events (Gold)	Total Extracted (Total correct)	Precision (%)	Recall (%)	F-measure (%)
Event ignoring normalized entities	58	69 (39)	56.52	67.24	61.41
Events with normalized entities	58	41 (25)	60.97	43.13	50.50

The first row in Table 6 corresponds to the evaluation where we considered the event annotations to be considered as correct even if the genes were not normalized to the correct Entrez gene symbols. We used the biological inference to judge if the extracted gene matches the gene definition annotated in PharmGKB. However we wish to clarify if the event is not represented in



the annotation we considered the text extraction to be a false positive as our underlying focus in this study is to evaluate the utility of literature mining in pathway curation. The second row in Table 6 considers the extraction to be correct only if the genes are normalized to the correct gene symbols. We observed an appreciable drop in the recall ( $>20\%$ ) and very little increase in precision ( $\sim 3\%$ ) when we consider gene normalized events, which illustrates that it is an important limitation in the performance of standardizing event extraction. Another limitation that we would like to point out is that our system being a rule-based one may require substantial manual effort to tune it to scale and improve its performance further.

## 5 Conclusions and future directions

Despite these limitations we believe that in this study we have made sincere efforts to explore and understand the limitations of a literature mining system in the context of extracting event descriptions which will be useful in finding literature evidences for actual pathway curation in a limited context of PharmGKB database. Our results are substantially lower than the recently reported studies (2, 3). However it is not fair to compare the performance of the system evaluated in this study with that of other systems as there is significant difference in the evaluation schema itself. Most of the previous studies evaluate the event annotation capability against the annotations at the textual level either abstracts (4, 15, 29) or full-text articles (30) aimed at benchmarking the text mining effort. However in this study, we explored the comparison of text-based extraction against events annotated in an independently curated pathway knowledge base. The performance of our system is comparable to the other state of the art system against text-based annotations (2, 3). This study further allowed us to identify the gaps between the current state of the art in literature mining and the demands of text mining assisted pathway curation. However we believe that our current system will be useful for finding the evidence needed for curation of the pathways. We plan to explore the following steps to improve text mining assisted pathway curation:

- 1) Improve the state of the art in gene normalization, which we hope to improve since we are working on this task in parallel for the BioCreative 4 Track3 (31);
- 2) Explore hybrid approaches by combining the rule-based system with machine learning approach to reduce the amount of manual effort required to tune the systems to new data sets;
- 3) Understand the pathway curation workflow and design annotation schema and corpora for pathway curation. The current available corpora limit the annotation to single abstracts or articles. Quite often we need to synthesize information across articles. But we realize that it is not possible without the understanding the pathway curation workflow;
- 4) Assess the needs of pathway curators to set more realistic and achievable text mining goals. We realize that working closely with the database curators and building an intuitive interface to facilitate pathway curation will not only help us understand the curation workflow but also help improve the state of the art in literature mining significantly.

## 6. Acknowledgments

The authors acknowledge that the study was supported by two grants: National Science Foundation ABI:0845523 and National Library of Medicine R01LM009959 grants. The authors also acknowledge the support received from Centre for Individualized Medicine, Mayo Clinic.

## References

1. Baumgartner WA, Cohen KB, Fox LM, Acquah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*. 2007;**23**(13):i41-i8.
2. Björne J, Ginter F, Pyysalo S, Tsujii Ji, Salakoski T. Complex event extraction at PubMed scale. *Bioinformatics*. 2010;**26**(12):i382-i90.
3. Björne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T. Extracting complex biological events with rich graph-based feature sets. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*; 2009: Association for Computational Linguistics; 2009. p. 10-8.
4. Liu H, Komandur, R., Verspoor, K. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In *proceedings, BioNLP-ST'11 Workshop*; 2011; 2011.
5. BioNLP Shared Task 2013. [cited; Available from: <http://2013.bionlp-st.org/>]
6. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*. 2005;**21**(suppl 2):ii252-ii8.
7. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC bioinformatics*. 2004;**5**(1):147.
8. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*. 2003;**19**(16):2155-7.
9. Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of biomedical informatics*. 2004;**37**(1):43-53.
10. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics*. 2005;**6**(Suppl 1):S1.
11. Lu Z, Kao H, Wei C, et al. The Gene Normalization Task in BioCreative III. *BMC Bioinformatics*. 2011; **12**(Suppl 8):S2.
12. Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii Ji. Overview of bionlp shared task 2011. *Proceedings of the BioNLP Shared Task 2011 Workshop*; Association for Computational Linguistics; 2011. p. 1-6.
13. Kim JD, Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J. Overview of BioNLP'09 shared task on event extraction. *Proceedings of the Workshop on BioNLP: Shared Task*; 2009; 2009. p. 1-9.
14. Bandy J, Milward D, McQuay S. Mining protein–protein interactions from published literature using Linguamatics I2E. *Protein Networks and Pathway Analysis*; Springer; 2009. p. 3-13.

15. Van Landeghem S, Ginter F, Van de Peer Y, Salakoski T. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. Proceedings of BioNLP 2011 Workshop; 2011: Association for Computational Linguistics; 2011. p. 28-37.
16. Kemper B, Matsuzaki T, Matsuoka Y, et al. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*. 2010;**26**(12):i374.
17. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*. 2012;**13**(1):108.
18. Schmidt CJ, Sun L, Arighi CN, et al. Pathway curation: Application of text-mining tools eGIFT and RLIMS-P. *Bioinformatics and Biomedicine Workshops (BIBMW)*, 2012 IEEE International Conference on; 2012: IEEE; 2012. p. 523-8.
19. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic acids research*. 2002;**30**(1):163-5.
20. Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*. 1995;**21**(4):543-65.
21. Kim JD, Ohta, T., Tateisi, Y., Tsujii, J. GENIA corpus : a semantically annotated corpus for bio-textmining. *Bioinformatics*. 2003;**19**(Suppl1):i180-i2.
22. Florian R, Ngai G. Fast transformation-based learning toolkit. Johns Hopkins University, <http://nlp.cs.jhu.edu/rflorian/fntbl/documentation.html>; 2001.
23. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. *Pac Symp Biocomput*; 2003; p. 427-438.
24. Kennedy C, Boguraev B. Anaphora for everyone: pronominal anaphora resolution without a parser. Proceedings of the 16th conference on Computational linguistics-Volume 1; 1996: Association for Computational Linguistics; 1996. p. 113-8.
25. PubMed database. [cited; Available from: <http://www.ncbi.nlm.nih.gov/pubmed/>
26. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*. 2005;**33**(suppl 1):D428-D32.
27. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: Molecular biology database and retrieval system. *Methods in enzymology*. 1996;**266**:141-62.
28. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. *Bioinformatics*. 2009;**25**(6):815-21.
29. Liu H, Hunter L, Kešelj V, Verspoor K. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. *PloS one*. 2013;**8**(4):e60954.
30. Garten Y, Altman R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*. 2009;**10**(Suppl 2):S6.
31. Biocreative IV 2013 [cited; Available from: <http://www.biocreative.org/tasks/biocreative-iv/track-3-CTD/>

# SPARSE GENERALIZED FUNCTIONAL LINEAR MODEL FOR PREDICTING REMISSION STATUS OF DEPRESSION PATIENTS

YASHU LIU<sup>†</sup>, ZHI NIE<sup>†</sup>, JIAYU ZHOU<sup>†</sup>, MICHAEL FARNUM<sup>‡</sup>, VAIBHAV A NARAYAN<sup>‡</sup>, GAYLE  
WITTENBERG<sup>‡</sup>, JIEPING YE<sup>†</sup>

<sup>†</sup>*Department of Computer Science and Engineering,  
Center for Evolutionary Medicine and Informatics, The Biodesign Institute,  
Arizona State University, Tempe, AZ 85287, USA*

<sup>‡</sup>*Johnson & Johnson Pharmaceutical Research & Development, LLC,  
Titusville, NJ, USA*

*E-mail: <sup>†</sup>{Yashu.Liu, Zhi.Nie, Jiayu.Zhou, Jieping.Ye}@asu.edu,  
<sup>‡</sup>{MFARNUM, VNaray16, GWittenb}@its.jnj.com*

Complex diseases such as major depression affect people over time in complicated patterns. Longitudinal data analysis is thus crucial for understanding and prognosis of such diseases and has received considerable attention in the biomedical research community. Traditional classification and regression methods have been commonly applied in a simple (controlled) clinical setting with a small number of time points. However, these methods cannot be easily extended to the more general setting for longitudinal analysis, as they are not inherently built for time-dependent data. Functional regression, in contrast, is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. In this paper, we propose a novel sparse generalized functional linear model for the prediction of treatment remission status of the depression participants with longitudinal features. Compared to traditional functional regression models, our model enables high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. Extensive experiments have been conducted on the Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) data set and the results show that the proposed sparse functional regression method achieves significantly higher prediction power than existing approaches.

*Keywords:* Depression, generalized functional linear model, STAR\*D, longitudinal analysis, fused Lasso, group Lasso

## 1. Introduction

The increasing life expectancy of the worldwide population has led to a growing number of patients with serious mental disease such as depression. Research on the diagnosis and prognosis of these diseases has received increasing attention in the biomedical domain. Depression, or major depression (MD) is a common mental disorder affecting estimated 350 million people worldwide, featured by symptoms such as depressed mood, loss of interest or pleasure, feelings of guilt or low self-worth.<sup>1</sup> It is expected to be the second leading cause of disability worldwide.<sup>2</sup> Though the efficacy of several antidepressant medications and therapies has been proven, a universal and long-term treatment of MD has not been well explored due to its high risk of relapses and recurrences.<sup>3</sup>

Like many other mental conditions, major depression affects people over time and it is notorious for the chronicity. Thus, the analysis of longitudinal data is one crucial step towards the understanding and prognosis of major depression. One valuable resource for such research

is the Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) trial initiated by National Institute of Mental Health (NIMH), which was originally designed for seeking the optimal combination and sequence of treatment strategies for non-psychiatric depressed patients.<sup>3</sup> Based on the evaluation of the therapeutic responses, participants in STAR\*D may receive up to 4 levels of treatments and their information such as symptomatic status, daily functioning, treatment side effects is collected during every clinical visit.

In STAR\*D, a range of clinical scales have been applied to evaluate or describe the severity of diseases. For instance, the 17-item Hamilton Rating Scale for Depression (HRSD<sub>17</sub>) is collected via telephone interview for research purposes.<sup>3</sup> The 16-item Quick Inventory of Depressive Symptomatology - Clinician Rated (QIDS-C<sub>16</sub>) provides the evidences for clinicians to decide whether the patients proceed to the next treatment level.<sup>3</sup> Exploring the longitudinal relationship between clinical measurements (input features) and therapeutic responses (outcomes) and detecting features with significant statistical power are two fundamental and important research questions. Several tools based on machine learning techniques have been developed for longitudinal study.<sup>4-7</sup>

In our paper, we adopt sparse functional regression for the longitudinal data analysis. Functional data (FD) refers to the data samples whose features are viewed as random functions or surfaces over one or more continuum such as time, spatial location.<sup>8,9</sup> For instance, the average daily temperatures observed in a weather station can be viewed as a functional data sample over time; the intensity or color composition of a brain image can be taken as a functional sample over spatial location.<sup>9</sup> Functional data analysis (FDA), an important branch of statistics, is referred to the statistical analysis built on functional data, where the random functions are assumed to be independent and smooth.<sup>8-10</sup> As the extension of classic regression methods to functional data, functional regression is used to estimate the relationship among functional features. Variant forms of functional regression are applicable in different problem setups. For example, it can be applied for regressing functional outcomes on scalar features;<sup>9,11-13</sup> it can also be applied on estimating relationships between functional features and scalar outcomes.<sup>14-16</sup> Under the assumption that the functional coefficient is sparse over time, the FLiRTI model was proposed by James *et al.*<sup>15</sup> and it showed better predictive power than regular functional regression models. However, the FLiRTI model is only limited to the settings with one functional feature. For higher flexibility, multivariate functional regression models were developed. To enhance interpretability of the multivariate functional regression model, Zhu *et al.*<sup>17</sup> and Gertheiss *et al.*<sup>18</sup> applied the group Lasso type constraint for curve (functional feature) selection. Zhu *et al.*<sup>17</sup> combined both functional features and scalar features together in their model, however it imposed smoothness of coefficient functions only by controlling the number of basis functions. Gertheiss *et al.*<sup>18</sup> introduced the sparsity-smoothness penalty for simultaneously selecting functional feature and controlling the smoothness of the coefficient functions, however it does not incorporate extra scalar features or achieve sparse feature effects over time. Fan *et al.*<sup>19</sup> proposed a functional additive regression (FAR) model which managed functional feature selection via concave penalties in both linear and non-linear settings, while the resulting solutions are not interpretable in term of functional feature effects over time. Therefore, there is a need to develop a general and interpretable formulation

of functional regression that simultaneously achieves functional feature selection, smoothness of functional coefficients and interpretable estimation of functional coefficients.

In this paper, we propose a novel sparse generalized functional linear model for longitudinal biomedical data analysis, which can be applied to predict the disease status based on longitudinal features. Specifically, we empower basic functional regression models to simultaneously identify features with significant predictive power across time points with the group Lasso penalty,<sup>20</sup> enforce smoothness of functional coefficients with the fused Lasso penalty<sup>21</sup> and achieve interpretable estimations of functional coefficients with the Lasso penalty.<sup>22</sup> Since the unknown coefficient matrix is a multiplication factor of the penalized term, the proposed formulation is challenging to solve. Our proposed algorithm integrates the Alternating Direction Method of Multipliers (ADMM)<sup>23</sup> and the accelerated gradient method (AGM)<sup>24,25</sup> to estimate the unknown coefficient matrix. We demonstrate the effectiveness and flexibility of the proposed formulations for longitudinal data analysis using STAR\*D data. Experimental results show that the proposed method achieves better prediction performance with longitudinal features than existing approaches.

The rest of the paper is organized as follows. We briefly introduce FDA and the basic functional regression model in section 2. We propose a novel sparse generalized functional linear model and present the algorithm to solve the proposed formulations in section 3. In section 4, we evaluate the proposed sparse generalized functional regression model on STAR\*D data and report the experimental results. We conclude our paper in section 5.

## 2. Basics of Functional Regression

### 2.1. Functional Data Analysis

Functional data is usually assumed to be generated by an underlying smooth function. In practice, a functional data sample consists of sequences of numerical values (or vectors) varying over a certain continuum. For instance, the series of QIDS-C<sub>16</sub> scores of a depression patient over his/her visiting time can be considered as a functional data sample. Fig. 1 gives an illustration of functional data. Sequences of QIDS-C<sub>16</sub> scores of 6 depression patients are recorded over 14 weeks. In the functional context, we assume each sequence of QIDS-C<sub>16</sub> scores is generated by an underlying function varying over time  $t$  (weeks).

One important issue in FDA is to recover the underlying function based on the sequences of observed numerical values. A common approach is to express the underlying function by a linear combination of basis functions using smoothing techniques. Specifically, given the evaluations of basis functions over time as features and the observed numerical values as outcomes, the coefficients of basis functions can be fitted using least square methods with roughness penalty.<sup>9</sup> However, the basis smoothing technique is only effective when the functional data is observed continuously or densely. When it comes to longitudinal data, observations are always sparse and irregular. Rice *et al.*<sup>26</sup> contrasted and compared FDA with longitudinal data analysis (LDA). James *et al.*<sup>27</sup> and Yao *et al.*<sup>28</sup> connected FDA and LDA by proposing approaches that estimate the underlying function of sparsely and irregularly observed functional data by exploiting both population and individual information. The former extended the basis smoothing technique with mixed effect models while the latter proposed the “PACE” method

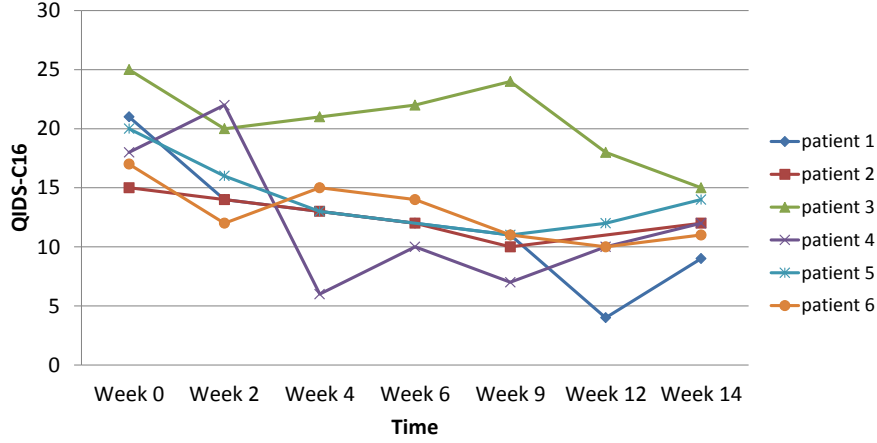


Fig. 1. Illustration of functional data. The QIDS-C<sub>16</sub> scores of 6 patients are recorded over 14 weeks. In the functional context, we assume each sequence of QIDS-C<sub>16</sub> scores is generated by an underlying function.

which involves the kernel smoothing technique and computing the conditional expectation. In our paper, we adopt PACE to estimate the underlying function of functional data.

## 2.2. Functional Regression Model with Functional Features and Scalar Outcomes

In classic statistical analysis, regression methods play an important role in analyzing the relationship between features (independent variables) and outcomes (dependent variables). FDA extends the philosophy of classic regression to functional data and develops functional regression which involves various models for different purposes.

When the features are functional and the outcomes are scalar, we have

$$Y_i = \alpha + \int_{\Omega_t} X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where  $n$  is the total number of samples,  $Y_i$  is a scalar outcome of the  $i$ th sample,  $X_i(t)$  is a 1-dimensional functional feature of the  $i$ th sample,  $\beta(t)$  is a functional coefficient,  $\Omega_t$  is the domain of continuum  $t$ , scalar  $\alpha$  is a bias term, and  $\epsilon_i$  corresponds to the scalar residual. Note that the model above only allows one functional feature, which greatly limits its application. A simple but useful extension to multiple functional features is

$$Y_i = \alpha + \sum_{k=1}^p \int_{\Omega_t} X_{ik}(t)\beta_k(t)dt + \epsilon_i, \quad i = 1, \dots, n, \quad (2)$$

where  $X_{ik}(t)$  is the  $k$ th functional feature of the  $i$ th sample, and  $\beta_k(t)$  is the functional coefficient corresponding to the  $k$ th functional feature.

## 3. Proposed Sparse Functional Regression Models

In this section, we propose a novel sparse generalized functional linear model which simultaneously selects useful features, enforces smoothness of functional coefficients and achieves interpretable estimations of functional coefficients. Suppose there are  $N$  samples and each sample has  $d$  scalar features  $s_1, s_2, \dots, s_d$  and  $p$  functional features  $x_1(t), x_2(t), \dots, x_p(t)$ . Then, for

the  $i$ th sample, we denote  $\mathbf{s}_i = (s_{i1}, s_{i2}, \dots, s_{id})^T$  and  $\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{ip}(t))^T$ . In matrix form, we let  $S \in \mathbb{R}^{N \times d}$  be the scalar data matrix with each row as a sample of  $d$  scalar features, *i.e.*,  $S_{i,\cdot} = \mathbf{s}_i^T = (s_{i1}, s_{i2}, \dots, s_{id})$ , and let  $X(t)$  be an  $N \times p$  matrix of functions where the  $i$ th row denotes the  $i$ th sample of  $p$  functional features *i.e.*,  $X_{i,\cdot}(t) = \mathbf{x}_i(t)^T = (x_{i1}(t), x_{i2}(t), \dots, x_{ip}(t))$ . Let  $\mathbf{y} = (y_1, y_2, \dots, y_N)^T \in \mathbb{R}^{N \times 1}$  be the vector of scalar outcomes, and  $\mathbf{w} = (w_1, w_2, \dots, w_d)^T$  be the coefficients of  $d$  scalar features and  $\mathbf{b}(t) = (\beta_1(t), \beta_2(t), \dots, \beta_p(t))^T$  be the vector of  $p$  functional coefficients. Moreover, we assume the functional coefficients can be represented by a set of  $k_b$  basis functions  $\Theta(t) = (\theta_1(t), \theta_2(t), \dots, \theta_{k_b}(t))^T$ , *i.e.*,  $\mathbf{b}(t) = B\Theta(t)$ , where  $B \in \mathbb{R}^{p \times k_b}$ . Then, for a known link function  $g(\cdot)$ , we have the generalized functional linear model

$$g(y_i) = \alpha + \sum_{g=1}^d s_{ig} w_g + \sum_{h=1}^p \int_{\Omega_t} x_{ih}(t) \beta_h(t) dt = \alpha + \mathbf{s}_i \mathbf{w} + \int_{\Omega_t} \mathbf{x}_i(t) \mathbf{b}(t) dt. \quad (3)$$

In matrix form, we have

$$g(\mathbf{y}) = \alpha \mathbf{1} + S \mathbf{w} + \int_{\Omega_t} X(t) \beta(t) dt = \alpha \mathbf{1} + S \mathbf{w} + \int_{\Omega_t} X(t) B \Theta(t) dt, \quad (4)$$

where  $\mathbf{1} \in \mathbb{R}^{N \times 1}$  is a column vectors of ones. When  $g(\cdot)$  is the identity function, *i.e.*,  $g(u) = u$ , the proposed model is functional linear regression. Then the optimization procedure involves minimizing the quadratic loss. When  $g(\cdot)$  is the sigmoid function, the proposed model turns out to be a functional logistic regression, *i.e.*,

$$\text{Prob}(\mathbf{y}|S, X) = \frac{1}{1 + \exp \left( -\mathbf{y} \odot \left( \alpha \mathbf{1} + S \mathbf{w} + \int_{\Omega_t} X(t) B \Theta(t) dt \right) \right)}, \quad (5)$$

where “ $\odot$ ” denotes the componentwise multiplication. Let  $\Theta \in \mathbb{R}^{k_b \times T}$  be the evaluation matrix of  $\Theta(t)$  at  $T$  time points, where  $\Theta_{\cdot,t} \in \mathbb{R}^{k_b \times 1}$  corresponds to the evaluation at time  $t$ . Then the unknown coefficients  $\alpha$ ,  $\mathbf{w}$  and matrix  $B$  can be obtained by minimizing the average logistic loss (negative log-likelihood function),

$$\mathcal{L}(\alpha, \mathbf{w}, B) = \frac{1}{N} \sum_{i=1}^N \log (1 + \exp (-y_i (\alpha + \mathbf{s}_i \mathbf{w} + \sum_t X(t) B \Theta_{\cdot,t}))). \quad (6)$$

The logistic loss is convex and smooth and can be solved via standard optimization methods.

When  $\beta_j(t) = 0$ , the changes of the  $j$ th functional feature has no effect on the outcome at time  $t$ . We apply the Lasso penalty<sup>22</sup> on  $B\Theta$ , *i.e.*,  $\|B\Theta\|_1 = \sum_{j,t} |B_{j,\cdot} \Theta_{\cdot,t}|$ , resulting in interpretable estimations, *i.e.*, many entries of  $B\Theta$  are zero. That is, the changes of many features have no effects on prediction at some time points. For feature selection purpose, we also introduce the group Lasso penalty<sup>20</sup> with  $\|B\Theta\|_{2,1} = \sum_{j=1,\dots,p} \|B_{j,\cdot} \Theta\|_2$  which enforces many rows of  $B\Theta$  to be zero. If the  $j$ th row of matrix  $B\Theta$  is zero, then the  $j$ th feature has no predictive power along with time. In addition, we employ the fused Lasso penalty<sup>21</sup> on matrix  $B\Theta$  to enforce the smoothness of the functional coefficients. The resulting sparse functional logistic regression model with scalar features and functional outcomes can be obtained by

$$\min_{\alpha, \mathbf{w}, B} \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|B\Theta\|_1 + \lambda_3 \|B\Theta R\|_1 + \lambda_4 \|B\Theta\|_{2,1}, \quad (7)$$

where  $R$  is a  $T$  by  $T - 1$  sparse matrix with  $R_{j,j} = 1, R_{j+1,j} = -1$ , and  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  are the tuning parameters. We solve problem (7) by alternately minimizing over  $\alpha$  and  $\mathbf{w}$  with  $B$



fixed, and minimizing over  $B$  with  $\alpha$  and  $\mathbf{w}$  fixed. Note that the penalty of  $\mathbf{w}$  only involves the Lasso penalty, and we have already known that the optimization in terms of  $\mathbf{w}$  (*i.e.*, sparse logistic regression problem) can be solved efficiently by the Accelerated Gradient Methods (AGM).<sup>24,25,29</sup>

The proposed sparse functional logistic regression model is much more challenging to solve than usual multi-task learning algorithms since the unknown coefficient matrix  $B$  is a multiplication factor of the penalized term  $B\Theta$ . In this paper, we integrate the Alternating Direction Method of Multipliers (ADMM)<sup>23</sup> and AGM<sup>24,25</sup> to solve  $B$ . When  $\alpha$  and  $\mathbf{w}$  are fixed, we write the objective (7) in the following form:

$$\begin{aligned} \min_B \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_2 \|Z\|_1 + \lambda_3 \|ZR\|_1 + \lambda_4 \|Z\|_{2,1} \\ \text{s.t. } B\Theta = Z. \end{aligned} \quad (8)$$

Then, the augmented Lagrangian function is given by

$$L_\rho(B, Z, \xi) = \mathcal{L}(\alpha, \mathbf{w}, B) + \lambda_2 \|Z\|_1 + \lambda_3 \|ZR\|_1 + \lambda_4 \|Z\|_{2,1} + \langle \xi, B\Theta - Z \rangle + \frac{\rho}{2} \|B\Theta - Z\|_F^2, \quad (9)$$

where  $\xi \in \mathbb{R}^{p \times T}$  is the lagrangian dual variable and  $\rho$  is a penalty parameter. The ADMM-based procedures for solving the unknown matrix  $B$  in the proposed sparse functional logistic regression at the  $k$ th iteration can be described as follows:<sup>23</sup>

$$B^{(k)} := \min_B \mathcal{E}(B) = \min_B \mathcal{L}(\alpha^{(k)}, \mathbf{w}^{(k)}, B) + \langle \xi^{(k-1)}, B\Theta - Z^{(k-1)} \rangle + \frac{\rho}{2} \|B\Theta - Z^{(k-1)}\|_F^2, \quad (10)$$

$$Z^{(k)} := \min_Z \lambda_2 \|Z\|_1 + \lambda_3 \|ZR\|_1 + \lambda_4 \|Z\|_{2,1} + \langle \xi^{(k-1)}, B^{(k)}\Theta - Z \rangle + \frac{\rho}{2} \|B^{(k)}\Theta - Z\|_F^2, \quad (11)$$

$$\xi^{(k)} := \xi^{(k-1)} + \rho(B^{(k)}\Theta - Z^{(k)}). \quad (12)$$

For the  $B$ -update step (10), the unknown matrix  $B$  can be solved by the accelerated gradient descent method<sup>24,25,29</sup> with gradient

$$\nabla_B \mathcal{E}(B) = -\frac{1}{N} U^T (\mathbf{1} - \mathbf{p}) + \rho(B\Theta + \frac{\xi^{(k-1)}}{\rho} - Z^{(k-1)})\Theta^T,$$

where

$$\begin{aligned} U &= [y_1 \sum_t X_{1,\cdot}(t)^T \Theta_{\cdot,t}^T, y_2 \sum_t X_{2,\cdot}(t)^T \Theta_{\cdot,t}^T, \dots, y_N \sum_t X_{N,\cdot}(t)^T \Theta_{\cdot,t}^T]^T, \\ \mathbf{p} &= \mathbf{1} ./ \left( \mathbf{1} + \exp \left( -\mathbf{y} \odot \left( \alpha^{(k)} + S\mathbf{w}^{(k)} + \sum_t X(t)B\Theta_{\cdot,t} \right) \right) \right), \end{aligned}$$

and “./” denotes the componentwise division. For the  $Z$ -update step (11), it has been shown that the proximal operator can be solved efficiently in two stages as stated in the following lemma,<sup>30</sup>

**Lemma 3.1.** *Given vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{1 \times T}$ , penalty parameters  $\gamma_1, \gamma_2, \gamma_3$ , and the sparse matrix  $R$  defined as above, let*

$$\begin{aligned} \mathcal{F}(\mathbf{v}) &= \arg \min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_1 \|\mathbf{u}\|_1 + \gamma_2 \|\mathbf{u}R\|_1, \\ \mathcal{G}(\mathbf{v}) &= \arg \min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_3 \|\mathbf{u}\|_2, \\ \mathcal{FG}(\mathbf{v}) &= \arg \min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 + \gamma_1 \|\mathbf{u}\|_1 + \gamma_2 \|\mathbf{u}R\|_1 + \gamma_3 \|\mathbf{u}\|_2. \end{aligned} \quad (13)$$

Then, the following relationship holds

$$\mathcal{FG}(\mathbf{v}) = \mathcal{G}(\mathcal{F}(\mathbf{v})). \quad (14)$$

The details of the proposed algorithm for solving (7) are summarized in Algorithm 1. Steps 3 and 4 are solved by AGM, and readers may refer to Liu *et al.*<sup>29</sup> for more details. In terms of solving the fused Lasso problem (step 6), readers may refer to Liu *et al.*<sup>31</sup> The ADMM parameter  $\rho$  is fixed as a constant in our experiment.

---

**Algorithm 1** Sparse Functional Logistic Regression

---

**Input:**  $\mathbf{y} \in \mathbb{R}^{N \times 1}$ ,  $S \in \mathbb{R}^{N \times d}$ ,  $X_{(j)} \in \mathbb{R}^{N \times p}$ ,  $j = 1, \dots, T$ ,  $\Theta \in \mathbb{R}^{k_b \times T}$ ,  $R \in \mathbb{R}^{T \times (T-1)}$ ,  $\rho \in \mathbb{R}$

**Output:**  $\alpha \in \mathbb{R}$ ,  $\mathbf{w} \in \mathbb{R}^{d \times 1}$ ,  $B \in \mathbb{R}^{p \times k_b}$

- 1: Initialize starting points  $\alpha^{(0)}$ ,  $\mathbf{w}^{(0)}$ ,  $B^{(0)}$ ,  $\xi^{(0)}$ ,  $Z^{(0)}$
  - 2: **for**  $k = 1 : \mathcal{K}$  **do**
  - 3:    $(\alpha^{(k)}, \mathbf{w}^{(k)}) := \arg \min_{\alpha, \mathbf{w}} \mathcal{L}(\alpha, \mathbf{w}, B^{(k-1)}) + \lambda_1 \|\mathbf{w}\|_1$
  - 4:    $B^{(k)} := \arg \min_B \mathcal{L}(\alpha^{(k)}, \mathbf{w}^{(k)}, B) + \langle \xi^{(k-1)}, B\Theta - Z^{(k-1)} \rangle + \frac{\rho}{2} \|B\Theta - Z^{(k-1)}\|_F^2$
  - 5:   **for**  $i = 1 : p$  **do**
  - 6:      $u_i := \arg \min_z \frac{\rho}{2} \|B_{i,\cdot}^{(k)} \Theta + \xi_{i,\cdot}^{(k-1)} / \rho - z\|_2^2 + \lambda_2 \|z\|_1 + \lambda_3 \|zR\|_1$
  - 7:      $Z_{i,\cdot}^{(k)} := \arg \min_z \frac{\rho}{2} \|u_i - z\|_2^2 + \lambda_4 \|z\|_2$
  - 8:   **end for**
  - 9:    $\xi^{(k)} := \xi^{(k-1)} + \rho(B^{(k)}\Theta - Z^{(k)})$
  - 10: **end for**
- 

## 4. Experiments

In this section, we evaluate the proposed sparse functional logistic regression model on the STAR\*D data set. We use the functional data analysis code<sup>a</sup> for the construction and evaluation of basis functions. We also use the PACE package<sup>b</sup> for estimating the underlying smooth functions of functional data.

### 4.1. STAR\*D Data Set

The STAR\*D project consists of four treatment levels aimed to help outpatients achieve depressive symptom remission with measurement-based care treatment.<sup>32</sup> Throughout STAR\*D study, the QIDS-C<sub>16</sub> score, which measures the general symptoms of depression, provides clinicians evidences for deciding the remission status of patients.<sup>32</sup> All the participants enrolled to STAR\*D receive the same antidepressant treatment at level 1, where the selective serotonin reuptake inhibitor citalopram is used. If the participant's therapeutic response is satisfactory, *i.e.*, QIDS-C<sub>16</sub>  $\leq 5$ , he/she will be recommended to the follow-up phase. If the initial therapy is not sufficiently effective on the participants, they will be recommended to the level 2 treatment. At level 2, participants will enter into a set of randomized clinical trials. In a similar

<sup>a</sup><http://www.psych.mcgill.ca/misc/fda/software.html>

<sup>b</sup><http://www.stat.ucdavis.edu/PACE/>

manner, participants who fail to achieve satisfactory responses will enter level 2A, level 3 and up to level 4. In this paper, we concentrate on the longitudinal analysis of the data collected from level 1 and level 2. Since all the participants receive the same treatment at level 1, the questions we aim to address in this paper are: Can we use the level 1 information to predict the participant's remission status at level 2? Which features are most important for the prediction of remission status? How do the important longitudinal features affect the prediction over time?

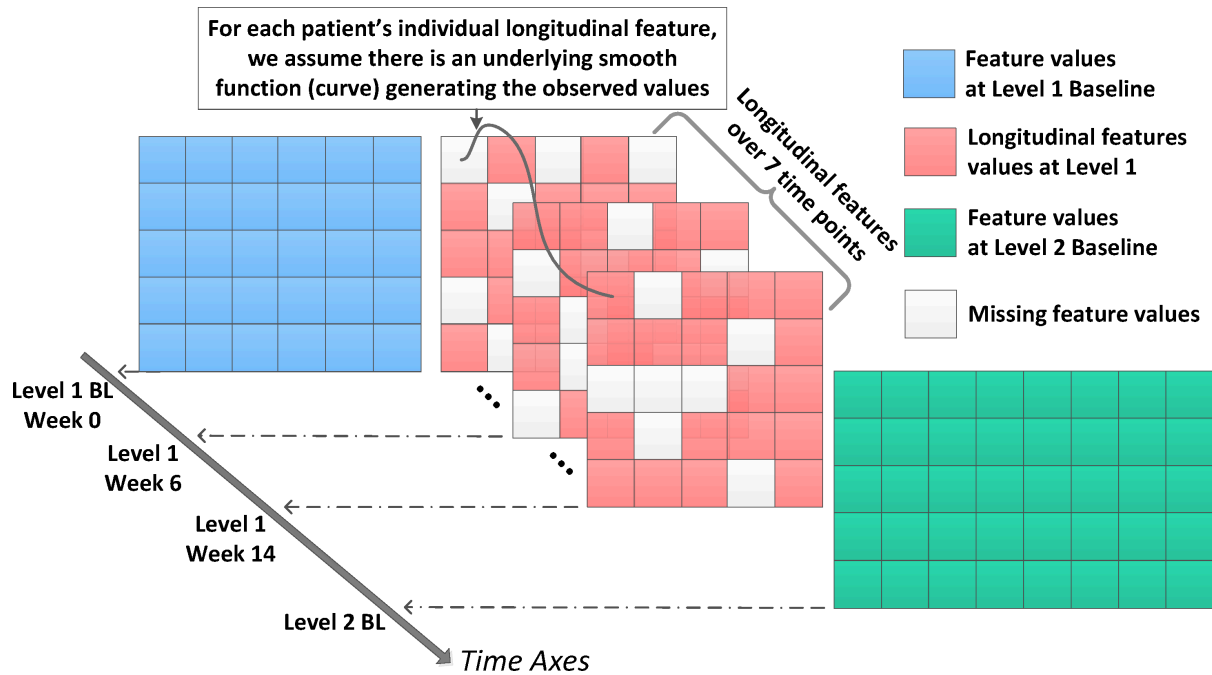


Fig. 2. This figure gives a brief description of the STAR\*D data used in our experiment. In our experiment, we use the scalar features at level 1 baseline and level 2 baseline and the longitudinal features at level 1. Note that the longitudinal data is observed sparsely and irregularly. Therefore, if we align the longitudinal data in a regular time grid as shown in the figure, some feature values will appear as the “missing” values. In our experiment, we first estimate the underlying curves using PACE<sup>28</sup> and then evaluate them on a dense grid of time points.

At each level of treatment, clinical visit information including medication names and doses, side effect intensity, burden and frequency, is collected every 2 or 3 weeks during the acute treatment stage. In our study, we name the time point based on the duration time from baseline to the clinical visit, *e.g.*, “W2” refers to the time point 2 weeks after baseline. At level 1, there are 7 time points scheduled for the regular visits, *i.e.*, W0, W2, W4, W6, W9, W12, W14. Fig. 2 gives a brief description of the STAR\*D used in our experiment. The red matrices in Fig. 2 refer to the longitudinal data stated above. Besides the longitudinal data, the enrolment information involving Cumulative Illness Rating Scale, demographics, HRSD, Medication History, Protocol Eligibility, Psychiatric History, and some level 1 W0 information without further follow-up are also available for the prediction. Those information refers to the level 1 baseline scalar features and corresponds to the blue matrix in Fig. 2. Moreover, we also

want to test the predictive power of the information collected at level 2 baseline (*i.e.*, week 0 at level 2) since the participants did not receive new treatments at that time. These level 2 baseline scalar features are diagramed as the green matrix in Fig. 2.

In STAR\*D, there are 730 participants in total entering level 2 with their level 1 longitudinal information recorded until W12 or W14. After eliminating extremely sparse observations, we have a total number of 596 samples with 202 longitudinal features (LV1.L) available for analysis. There are 1438 level 1 baseline scalar features (LV1.S) and 1667 level 2 baseline scalar features (LV2.S) where the missing values for categorical features are imputed by zeros and the missing values for continuous features are imputed by the mean values. For research purposes, the 16-item Quick Inventory of Depressive Symptomatology – Self-Report (QIDS-SR<sub>16</sub>) is used as outcomes in many existing studies.<sup>32</sup> In our experiments, we also adopt QIDS-SR<sub>16</sub> as the criterion of remission, and define QIDS-SR<sub>16</sub>  $\leq 5$  as remission and a QIDS-SR score of  $> 5$  as non-remission. We evaluate our proposed sparse functional logistic regression model on differentiating the remission cohort from non-remission cohort with available level 1 and level 2 features. Moreover, we classify the remission samples and a subgroup of non-remission samples whose QIDS-SR<sub>16</sub>  $\geq 11$ ; this subgroup is sometimes referred to as severe depression.<sup>32</sup> Detailed sample statistics are shown in Table 1.

Table 1. The sample statistics of the STAR\*D data used in our experiments. Group (All) refers to all qualified samples for the experiment; and Group (Sub) refers to the remaining samples after removing those with QIDS-SR<sub>16</sub> between 6 and 10.

Cohort	Remission (+)	Non-remission (–)	Total	Data Name	Dim
Group (All)	240	356	596	LV1.S	1438
Group (Sub)	240	161	401	LV1.L	202
				LV2.S	1667

#### 4.2. Predicting Remission Status at Level 2

We compare the proposed sparse functional logistic regression with two classic multivariate classifiers, *i.e.*, Random Forest and sparse logistic regression on exactly the same training and testing sets. Our report presents the average accuracy, sensitivity and specificity and the corresponding standard deviations obtained from the 5-fold cross-validation. In all the experiments, both the parameters of sparse functional logistic regression and the sparse logistic regression are tuned via 5-fold cross-validation in the training process. We use B-spline basis functions in our proposed sparse functional logistic regression model, which are the common choice for approximating non-periodic functions.<sup>9</sup>

We first conduct the classification experiment on the level 1 longitudinal data. For classifiers such as Random Forest and sparse logistic regression, the input data is the the average of the longitudinal features over time. The detailed report is shown in Table 2. Compared with Random Forest and sparse logistic regression, the classification performance achieved by the proposed sparse functional logistic regression is consistently better demonstrating the effectiveness of the sparse functional logistic regression in capturing the temporal information. In addition, we observe that, when the sparse functional logistic regression is applied, the level

Table 2. Comparisons of classification performance between Random Forest, sparse logistic regression and sparse functional logistic regression on the longitudinal STAR\*D data.

Experimental Results Using All Samples			
	LV1.L	LV1.S+LV1.L	LV1.S+LV1.L+LV2.S
Random Forest			
Accuracy (%)	68.63 $\pm$ 1.16	69.47 $\pm$ 2.23	68.46 $\pm$ 3.12
Sensitivity(%)	68.83 $\pm$ 3.31	68.83 $\pm$ 4.28	67.71 $\pm$ 5.17
Specificity(%)	68.33 $\pm$ 2.72	70.42 $\pm$ 2.72	69.58 $\pm$ 2.38
Sparse Logistic Regression			
Accuracy (%)	65.94 $\pm$ 2.31	66.28 $\pm$ 3.19	63.76 $\pm$ 2.30
Sensitivity(%)	64.61 $\pm$ 5.19	66.02 $\pm$ 4.92	60.41 $\pm$ 4.43
Specificity(%)	67.92 $\pm$ 6.00	66.67 $\pm$ 2.08	68.75 $\pm$ 6.91
Sparse Functional Logistic Regression			
Accuracy (%)	<b>69.79 <math>\pm</math> 2.38</b>	<b>70.30 <math>\pm</math> 1.60</b>	<b>70.30 <math>\pm</math> 1.95</b>
Sensitivity(%)	<b>70.83 <math>\pm</math> 5.31</b>	<b>72.50 <math>\pm</math> 5.78</b>	<b>73.33 <math>\pm</math> 6.97</b>
Specificity(%)	<b>69.10 <math>\pm</math> 3.20</b>	<b>68.82 <math>\pm</math> 2.48</b>	<b>68.27 <math>\pm</math> 3.12</b>
Experimental Results Using Samples With QIDS.C <sub>16</sub> $\leq$ 5 and QIDS.C <sub>16</sub> $\geq$ 11			
	LV1.L	LV1.S+LV1.L	LV1.S+LV1.L+LV2.S
Random Forest			
Accuracy (%)	73.84 $\pm$ 6.66	74.09 $\pm$ 6.56	74.08 $\pm$ 4.74
Sensitivity(%)	75.25 $\pm$ 11.50	75.23 $\pm$ 8.44	73.35 $\pm$ 10.46
Specificity(%)	72.92 $\pm$ 4.89	73.33 $\pm$ 6.32	74.58 $\pm$ 2.72
Sparse Logistic Regression			
Accuracy (%)	72.58 $\pm$ 3.99	73.08 $\pm$ 4.23	74.82 $\pm$ 4.53
Sensitivity(%)	68.37 $\pm$ 7.11	70.87 $\pm$ 7.41	75.19 $\pm$ 9.74
Specificity(%)	75.42 $\pm$ 3.09	74.58 $\pm$ 3.09	74.58 $\pm$ 2.28
Sparse Functional Logistic Regression			
Accuracy (%)	<b>77.81 <math>\pm</math> 4.11</b>	<b>77.82 <math>\pm</math> 4.89</b>	<b>77.07 <math>\pm</math> 4.54</b>
Sensitivity(%)	<b>79.17 <math>\pm</math> 2.08</b>	<b>78.75 <math>\pm</math> 4.01</b>	<b>77.92 <math>\pm</math> 3.78</b>
Specificity(%)	<b>75.81 <math>\pm</math> 10.77</b>	<b>76.46 <math>\pm</math> 9.51</b>	<b>75.83 <math>\pm</math> 10.45</b>

1 and level 2 baseline scalar features are not helpful for improving classification performance. We obtain similar observations when applying Random Forest and sparse logistic regression. Since only using longitudinal data at level 1 leads to satisfactory classification performance, we may conclude that most of the information in the level 1 and level 2 baseline data is captured by the longitudinal data at level 1. The experimental results further demonstrate the importance of mining longitudinal data.

Besides the superior predictive performance, the sparse functional logistic regression is also capable of selecting significant longitudinal features and giving interpretable solutions. In our experiments, most meaningful and interesting longitudinal features including side effect frequency, side effect burden, and QIDS-C current score, are selected as a result of the group sparsity constraint. Moreover, the functional coefficients can be visualized to provide deeper insights for understanding the effects of longitudinal features on predicting remission status over time. In Fig. 3, we show 6 important functional coefficients obtained from the task of differentiating participants with QIDS-SR<sub>16</sub>  $\leq$  5 from those with QIDS-SR<sub>16</sub>  $\geq$  11. From the figure, we can see that the effect of FG-FISGQ = 0 (side effect frequency) decreases from W0 to W6 and then increases over time. For feature FG-GRSEB = 0 (side effect burden),

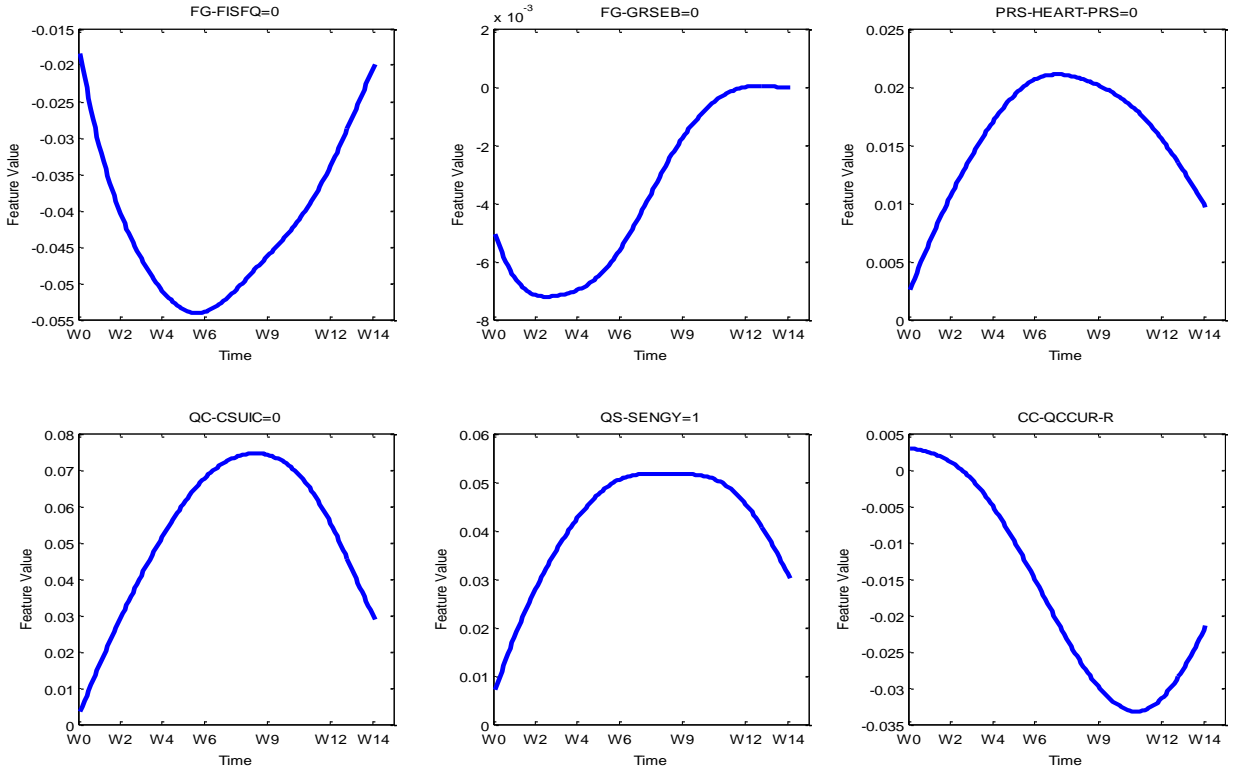


Fig. 3. The functional coefficients of some selected longitudinal features via sparse functional logistic regression. Feature FG-FISQ= 0 relates to side effect frequency; feature FG-GRSEB= 0 relates to side effect burden; feature PRS-HERAT-PRS= 0 relates to clinical measurements of heart; feature QC-CSUIC= 0 relates to QIDS Suicidal ideation; feature QS-SENGY= 0 relates to QIDS Energy/fatigability; and feature CC-QCCUR=  $R$  relates to QIDS-C current score.

we observe that the effects are zero during W12 to W14, which indicates the corresponding feature values during that period make no contributions to the prediction. The sparse and interpretable results are due to the use of Lasso and fused Lasso penalty. From all the plots, we also observe that all the obtained functional coefficients are smooth, which demonstrates the effectiveness of the fused lasso penalty in controlling the smoothness of coefficient functions.

## 5. Conclusions

In this paper, we propose a novel sparse generalized functional linear model for the longitudinal analysis of STAR\*D data. Compared to traditional functional regression models, our model has the advantages of simultaneously achieving high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. We conduct extensive experiments on the STAR\*D data set and the experimental results demonstrate that the proposed sparse functional regression model achieves significantly higher longitudinal prediction power than existing approaches.

Since the proposed models have shown great effectiveness in capturing temporal information, we intend to apply them to investigate the predictive effects of the biomarkers on other longitudinal problems such as the progression of Alzheimer's Disease (AD). Moreover, we plan

to further study the theoretical properties of the proposed models in the future.

## 6. Acknowledgments

This work is support by NIH (R01 LM010730) and NSF (IIS-0953662, MCB-1026710, and CCF-1025177).

## References

1. World Federation for Mental Health, *Archives of Neurology* (2012).
2. A. Haden and B. Campanini, *The World Health Report 2001: Mental health: new understanding, new hope* (World health organization (WHO), 2001).
3. M. Fava, *et al.*, *Psychiatric Clinics of North America* **26**, 457 (2003).
4. H. Wang, *et al.*, *Bioinformatics* **28**, i619 (2012).
5. H. Wang, *et al.*, High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction, in *NIPS*, 2012.
6. J. Zhou, L. Yuan, J. Liu and J. Ye, A multi-task learning formulation for predicting disease progression, in *ACM SIGKDD*, 2011.
7. J. Zhou, J. Liu, V. A. Narayan and J. Ye, *NeuroImage* **78**, 233 (2013).
8. H.-G. Müller, *StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies*.
9. J. Ramsay and B. Silverman, *Functional Data Analysis*, Springer Series in Statistics (Springer, 2005).
10. F. Ferraty, P. Vieu, *Nonparametric functional data analysis: theory and practice* (Springer, 2006).
11. Q. Shen and J. Faraway, *Statistica Sinica* **14**, 1239 (2004).
12. J.-M. Chiou, H.-G. Muller, J.-L. Wang *et al.*, *Statistica Sinica* **14**, 675 (2004).
13. X. Yang, Q. Shen, H. Xu and S. Shoptaw, *Statistics in medicine* **26**, 1552 (2007).
14. T. T. Cai and P. Hall, *The Annals of Statistics* **34**, 2159 (2006).
15. G. M. James, J. Wang and J. Zhu, *The Annals of Statistics* **37**, 2083 (2009).
16. H. Cardot, *et al.*, *Computational statistics & data analysis* **51**, 4832 (2007).
17. H. Zhu and D. D. Cox, *Lecture Notes-Monograph Series*, 173 (2009).
18. J. Gertheiss, A. Maity and A.-M. Staicu, *Stat* (2013).
19. Y. Fan and G. M. James, Functional additive regression, *Under Review*, 2012.
20. J. Friedman, T. Hastie and R. Tibshirani, *arXiv preprint arXiv:1001.0736* (2010).
21. R. Tibshirani, *et al.*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 91 (2004).
22. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)*, 267 (1996).
23. S. Boyd, *et al.*, *Foundations and Trends® in Machine Learning* **3**, 1 (2011).
24. Y. Nesterov, A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ , in *Soviet Mathematics Doklady*, (2)1983.
25. Y. Nesterov, *Mathematical Programming* **103**, 127 (2005).
26. J. A. Rice, *Statistica Sinica* **14**, 631 (2004).
27. G. M. James, T. J. Hastie and C. A. Sugar, *Biometrika* **87**, 587 (2000).
28. F. Yao, *et al.*, *Journal of the American Statistical Association* **100**, 577 (2005).
29. J. Liu, J. Chen and J. Ye, Large-scale sparse logistic regression, in *ACM SIGKDD*, 2009.
30. J. Zhou, J. Liu, V. A. Narayan and J. Ye, Modeling disease progression via fused sparse group lasso, in *ACM SIGKDD*, 2012.
31. J. Liu, *et al.*, An efficient algorithm for a class of fused lasso problems, in *ACM SIGKDD*, 2010.
32. A. Rush, *et al.*, *American Journal of Psychiatry* **163**, 1905 (2006).

## **DEVELOPMENT OF A DATA-MINING ALGORITHM TO IDENTIFY AGES AT REPRODUCTIVE MILESTONES IN ELECTRONIC MEDICAL RECORDS**

JENNIFER MALINOWSKI

*Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall  
Nashville, TN 37232, USA*

*Email: [jennifer.malinowski@vanderbilt.edu](mailto:jennifer.malinowski@vanderbilt.edu)*

ERIC FARBER-EGER

*Center for Human Genetics Research, Vanderbilt University, 1207 17<sup>th</sup> Avenue, Suite 300  
Nashville, TN 37232, USA*

*Email: [eric.h.farber-eger@vanderbilt.edu](mailto:eric.h.farber-eger@vanderbilt.edu)*

DANA C. CRAWFORD

*Department of Molecular Physiology and Biophysics, Center for Human Genetics Research,  
2215 Garland Avenue, 519 Light Hall  
Nashville, TN 37232, USA*

*Email: [crawford@chgr.mc.vanderbilt.edu](mailto:crawford@chgr.mc.vanderbilt.edu)*

Electronic medical records (EMRs) are becoming more widely implemented following directives from the federal government and incentives for supplemental reimbursements for Medicare and Medicaid claims. Replete with rich phenotypic data, EMRs offer a unique opportunity for clinicians and researchers to identify potential research cohorts and perform epidemiologic studies. Notable limitations to the traditional epidemiologic study include cost, time to complete the study, and limited ancestral diversity; EMR-based epidemiologic studies offer an alternative. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study, as part of the Population Architecture using Genomics and Epidemiology (PAGE) I Study, has genotyped more than 15,000 patients of diverse ancestry in BioVU, the Vanderbilt University Medical Center's biorepository linked to the EMR (EAGLE BioVU). We report here the development and performance of data-mining techniques used to identify the age at menarche (AM) and age at menopause (AAM), important milestones in the reproductive lifespan, in women from EAGLE BioVU for genetic association studies. In addition, we demonstrate the ability to discriminate age at naturally-occurring menopause (ANM) from medically-induced menopause. Unusual timing of these events may indicate underlying pathologies and increased risk for some complex diseases and cancer; however, they are not consistently recorded in the EMR. Our algorithm offers a mechanism by which to extract these data for clinical and research goals.



## 1. Introduction

### 1.1 *Women's health and the reproductive lifespan*

Though women comprise more than 50% of the US population[1] and there are notable differences in the incidences and severity of diseases between men and women, from Alzheimer's disease[2] to inflammatory arthritis[3], only in the last few decades has the importance of women's health and physiologic differences between males and females in the research setting come to the forefront of researchers and government agencies[4]. Age at menarche (AM) and age at menopause (AAM) define the boundaries of the reproductive lifespan in women. The timing of these events has also been linked to numerous diseases and complex traits [5]. Fertility is directly impacted by the length of the reproductive lifespan. Earlier AM and later AAM have been associated with heightened risks for breast, ovarian, and endometrial cancers, elevated blood pressure, and increased glucose intolerance, driven by a significant extent by the additional exposure to circulating estrogens over an extended reproductive lifespan [6]. Early AAM has been associated with increased risk for cardiovascular disease [7]. More directly, extremely early or late attainment of these reproductive milestones can indicate underlying pathologies, such as pituitary diseases, hormone imbalances, and nutritional insufficiencies [5].

National surveys have calculated the average AM to be 12.4 years and age at natural menopause (ANM) at 51 years [8]. The genetic contribution to the timing of menarche and natural menopause is estimated to be approximately 0.50, however variants identified through numerous genome-wide association studies (GWAS) account for <10% of the observed variation in either AM or ANM [8]. Cross-sectional and longitudinal studies have shown recent secular trends in the earlier attainment of pubertal milestones (breast development, appearance of pubic hair, menarche) from the 1960s to present and later age at natural menopause [9]. The earlier AM is accelerated in girls of African American and Hispanic ancestry, a bias that remains after adjusting for socioeconomic variables and body mass index (BMI) [10]. The difference observed in the timing of reproductive events across ethnicities highlights the importance of conducting research in diverse populations—a challenging enterprise given the limited diversity in cohorts available for women's health outcomes research.

### 1.2 *Research use of electronic medical records*

Electronic medical/health records (EMRs/EHRs) are becoming more widely used in clinical practice and hospital settings. Motivated in part by the 'meaningful use' requirement for supplemental reimbursements for Medicare and Medicaid claims through the Health Information Technology for Economic and Clinical Health (HITECH) Act, widespread adoption of EMR technology is expected to improve patient outcomes and streamline health care processes and may be helpful in the goal of "personalized medicine" [11-14]. A significant measure of 'meaningful use' is the recording of patient data (e.g., demographic, medication allergy, smoking status, vital signs) as structured data [12]. Additional measurements of 'meaningful use' include the dissemination of clinical quality measurements to states or other governmental oversight agencies.

Immunization and reportable disease statistics are two examples of EMR data that can be leveraged for public health research [15].

The rich phenotypic data existing in EMR systems allows clinicians and researchers to identify potential cohorts, while EMRs that are linked to biobanks extend this framework to genotype-phenotype association studies. Traditional epidemiologic studies are costly and require significant amounts of time to complete; furthermore, these studies may not include sufficient numbers of individuals from diverse ancestries. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study seeks to address these limitations by enabling high-throughput identification and generalization of genotype-phenotype associations in ethnically diverse research populations. Accessing data from EMRs for use in research may prove to be a cost effective alternative to traditional ascertainment and data collection. One challenge to research use of EMR-derived data is the lack of consistency in recording certain types of data in the EMR. Despite the obvious health implications, AM and AAM/ANM are not recorded consistently or in a standardized manner in the EMR. This presents a challenge for researchers and suggests algorithm development is a necessary first step in developing a resource for women's health studies in diverse populations.

### **1.3 BioVU**

BioVU is the Vanderbilt University Medical Center (VUMC) biorepository linked to the EMR system. Beginning in 2007, discarded blood samples from routine clinical testing have the DNA extracted, stored, and linked to a de-identified version of the EMR termed the Synthetic Derivative (SD). As of mid-2012, more than 150,000 samples have been collected for BioVU, including more than 16,000 pediatric samples. Patients are given the opportunity to opt-out of BioVU at any time. Once a sample has been accepted into the system, a unique ID is generated through a one-way hash mechanism and linked to that patient's SD. The SD removes or de-identifies Health Insurance Portability and Accountability Act (HIPAA) information, such as names, geographical locations, and social security numbers, and replaces dates with dates that have been randomly shifted by up to six months. The date shifting is consistent within a single SD record. The SD enables researchers to examine genome-phenome associations and identify cohorts for research.

## **2. Methods**

### **2.1. Population**

As part of the Population Architecture using Genomics and Epidemiology (PAGE) I Study, EAGLE genotyped all non-European descent patients in BioVU (EAGLE BioVU, n=15,863) on the MetaboChip, a custom genotyping array with an emphasis on cardiovascular disease and metabolic traits. This array also includes over 2200 SNPs associated at genome-wide significance to any trait published in the NHGRI GWAS catalog as of August 2009, with additional proxy SNPs chosen based on linkage disequilibrium (LD) in both CEU and YRI HapMap II datasets [16]. Overall, 11,521 African Americans, 1,714 Hispanics, 1,122 Asians and others were

genotyped on the MetaboChip by EAGLE. For the AM study, all females age >6 in EAGLE BioVU as of January 31, 2013 were eligible for inclusion. For the AAM study, all females >18 years were eligible for inclusion; for the ANM study, only women ages  $\geq 41$  were eligible for inclusion. All patients were of diverse ethnicity.

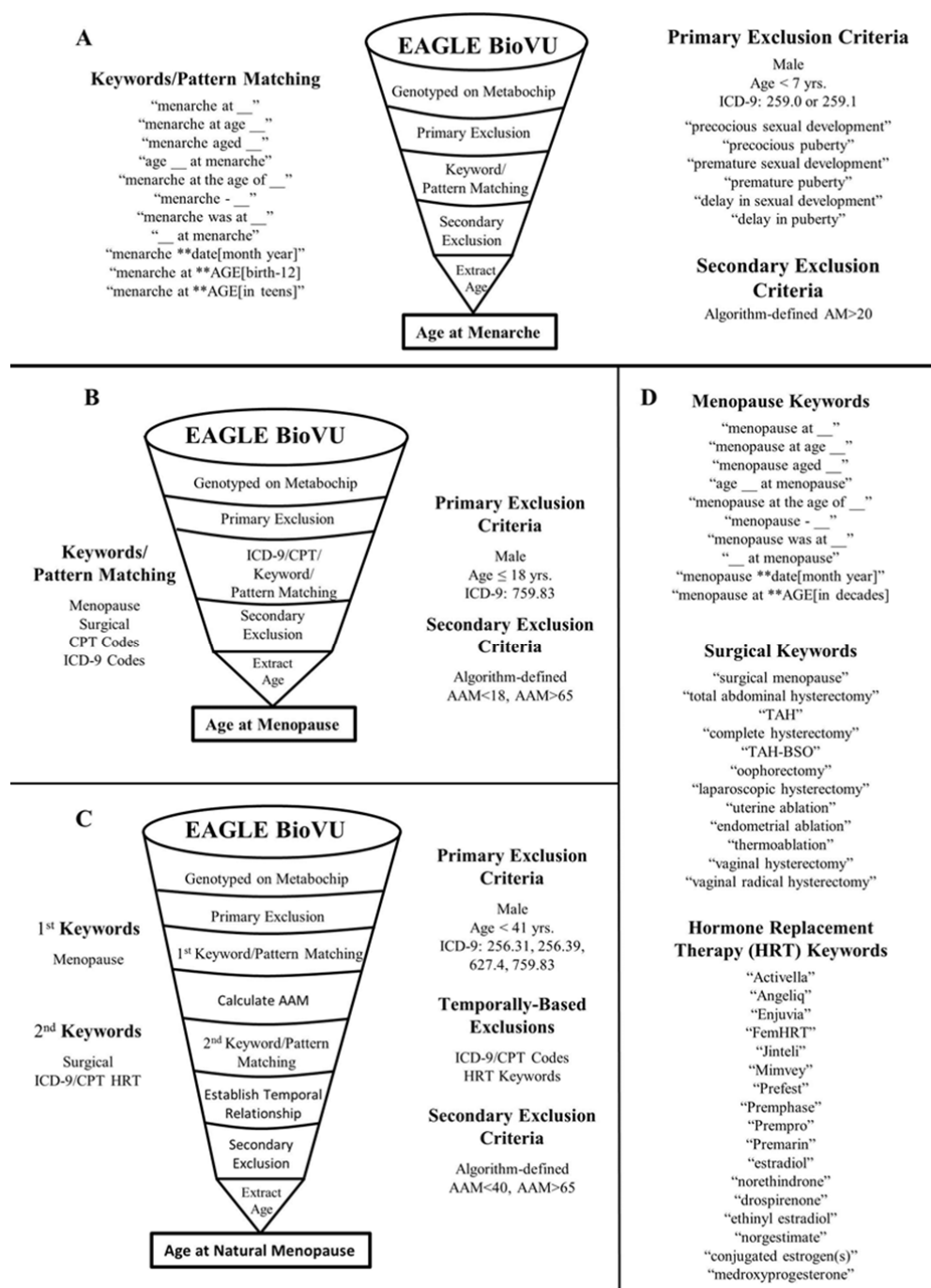
## 2.2. Algorithm development

We developed a flow chart to visualize the inclusion/exclusion processes for the algorithms (Fig. 1A (AM) and Fig. 1B/C (AAM/ANM)). AM and age at menopause or age at natural menopause (AAM/ANM) are not consistently recorded in the EMR system at VUMC; individuals may enter BioVU through numerous outpatient clinics. The lack of a pre-specified field for AM and AAM/ANM in the EMR necessitated a combination of free text data mining using regular expressions/pattern matching, billing (ICD-9) codes, and procedure (CPT) codes to identify AM and AAM/ANM in the subsequently generated SD. All analysis for this study was performed using the SD.

### 2.2.1 Age at menarche (AM)

Primary exclusion criteria for AM phenotype consisted of four components: age <7 years, male sex, ICD-9 codes for delayed puberty/sexual development (259.0) and precocious puberty/sexual development (259.1), and keywords (Figure 1A). Inclusion of any of the preceding criteria in the SD resulted in exclusion for the AM study. As part of the de-identification data scrubbing to convert a patient's EMR to the SD, ages and dates may be masked and listed as "birth-12" or "in teens." Dates and ages which are not masked were date shifted by up to six months forward or backward from the actual date.

To identify a listed AM for an individual, we utilized pattern matching to seek instances with menarche keyword phrases (Figure 1A). Numbers and dates were allowed to be included as numerals only. Instances where the AM was listed as a date used the subject's birthdate to calculate the age (in years) at menarche. In cases of ties, where more than one AM was identified and recorded an equal number of times in the SD, the AM was determined to be the one listed first in the SD. If the algorithm identified multiple versions of the AM (an exact age, an age calculated from a date, or a de-identified age), a hierarchy was used to determine the AM for the output, where an exact age or date was prioritized over de-identified age ranges. Instances where multiple different ages were listed in the SD as AM defaulted to the age listed most frequently. We considered situations where the algorithm identified an exact AAM and a de-identified AAM range containing the exact AAM to be the same for purpose of calculating sensitivity, specificity, and positive predictive value (PPV), but different for the purpose of calculating accuracy. The resulting output file contained the subject's unique research id (RUID), date of birth, and either an algorithm-generated AM or null value.



**Figure 1.** Flow chart for (A) age at menarche (AM), (B) age at menopause (AAM), (C) age at natural menopause (ANM), and (D) keywords for AAM and ANM algorithms.

### 2.2.2 Age at menopause (AAM)

For an algorithm to identify all post-menopausal women and their age at menopause (AAM), we initially excluded all males, set a minimum age of 18 years, and excluded patients with a Fragile X diagnosis (ICD-9 759.83) (Figure 1B). Pattern matching was utilized to find keyword phrases similar to those used in the menarche algorithm, substituting “menopause” for “menarche” (Figure 1D). Furthermore, we included keywords pertaining to surgical procedures that induce cessation of menses/menopause (Figure 1D). We excluded instances where the word “possible” immediately preceded a keyword. For instances where the SD had scrubbed the exact age, decade-specific results (e.g. “in 30s”, “in 50s”) were captured by our algorithm. CPT and ICD-9 (Table 1) codes were used to identify women with surgical menopause or menses-ceasing procedures.

Table 1. CPT and ICD-9 codes used for menopause (AAM/ANM) algorithm development.

CPT codes			ICD-9 codes		
58150	58285	58548	65.5	68.3	68.69
58152	58290	58550	65.51	68.31	68.7
58180	58291	58552	65.52	68.39	68.71
58200	58292	58553	65.53	68.4	68.79
58260	58293	58554	65.64	68.41	68.9
58262	58294	58563	65.6	68.49	
58263	58353	58570	65.61	68.5	
58267	58541	58571	65.62	68.51	
58270	58542	58572	65.63	68.59	
58275	58543	58573	65.64	68.6	
58280	58544		68.23	68.61	

After SD review of initial algorithms and subject matter knowledge, we implemented secondary exclusion criteria based on the algorithm-identified AAM and excluded subjects with a calculated  $AAM < 18$  or  $AAM > 65$  (Figure 1B). A hierarchy was used to determine the AAM for the output, with an exact age or date identified by keyword or pattern matching and ICD-9/CPT codes prioritized over de-identified age ranges. In rare instances where the algorithm identified more than one AAM for a subject, the age recorded most frequently was determined to be the AAM for that patient. In cases of ties, where more than one AAM was identified and recorded an equal number of times in the SD, the AAM was determined to be the one listed first in the SD. We considered situations where the algorithm identified an exact AAM and a de-identified AAM range containing the exact AAM to be the same for purpose of calculating sensitivity, specificity, and PPV, but different for the purpose of calculating accuracy. The resulting output file contained the subject’s unique research id (RUID), date of birth, race/ethnicity, either an algorithm-

generated AAM or null value, the method by which the AAM was calculated (e.g., from ICD-9 code, keyword), and the date in the SD corresponding to the AAM identification.

### 2.2.3 Age at natural menopause (ANM)

To discriminate age at natural menopause (ANM) from all instances of menopause (AAM), we extended the AAM algorithm to exclude women aged <41 years, men, and subjects with ICD-9 codes signifying premature ovarian failure/premature menopause (256.31), artificially induced menopause (627.4), ovarian failure (256.39), and Fragile X syndrome (759.83) (Figure 1C). We used pattern matching with the menopause keywords to identify an age at menopause (Figure 1D). We did not use ICD-9 codes, CPT codes, or keywords associated with procedures that induce menopause to identify subjects for the ANM cohort.

Medication delivery and prescriptions are captured by the EMR at VUMC and are included in the SD. To ascertain the temporal relationship between AAM and menopause-inducing/menses-ceasing surgery or hormone replacement therapy (HRT) use, we first calculated the AAM with the alternate algorithm (Figure 1C). Surgery-inducing menopause, determined through CPT and/or ICD-9 codes or keywords, and HRT were not exclusion criteria unless the first instance of surgery or HRT occurred prior to the extended algorithm-identified AAM. Keyword pattern matching was performed using surgical keywords (Figure 1D). We used a combination of brand-name and generic names for HRT identification (Figure 1D). If AAM was identified and no keywords or CPT/ICD-9 codes were found to indicate artificially induced menopause, the subject was deemed to have undergone natural menopause. If surgery or HRT occurred after the algorithm-determined ANM, the subject was also considered to have undergone natural menopause. If the subject had either surgery or used HRT prior to menopause, they were excluded from the cohort and the resulting output was a null value.

We implemented secondary exclusion criteria (Figure 1C) based on the algorithm-identified age at menopause and excluded subjects with a calculated ANM<18 or ANM>65 based on subject matter knowledge and review of early versions of our algorithms. A hierarchy was used to determine the ANM for the output. If the algorithm determined more than one ANM for a subject, we used the same procedure as described above to determine the final ANM generated by our query. We again considered situations where the algorithm identified an exact ANM and a de-identified ANM range containing the exact ANM to be the same for purpose of calculating sensitivity, specificity, and PPV, but different for the purpose of calculating accuracy. The resulting output file contained the subject's unique research id (RUID), date of birth, race/ethnicity, either an algorithm-generated ANM or null value, the method by which the ANM was calculated (e.g., from exact date, de-identified age), and the date in the SD corresponding to the ANM identification.

### 2.3. Manual review

To determine the sensitivity, specificity, PPV, and accuracy of the AM, AAM, and ANM algorithms, extensive manual chart review was performed by a single individual for consistency. Each algorithm output contained three types of values: exact ages, de-identified ages, and null values. For each algorithm, a random number generator was used to randomize RUIDs within each of the three types of output and the subjects were then sorted in ascending value by the random number. The first 50 subjects in the exact age and de-identified age categories and the first 100 subjects with a null value had their SD reviewed manually to determine the AM, AAM, or ANM. Sensitivity, specificity, PPV and accuracy were calculated by comparing the automated algorithm result to the manual review result for each subject.

## 3. Results

### 3.1 Population characteristics

A total of 10,051 females were genotyped on the MetaboChip in BioVU by EAGLE for various studies. We identified an age for menarche (exact or de-identified) in 1,618 individuals. For the AAM algorithm, we identified an AAM (exact age or de-identified decade) for 1281 individuals. We identified 83 individuals with an ANM (exact or de-identified decade) (Table 2). The algorithm-extracted mean AM in our population was 12.7 (+/- 2.1 ) yrs. The mean AAM in our population was 44.6 (+/- 9.8) yrs. and the mean ANM was 49.7 (+/- 5.6) yrs. (Table 2). Approximately half of the algorithm extracted AM (54.7%) and ANM (47.0%) were exact ages, while the majority of AAM (92.5%) were exact ages (Table 2).

Table 2. Population characteristics for women with algorithm-identified age at menarche (AM), age at menopause (AAM), and age at natural menopause (ANM) from EAGLE BioVU. Abbreviations: standard deviation (sd), years (yrs).

A	M	AAM	ANM
N, total	1618	1281	83
exact age (n)	885	1185	39
de-identified age (n)	733	96	44
Age at event, mean +/- sd (yrs)	12.7 (2.1)	44.6 (9.8)	49.7 (5.6)
Age range at event (yrs)	8-20	18-65	40-65
Race/ethnicity (n)			
African American	1232	1112	62
Hispanic	120	45	4
Asian	115	66	11
Other	151	58	6

### 3.2 AM algorithm performance

We manually reviewed 200 SD entries for the AM algorithm to determine sensitivity, specificity, PPV, and accuracy. Of the 100 subjects with an algorithm-specified AM, 94 were confirmed by manual review. For the 100 subjects without an AM captured by the algorithm, 99 were not found to have an identifiable AM upon manual review. The AM algorithm had a sensitivity and specificity of 99.0% and 94.3%, respectively, and a PPV of 94.0% (Table 3). We calculated the accuracy of the algorithm by comparing the results for the 94 subjects with both manually identified and algorithm identified AMs, requiring identical results for concordance. Of these 94 subjects, we found 87 where the AM matched in both manual and algorithm identification for an accuracy of 92.6% (Table 4). We observed instances where the algorithm calculated an exact AM (e.g., 8) and manual review found a de-identified AM (e.g., birth-12), or vice-versa. If we allow these to be concordant, accuracy increases to 94.7%.

Table 3. Performance of the age at menarche (AM), age at menopause (AAM), and age at natural menopause (ANM) algorithms in women from EAGLE BioVU.

Abbreviations: positive predictive value (PPV).

Sensitivity	ity	Specificity	Accuracy	PPV
AM (n=200)	99.0%	94.3%	92.6%	94.0%
AAM (n=200)	94.4%	85.6%	52.4%	84.0%
ANM (n=183)	89.8%	75.8%	75.5%	63.9%

### 3.3 AAM algorithm performance

For the AAM algorithm, we manually reviewed 200 SD entries to determine sensitivity, specificity, PPV, and accuracy. Of the 100 subjects with an algorithm-identified AAM, we identified 82 with AAM via manual review. Only five of the 100 subjects without an algorithm-identified AAM were found to have an identifiable AAM with manual review. Overall, our algorithm was found to have 94.4% sensitivity, 85.6% specificity, and a PPV of 84.0% (Table 3). We also calculated the accuracy of our AAM algorithm by comparing the algorithm-obtained AAM to the manual review-obtained AAM. We observed a 52.4% exact concordance within our 82 subjects with AAMs calculated from both manual review and the algorithm. If we allowed a de-identified age range encompassing an exact age to be considered concordant with the exact age obtained from the other method, our accuracy improved to 61.9%.

### 3.4 ANM algorithm performance

The ANM algorithm identified 83 individuals with an ANM; therefore, we manually reviewed 183 SD entries to determine the specificity, sensitivity, PPV, and accuracy of our ANM algorithm. Of the 100 individuals with no algorithm-identified ANM, manual review of the SD found 6 instances with an identifiable ANM (Table 3). Of the 83 individuals with an algorithm-specified ANM,



manual review confirmed 53. Overall, the sensitivity and specificity of the ANM algorithm were 89.8% and 75.8%, respectively, and the PPV was 63.9%. Of the 53 subjects with both algorithm- and manually-identified ANM, 40 were an exact match, yielding an accuracy of 75.5%. We again observed instances where the algorithm yielded an exact age, but manual review of the SD obtained only a de-identified ANM range that encompassed the exact age, and vice-versa; if we considered these as concordant, our accuracy increased to 81.1%.

#### 4. Conclusion

Menarche and menopause are the bookends of the reproductive lifespan in women. The timing of these events may increase risk for various complex disorders and cancers, such as osteoporosis and breast cancer [5]. Precocious or delayed menarche may signal the occurrence of hormonal imbalance, inadequate nutrition or caloric intake, or pituitary diseases [5]. The timing of menopause directly affects reproductive capabilities. In addition, premature menopause may result from hormonal imbalances, genetic disorders such as Fragile X Syndrome, metabolic disorders, or autoimmune diseases such as thyroid disease or rheumatoid arthritis [17]. Though the timing of menarche and menopause may increase risk for disease or indicate underlying pathologies, this information is not consistently included in electronic health records, leading to missed opportunities to inform clinical care and represents a challenge to clinicians and researchers alike.

Data-mining EMRs has been used to identify cohorts for research studies [18-21], determine smoking status [22], and predict disease, such as sepsis [23]. Our development of algorithms to extract these important data is notable for the emphasis on diverse populations and attention to women's health, both historically underrepresented in health outcomes research. The menarche (AM) and menopause (AAM) algorithms have PPV>80% and high specificity and sensitivity, though accuracy of the AAM algorithm was just over 50%. The age at natural menopause (ANM) algorithm had moderately high (>75%) sensitivity and specificity but the lowest PPV, at 63.9%. However, the accuracy of the ANM algorithm bested that of the AAM (75.5% vs. 52.4%, respectively). In addition, the algorithm-extracted ages at menarche, menopause, and natural menopause are consistent with published research, validating our methodology.

Several factors may have reduced the performance of our menopause algorithms. We observed many instances where the ages calculated by the algorithm and by manual review differed by one year. This may have been the result of the date-shifting done within each individual's SD for de-identification purposes. If the method for calculating the age differed between the methods, it is possible this could result in the observed one-year difference. When we allowed a +/- 1 year difference in the algorithm and manual identified AAM and ANM, the accuracy of our algorithms improved to 70.2% and 90.6%, respectively. The timing of menopause is challenging to identify, as the menstrual cycle becomes more erratic as a woman moves through perimenopause into menopause. Months may lapse between cycles; hormone levels may change substantially. In addition, the normal menopausal age range is quite large, taking place between the ages of 40 and 60. These factors challenge the accurate dating of the onset of menopause.

Furthermore, an algorithm designed to identify the age at menopause may not accurately reconcile multiple mentions in an EMR of menopause. Discerning between natural menopause and medically/surgically induced menopause is an additional challenge. Our extensive list of time-dependent exclusions for HRT and surgical procedures was not exhaustive and may have led to the algorithm identifying an ANM where manual review identified HRT and/or a procedure artificially inducing menopause. Correctly identifying the temporal relationship between attainment of natural menopause and surgical procedures that result in menopause may perform inconsistently in the absence of these data in structured fields in an EMR. Addressing some of these issues by including structured fields for age at menarche, age at menopause, and type of menopause (natural/medical), and standardizing the reporting of these data could greatly improve the performance of our algorithms.

We have demonstrated the performance of algorithms designed to extract the age at menarche and age at menopause from the Synthetic Derivative, a de-identified version of the electronic medical record at Vanderbilt University Medical Center. Furthermore, we have developed an algorithm to discriminate naturally occurring menopause from artificially-induced menopause. Our method combining text-mining for regular expressions and pattern matching, and structured data derived from the EMR to obtain the age at menarche and the age at menopause is likely to be easily transferable to other institutions, given the simplicity of the approach. Overall, these algorithms provide an opportunity for researchers and clinicians to obtain these valuable, though inconsistently reported data.

## Acknowledgments

This work was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

## References

1. Bureau of Census, C2010BR-03 (2011).
2. K. Irvine, K.R. Laws, T.M. Gale, and T.K. Kondel, *J Clin Exp Neuropsychol*. doi:10.1080/13803395.2012.712676 (2012).
3. C. Barnabe, L. Bessette, C. Flanagan, *et al.*, *J Rheumatol* **39**, 1221 (2012).
4. C. Taylor, *J Womens Health* **3**, 143 (1994).
5. P. Hartge, *Nat Genet*. **41**, 637 (2009).
6. V. Dvornyk and U.H. Waqar, *Hum Reprod Update* doi:10.1093/humupd/dmr050 (2012).

7. W.A. Rocca, B. R. Grossardt, V.M. Miller, *et al.*, *Menopause* **19**, 272 (2012).
8. C. He and J.M. Murabito, *Mol Cell Endocrinol*, doi:10.1016/j.mce.2012.05.003 (2013).
9. P. Kaplowitz, *Curr Opin Obstet Gynecol* **18**, 487 (2006).
10. T. Wu, P. Mendola, and G.M. Buck, *Pediatrics* **110**, 752 (2002).
11. D. Blumenthal and M. Tavenner, *N Engl J Med*. **363**, 501 (2010).
12. D. Blumenthal, *N Engl J Med*. **362**, 382 (2010).
13. D. Blumenthal, *N Engl J Med*. **365**, 2426 (2011).
14. A.K. Jha, C.M. DesRoches, E.G. Campbell, *et al.*, *N Engl J Med*. **360**, 1628 (2009).
15. R. Kukafka, J.S. Ancker, C. Chan, *et al.*, *J Biomed Inform.* **40**, 398 (2007).
16. B.F. Voight, H.M. Kang, J. Ding, *et al.*, *PLoS Genet.* doi:10.1371/journal.pgen.1002793 (2012).
17. T.C. Okeke, U.B. Anyaehie, and C.C. Ezenyeaku, *Ann Med Health Sci Res.* **3**, 90 (2013).
18. K.M. Newton, P.L. Peissig, A.N. Kho, *et al.*, *J. Am. Med. Inform. Assoc.* doi: 10.1136/amiajnl-2012-000896 (2012).
19. M. J. Stratton-Loeffler, J. C. Lo, R.L. Hui, *et al.*, *J. Manag. Care Pharm.*, **18**, 497 (2012).
20. J. Warren, J. Kennelly, D. Warren, *et al.*, *Stud. Health Technol. Inform.* **178**, 228 (2012).
21. J. S. Brownstein, S.N. Murphy, A.B. Goldfine, *et al.*, *Diabetes Care* **33**, 526 (2010).
22. L.K. Wiley, A. Shah, H. Xu, W.S. Bush. *J. Am. Med. Inform. Assoc.* doi: 10.1136/amiajnl-2012-001557 (2013).
23. S. Mani, A. Ozdas, C. Aliferis, *et al.*, *J. Am. Med. Inform. Assoc.* 10.1136/amiajnl-2013-001854 (2013).

# AN EFFICIENT ALGORITHM TO INTEGRATE NETWORK AND ATTRIBUTE DATA FOR GENE FUNCTION PREDICTION

SHANKAR VEMBU

*Donnelly Center for Cellular and Biomolecular Research,  
University of Toronto, Toronto, ON, Canada  
E-mail: shankar.vembu@utoronto.ca*

QUAID MORRIS

*Donnelly Center for Cellular and Biomolecular Research,  
Banting and Best Department of Medical Research,  
Department of Molecular Genetics,  
Edward S. Rogers Sr. Department of Electrical and Computer Engineering,  
Department of Computer Science,  
University of Toronto, Toronto, ON, Canada  
E-mail: quaid.morris@utoronto.ca*

Label propagation methods are extremely well-suited for a variety of biomedical prediction tasks based on network data. However, these algorithms cannot be used to integrate feature-based data sources with networks. We propose an efficient learning algorithm to integrate these two types of heterogeneous data sources to perform binary prediction tasks on node features (e.g., gene prioritization, disease gene prediction). Our method, *LMGraph*, consists of two steps. In the first step, we extract a small set of “network features” from the nodes of networks that represent connectivity with labeled nodes in the prediction tasks. In the second step, we apply a simple weighting scheme in conjunction with linear classifiers to combine these network features with other feature data. This two-step procedure allows us to (i) learn highly scalable and computationally efficient linear classifiers, (ii) and seamlessly combine feature-based data sources with networks. Our method is much faster than label propagation which is already known to be computationally efficient on large-scale prediction problems. Experiments on multiple functional interaction networks from three species (mouse, fly, *C.elegans*) with tens of thousands of nodes and hundreds of binary prediction tasks demonstrate the efficacy of our method.

*Keywords:* gene function prediction; graph-based learning; label propagation; ensemble learning.

## 1. Introduction

Network-based prediction algorithms are widely used in biomedical prediction tasks.<sup>1–3</sup> These prediction tasks often share a number of properties – a small number of labeled nodes (e.g., genes or patients), a large number of unlabeled nodes, and sparse connectivity among the nodes – that make label propagation algorithms particularly well-suited to the domain. In particular, algorithms proposed by Zhu *et al.*<sup>4</sup> and Zhou *et al.*<sup>5</sup> have only a single free parameter and permit very efficient implementations, and can therefore be applied to very large prediction problems with very little labeled data. Despite their simplicity, these algorithms perform surprisingly well on prediction benchmarks, see for example, Refs. 6 and 7.

However, “feature-based” data are often available for individual nodes in the networks – for example, gene features could include the presence of particular protein domains, sequence conservation levels, associations with disease, phenotypes associated with its deletion mutants.

Representing this feature data by a network-based similarity measure requires grouping features and measuring similarity among feature profiles. This approach loses information about individual feature values, as well as generating a dense similarity network that slows down label propagation algorithms. In this paper, we describe a new algorithm related to label propagation which retains many of its advantages while also allowing heterogeneous feature and network data to be integrated into a common framework.

Although the algorithm we describe can be applied to any domain, for concreteness and because of the existence of comprehensive benchmark data, we consider the problem of predicting gene function from heterogeneous genomic and proteomic data sources.<sup>8–11</sup> Here, one is given a set of genes (query) with a given annotation, and asked to find genes similar to the query. The classic example of this type of problem is predicting Gene Ontology (GO) annotations but could also involve predicting disease associated genes. Functional interaction networks are a widely used representation to capture information about shared gene function present in genomic and proteomic data sources.<sup>8,11</sup> A popular approach to solving this problem is to combine these networks into a composite network<sup>6,12</sup> and, along with a set of labels that describe the gene function, use them as inputs to a graph-based learning algorithm such as label propagation.<sup>4,5</sup> The main advantage of these methods is that they are computationally efficient. Both label propagation and the method of Tsuda *et al.*<sup>12</sup> admit a solution of the form  $P^{-1}q$ , where  $P$  is a sparse matrix, and can be computed by solving a sparse linear system whose time complexity is almost linear in the number of non-zero entries in  $P$ .<sup>13</sup>

Despite being computationally efficient, the algorithms proposed by Tsuda *et al.*<sup>12</sup> and Mostafavi *et al.*<sup>6</sup> cannot be used to integrate feature-based data sources (attributes) with networks. A natural solution to this problem is to construct a similarity graph<sup>a</sup> (preferably sparse) from the feature-based data. This can be done by first computing a kernel matrix from the features, for example, using the dot-product kernel or the radial basis function (RBF) kernel, and then by using an appropriate method to sparsify the dense kernel matrix. However, as mentioned above, the main drawback of this approach is the potential loss of information during the graph construction step and the inability to produce interpretable models. By interpretable models, we mean linear prediction models learned from feature-based data sources that allow us to assess the importance of the learned weights/parameters.

Another solution is to use multiple kernel learning (MKL).<sup>14</sup> Given a set of kernels  $\{K_d\}$ , the goal of MKL is to learn a (linear) combination of kernels,  $K = \sum_d \mu_d K_d$  (where  $\mu_d \geq 0$  are the weights assigned to the individual kernels), along with the classifier parameters. Although there has been a lot of progress in designing efficient optimization methods for MKL (see, for example, Refs. 15 and 16, and references therein), these methods are not efficient to solve the specific problem of learning from multiple graphs for several reasons. In order to use MKL on graphs, we have to first compute a kernel on graphs.<sup>17</sup> Unfortunately, the resulting kernel matrix is dense and storing a pre-computed kernel matrix is infeasible for graphs with tens of thousands of nodes. Also, it is not possible to compute graph kernels “on-the-fly” unlike, for example, an RBF kernel, thereby forcing us to store the entire kernel matrix in memory. Furthermore, training a kernelized classifier (for example, non-linear SVMs) is computationally

<sup>a</sup>We use the terms graph and network interchangeably.

more expensive than training a linear classifier, and has the drawback of not being able to produce interpretable models. Although several advances have been made in machine learning to scale linear classifiers (see, for example, Ref. 18), large-scale learning of kernelized (non-linear) classifiers still remains a difficult problem to solve.

We propose a computationally efficient two-step procedure to integrate multiple functional interaction networks and feature-based data sources for gene function prediction. First, we extract a small set of discriminative features from the network nodes. Then, we apply a simple weighting scheme in conjunction with linear classifiers to combine these features. When compared to the methods proposed by Tsuda *et al.*<sup>12</sup> and Mostafavi *et al.*,<sup>6</sup> our method has the advantage of being able to combine networks with feature-based data sources. Furthermore, our method allows us to learn highly scalable and efficient linear classifiers for gene function prediction from tens of thousands of nodes and hundreds of GO biological process categories. Using our method, we were able to train classifiers much faster than label propagation which is already known to be computationally efficient on large-scale prediction problems.

### 1.1. Preliminaries

Given  $k$  undirected graphs,  $G_d = (V, E_d)$ ,  $d \in \{1, \dots, k\}$ , each of them having  $n$  nodes, and a set  $V_\ell \subset V$  of labeled nodes, the goal is to learn a binary classifier  $f : V \rightarrow \{0, 1\}$  to predict node labels by using (edge) information from all the graphs. For single graphs, the standard approach to learn such a classifier is to propagate labels in the graph.<sup>4,5</sup> Let  $W = (w_{ij})_{i,j=1,\dots,n}$  denote the weighted adjacency matrix of the graph,  $D$  denote the diagonal degree matrix whose entries are  $d_{ii} = \sum_j w_{ij}$ ,  $\forall i \in \{1, \dots, n\}$ . Let  $L$  denote the unnormalized graph Laplacian defined as  $L = D - W$ . Label propagation is reduced to the following optimization problem:

$$\begin{aligned} \hat{f} &= \operatorname{argmin}_{f \in \mathbb{R}^n} \sum_{i \in V_\ell} (f_i - y_i)^2 + \lambda \sum_{i,j \in E} w_{ij} (f_i - f_j)^2 \\ &= \operatorname{argmin}_{f \in \mathbb{R}^n} \sum_{i \in V_\ell} (f_i - y_i)^2 + \lambda f^\top L f, \end{aligned} \quad (1)$$

where  $y$  is the label vector and  $\lambda > 0$  is a regularization parameter. The estimate  $\hat{f}$  can be used to score/rank the nodes where higher scores imply higher confidence in the classifier to assign a positive label to the nodes. Label propagation is a transductive learning algorithm where unlabeled examples are used for training. Since gene function prediction is a highly unbalanced classification problem, we redefine the labels to take one of three values from  $\{-1, +1, u\}$  and label all the unlabeled nodes with  $u = (n^+ - n^-)/n$ , where  $n^+$  and  $n^-$  are the number of positive and negative labels respectively.<sup>6</sup> The solution to this problem can be computed by solving the sparse system of equations:  $(L + \lambda \mathbf{I})\hat{f} = y$ , where  $\mathbf{I}$  denotes the identity matrix.

To combine multiple graphs, we construct a composite graph with adjacency matrix  $W = \sum_{d=1}^k \mu_d W_d$ , where  $\mu_d \geq 0$  are the weights assigned to the individual graphs. If these weights are known, then we can compute the corresponding composite Laplacian and plug it into the optimization problem (1). Tsuda *et al.*<sup>12</sup> showed that the weights  $\{\mu_d\}$  can be

computed by solving for

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} y^{\top} \left( \mathbf{I} + \sum_{d=1}^k \mu_d L_d \right)^{-1} y, \quad \text{s.t.} \quad \sum_d \mu_d \leq \lambda.$$

Mostafavi *et al.*<sup>6</sup> proposed another algorithm to compute the network weights and showed that it performed better than the method of Tsuda *et al.*<sup>12</sup> In this algorithm, network weights are estimated by solving the following constrained linear regression problem:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \|T - \sum_d \mu_d W_d\|_2^2, \quad \text{s.t.} \quad \mu_d \geq 0, \quad \forall d \in \{1, \dots, k\}, \quad (2)$$

where  $T$  is the target matrix whose entries are  $t_{ij} = (n^+/n)^2$  if genes  $i$  and  $j$  are both positive and  $-n^+n^-/n^2$  if genes  $i$  and  $j$  have opposite labels.

## 2. Methods

### 2.1. *Extracting features from graph nodes*

We first describe a method to extract discriminative features from graphs, where by discriminative we mean that the feature extraction method takes label information into account. Our feature extraction method is based on the 3Prop algorithm proposed by Mostafavi *et al.*<sup>19</sup> that labels the nodes of graphs using only three degrees of propagation. Let  $P = D^{-1}W$  denote the transition probability matrix of the graph. The entries  $p_{ij}$  of  $P$  are the probability that a random walk of length one starting from node  $i$  ends at node  $j$ . The  $r$ -step probability matrix  $P^r$  can be similarly interpreted as the random walk probabilities of length  $r$ . Let  $y$  denote the 0/1 vector of labels used for training. Now, if we compute the matrix-vector product  $x^{(r)} = P^r y$ , then the  $i$ -th element  $x_i^{(r)}$  can be interpreted as the probability that a random walk of length  $r$  from node  $i$  ends at a positively labeled node. Mostafavi *et al.*<sup>19</sup> argued and demonstrated empirically that random walks of length at most three suffice to propagate labels in biological networks. We use these probabilities as features for the nodes in the graph, i.e., for every node  $i \in V$ , we form the three-dimensional 3Prop feature vector  $[x_i^{(1)}, x_i^{(2)}, x_i^{(3)}]$ . Note that the probability matrix  $P$  is asymmetric. A symmetric version can be computed by setting  $P = D^{-1/2} W D^{1/2}$ . Algorithm 2.1 describes the feature extraction method. It is important to note that this feature extraction method is computationally highly efficient. As shown in Step 2 in Algorithm 2.1, it is not necessary to explicitly compute  $P^{(r)}$  using dense matrix-matrix products; instead, it can be computed efficiently in a recursive manner using only sparse matrix-vector products. Given these features, we learn a binary linear classifier  $h : \mathcal{X} \rightarrow \{0, 1\}$ , where  $\mathcal{X}$  denotes the feature space, parameterized by a weight vector  $w \in \mathbb{R}^3$  and make predictions as  $h(x) = w^{\top} x$ . Note that  $h(\cdot)$  is essentially a scoring function that can be used to score/rank nodes according to how confident the classifier is for the nodes to have a positive label.

### 2.2. *Combining multiple graphs*

We extract 3Prop features from the nodes of all the  $k$  graphs using the feature extraction method described in Algorithm 2.1. The next step is to learn a classifier by combining all

**Algorithm 2.1** 3Prop - Feature extraction from graph nodes

**Input:** Graph  $G = (V, E)$  ( $|V| = n$ ) with adjacency matrix  $W$  and degree matrix  $D$ , label vector  $y \in \{0, 1\}^n$

**Output:** Feature matrix  $X \in \mathbb{R}^{n \times 3}$

- 1: Compute  $P = D^{-1}W$  (asymmetric) or  $P = D^{-1/2}WD^{-1/2}$  (symmetric)
- 2: Compute  $x^{(1)} = Py$ ,  $x^{(2)} = Px^{(1)}$ ,  $x^{(3)} = Px^{(2)}$
- 3: **return**  $X = [x^{(1)}|x^{(2)}|x^{(3)}]$

these feature sets. A principled solution to this problem is to use multiple kernel learning (MKL). It is important to note that we do not have to design a kernel given the features extracted from the graph nodes. In other words, MKL can be applied with a *linear* kernel. This greatly reduces the computational complexity of learning classifiers in the MKL framework and allows us to scale our method to graphs with thousands of nodes. Although several advances have been made to solve the MKL problem, it has been found that a uniform weighting of kernels is a hard baseline to outperform.<sup>20,21</sup> Cortes *et al.*<sup>21</sup> proposed a simple yet computationally efficient algorithm that was shown to perform better than uniform weighting and also traditional MKL methods.<sup>14</sup> The algorithm consists of two steps: first, independent classifiers are trained using each of the given kernels; then, the weights of these classifiers are determined using an appropriate weighting scheme. We use a similar approach in our algorithm and describe the two steps below.

We use regularized least-squares regression (RLSR) since the loss function minimized by label propagation is squared loss noting, however, that any loss function such as the hinge loss (SVM), the logistic loss (logistic regression) or the ranking loss (RankSVM<sup>22</sup>) can be used. We solve the following set of (independent) optimization problems to estimate the parameters of the classifiers, one for each of the  $k$  graphs:

$$\hat{w}_d = \underset{w}{\operatorname{argmin}} \sum_i (y_i - w^\top x_{id})^2 + \lambda \|w\|_2^2, \quad \forall d \in \{1, \dots, k\},$$

where we have used  $x_{id} \in \mathbb{R}^3$  to denote the 3Prop feature vector for the  $i$ -th node in graph  $d$ . The solution to this problem can be computed in closed form as  $\hat{w}_d = (X_d^\top X_d + \lambda \mathbf{I})^{-1} X_d^\top y$ , where  $X_d \in \mathbb{R}^{n \times 3}$  is the feature matrix corresponding to the graph  $G_d$  (cf. Algorithm 2.1). Note that computing this solution requires the inversion of a small  $3 \times 3$  matrix. In practice, we found this method to be much faster than label propagation (1).

We then evaluate the performance of these classifiers on a separate validation set and use this performance measure directly as the network weights  $\{\mu_d\}$ . Specifically, we use the area under the ROC curve (AUC) as the performance measure, which is defined as follows:

$$\text{AUC}(y, \hat{y}) \propto \sum_{(i,j): y_i > y_j} \left( [\hat{y}_i > \hat{y}_j] + \frac{1}{2} [\hat{y}_i = \hat{y}_j] \right), \quad (3)$$

where  $y$  and  $\hat{y}$  are the target and the predicted label vectors respectively, and  $[p] = 1$  if  $p$  is True and 0 otherwise. We use AUC since the data sets in this domain are typically highly unbalanced and the use of other performance measures such as 0/1 error is not useful in such



**Algorithm 2.2** LMGraph - Learning from Multiple Graphs

---

**Input:**  $k$  undirected graphs,  $G_d = (V, E_d)$  ( $|V| = n$ ) with adjacency matrix  $W_d$ , label vector  $y \in \{0, 1\}^n$ , training and validation set indices  $t, v \subset \{1, \dots, n\}$

**Output:** classifier parameters  $\{w_1, \dots, w_k\}$ , network weights  $\{\mu_1, \dots, \mu_k\}$

---

- 1: For all  $i \in \{1, \dots, n\}$ :  $\tilde{y}_i = y_i$  if  $i \in t$ , 0 otherwise  $\{\text{training labels}\}$
  - 2: **for**  $d = 1 \dots k$  **do**
  - 3:    $X_d = \text{3Prop}(G_d, \tilde{y})$   $\{\text{extract features}\}$
  - 4:    $w_d = \text{RLSR}((X_d)_t, (\tilde{y})_t)$   $\{\text{train classifier}\}$
  - 5:   For all  $i \in \{1, \dots, n\}$ :  $\tilde{y}_i = y_i$  if  $i \in v$ , 0 otherwise  $\{\text{validation labels}\}$
  - 6:    $\mu_d = \text{AUC}((\tilde{y})_v, (X_d)_v w_d)$   $\{\text{estimate network weights}\}$
  - 7: **end for**
  - 8: **return**  $\{w_1, \dots, w_k\}, \{\mu_1, \dots, \mu_k\}$
- 

scenarios. We note that Cortes *et al.*<sup>21</sup> used a learning method (for example, SVM, Lasso or RLSR) to learn the weights  $\{\mu_d\}$  of the independent models using their predictions on the validation set as “features.” While it is indeed possible to optimize AUC<sup>22</sup> and learn the weights  $\{\mu_d\}$ , we refrained from following this approach because training a RankSVM is computationally expensive. Our own experience with learning these weights by solving the constrained least-squares regression problem:  $\text{argmin}_{0 \leq \mu \leq 1} \sum_i (\sum_d \mu_d h_d(x_i) - y_i)^2$  did not result in performance gains when compared to simply using AUC for the network weights.

Algorithm 2.2 describes our method, LMGraph, for gene function prediction from multiple graphs. The final predictions are made according to  $\hat{y}(x) = \sum_d \mu_d h_d(x) = \sum_d \mu_d \cdot (w_d^\top x)$ . Given this algorithm, it is straightforward to integrate feature-based data sources with the functional interaction networks – we simply combine any additional feature-based data with the 3Prop feature sets extracted using Algorithm 2.1 and train LMGraph described in Algorithm 2.2 using these feature sets.

### 3. Results and Discussion

#### 3.1. Data sets and experimental setup

The data sets in our experiments consists of multiple functional interaction networks in three species – mouse, fly and *C.elegans*. The mouse data set consists of seven networks with 19,559 genes constructed from gene expression, protein interaction and domain composition data. To demonstrate the integration of feature-based data with networks, we also included 6,273 protein domain features extracted from Ensembl.<sup>23</sup> The fly data set consists of 28 networks with 13,457 genes constructed from genetic and physical interaction, co-expression, and co-localization data. The *C.elegans* data set consists of 30 networks with 18,946 genes constructed from co-expression and shared protein domain data. All the network and feature-based data sources were downloaded from the GeneMANIA prediction server.<sup>24</sup>

To evaluate the gene function prediction task, we use the GO biological process (BP) function categories<sup>25</sup> as target labels. In all the experiments, we use all the categories with at least 30 annotations. This resulted in 954, 963 and 724 categories (binary prediction tasks)

for the mouse, fly and *C.elegans* data sets respectively.

We report results from three main experiments: (i) In the first experiment, we evaluate the predictive ability of classifiers trained with 3Prop features extracted from graphs in direct comparison to label propagation (cf. (1)). In this experiment, we combine the networks using uniform weights. (ii) In the second experiment, we compare our algorithm to learn from multiple graphs, LMGraph, as described in Algorithm 2.2 with both label propagation and regularized least-squares regression trained on a composite network where the networks are combined using uniform weights. We use 3Prop features in all the comparisons. (iii) In the final experiment, we combine feature-based data and networks from the mouse data set and evaluate the performance of LMGraph.

In all the experiments, we split the data sets in a stratified manner into training, validation and test sets in the ratio of 3:1:1. We report the performance of the algorithms measured in terms of the average ranking error, i.e., 1-AUC (cf. (3)), on the test data sets from five such trials in all the experiments. We use the validation sets to tune the regularization parameter, selected from the set  $\{2^{-14}, 2^{-12}, \dots, 2^6, 2^8\}$ , in label propagation and regularized least-squares regression. We evaluate the statistical significance of the results based on the Wilcoxon signed-rank test when comparing the AUC for all the GO categories resulting from pairs of algorithms. In all the figures below, double asterisk (\*\*) indicates that the predictive performance gains due to our method are significant when compared to the competing/baseline methods with  $p \leq 0.005$ ; single asterisk (\*) indicates that the differences in performance are not significant.

### 3.2. Justification for 3Prop features

We begin with an empirical justification for extracting node features from random walks of length at most three. As mentioned before, Mostafavi *et al.*<sup>19</sup> argued and demonstrated empirically that random walks of length three suffice to propagate labels in several types of networks, including functional interaction networks. In most of these networks, random walks quickly converge to the stationary distribution over the nodes, and therefore longer random walks do not carry any discriminative information useful for labeling the nodes. Let  $\pi \in \mathbb{R}^n$  denote the stationary distribution with the property  $\lim_{r \rightarrow \infty} P^r = \mathbf{1}\pi^\top$  ( $\mathbf{1}$  is vector whose elements are all one), and is computed as  $\pi = d / \sum_i d_i$ , where  $d_i$  is the degree of the node  $i$ . The total variation distance between probability distributions  $p$  and  $q$  is defined as  $\delta(p, q) := (1/2) \sum_i |p_i - q_i|$ . This distance can be used to study the convergence of random walks. For a random walk starting from node  $i$ , we compute the total variation distance between the distribution  $e_i^\top P^r$  ( $e_i$  is a vector with 1 at position  $i$  and 0 elsewhere) and the stationary distribution  $\pi$ . The smallest value of  $r$  at which this distance falls below a fixed value  $\epsilon$  is known as the mixing time of the random walk, which gives us a measure of how close the distribution for random walk of length  $r$  is to the stationary distribution  $\pi$ . Typically,  $\epsilon$  is chosen to be 0.25. The total variation distance computed for varying random walk lengths is shown in Figure 1 for mouse, fly and *C.elegans* networks for 100 different random walks starting from 100 randomly chosen nodes. Each gray line shows the effect of random walk length on the total variation distance for a random walk starting from a random node, and the red line shows the median of 100 such random walks. For each species, we combine all

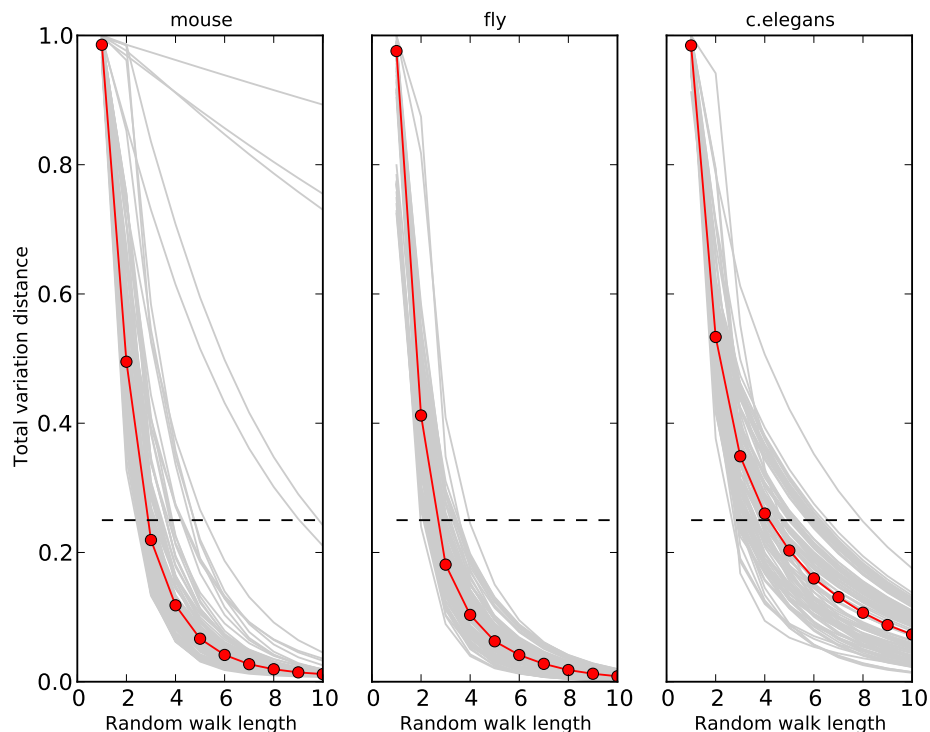


Fig. 1. Total variation distance for varying random walk lengths for mouse, fly and *C.elegans* networks, computed for 100 different random walks. Each gray line indicates a random walk starting from a random node, and the red line shows the median. The dashed line corresponds to a total variation distance of 0.25.

the networks with uniform weights and compute the transition probability matrix  $P = D^{-1}W$  from this combined network. From the figure, we observe that for  $r = 3$ , the total variation distance has dropped below or close to 0.25 for all the networks, thus confirming the findings of Mostafavi *et al.*<sup>19</sup> for the data sets used in our experiments.

### 3.3. 3Prop vs. LProp

We compare the performance of label propagation (LProp) with regularized least-squares regression (RLSR) trained with asymmetric and symmetric 3Prop features. In this experiment, we first combine all the networks for a given species with uniform weights to construct a single composite network. We then extract 3Prop features (cf. Algorithm 2.1) from this composite network and train RLSR with these features. For label propagation, we optimize the objective function (1) on the composite network. In Figure 2, we show the ranking errors<sup>b</sup> for both these methods trained on mouse, fly and *C.elegans* networks across all the GO biological process function categories. For all the networks, RLSR trained with both asymmetric and symmetric 3Prop features performs significantly better than or its performance is on par with label propagation.

<sup>b</sup>We do not show absolute AUC scores in the figures due to space constraints. In general, we observed percentage decrease in ranking errors across a wide range of AUC scores in all the experiments.

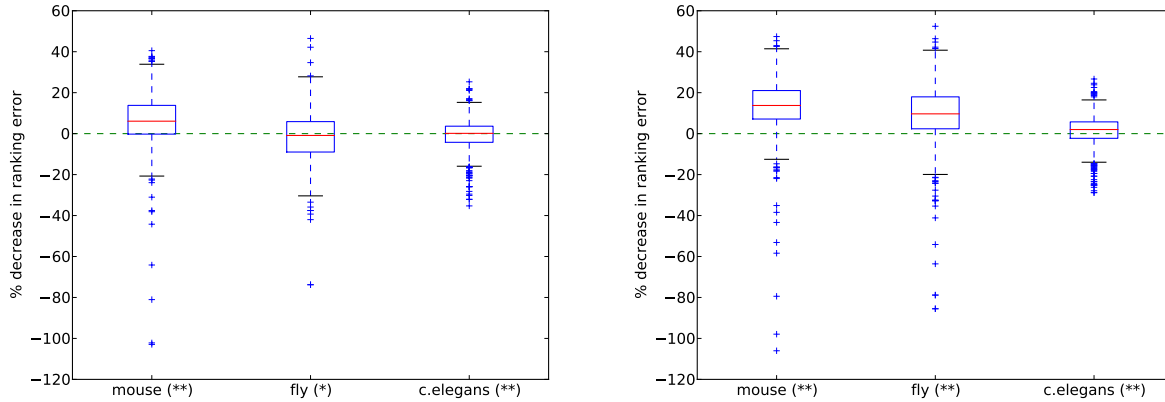


Fig. 2. Performance of label propagation and RLSR trained with asymmetric (*left*) and symmetric (*right*) 3Prop features on mouse, fly and *C.elegans* networks. The box plots show the percentage decrease in ranking error across all the GO biological process function categories.

This experiment clearly demonstrates the predictive ability of classifiers trained with 3Prop features and their potential as an alternative to label propagation for graph-based semi-supervised learning. We also observed that training a classifier with 3Prop features is computationally more efficient than label propagation even on sparse networks. As an example, on the mouse networks with 19,559 genes and a biological process function category with 100 annotations, the training time of RLSR with asymmetric and symmetric 3Prop features were 2.64s and 3.02s respectively for one trial which includes the tuning of regularization parameter, whereas label propagation took 35.29s for the same task on a 2 x 2.66 GHz dual-core processor with 4 GB of memory.

### 3.4. LMGraph vs. LProp and 3Prop

In this experiment, we first compare the performance of LMGraph with label propagation trained on composite networks combined using uniform weights. The results are shown in Figure 3 for asymmetric and symmetric 3Prop features. On all the networks, LMGraph performs significantly better than label propagation for both asymmetric and symmetric 3Prop features. On the *C.elegans* data set, we found that a slightly different version of Algorithm 2.2 wherein we first use the estimated weights  $\{\mu_d\}$  to construct a composite network and then extract 3Prop features from the resulting network, followed by training an RLSR with these features performed better than the ensemble method described in Algorithm 2.2.

We also compare the performance of LMGraph with RLSR trained using 3Prop features extracted from the composite networks combined using uniform weights. The results are shown in Figure 4 for asymmetric and symmetric 3Prop features. Here, we found that LMGraph did not significantly boost the predictive performance, especially for symmetric 3Prop features and the *C.elegans* networks. However, we would like to emphasize that RLSR trained on composite networks combined using uniform weights, where we first combine the networks and then extract 3Prop features, is a strong baseline. In fact, this is a much stronger baseline than training RLSR with all the  $3 \times K$  3Prop features and this was verified by us in our

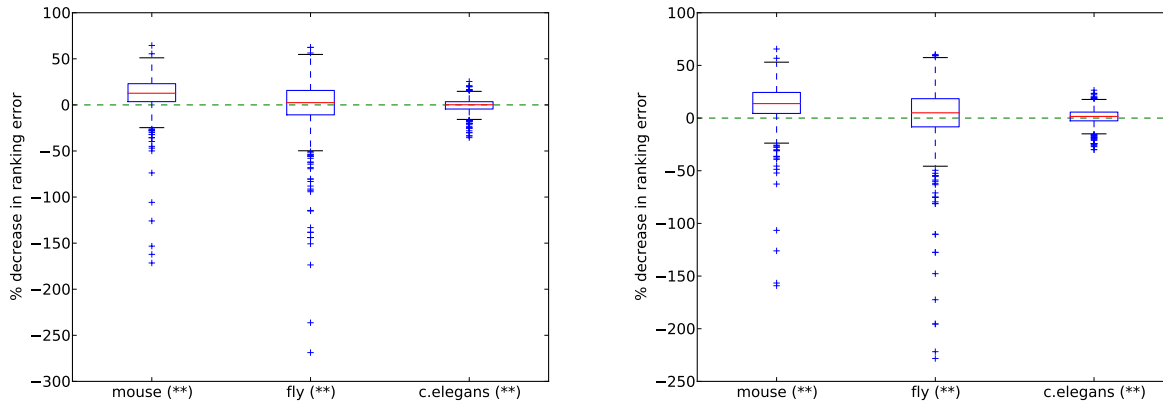


Fig. 3. Performance of LMGraph trained with asymmetric (*left*) / symmetric (*right*) 3Prop features and label propagation on mouse, fly and *C.elegans* networks.

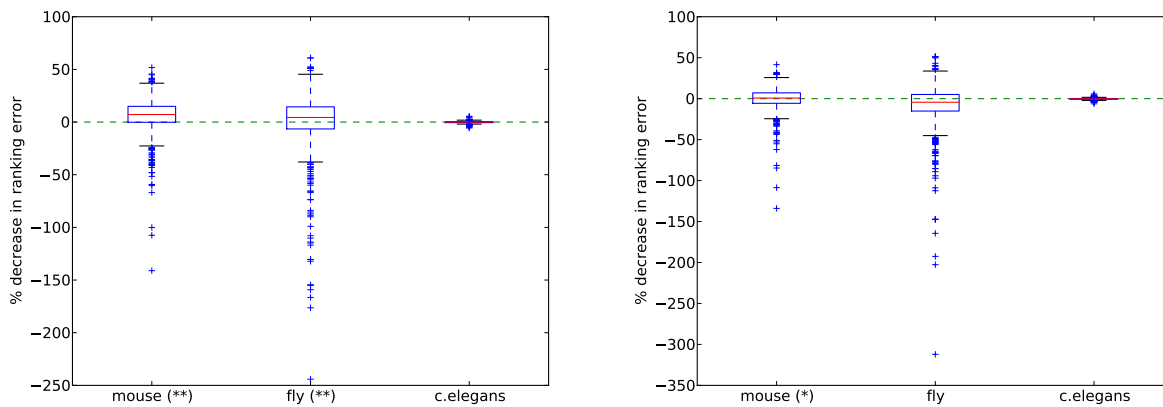


Fig. 4. Performance of LMGraph and RLSR trained with asymmetric (*left*) and symmetric (*right*) 3Prop features on mouse, fly and *C.elegans* networks.

experiments. Indeed, as is clearly evident from Figure 2 in the previous experiment, we see that these classifiers with a simple uniform weighting scheme performed significantly better than label propagation trained using the same composite networks.

### 3.5. *LMGraph with features*

In the final experiment, we integrate protein domain features into our learning algorithm to demonstrate the benefits of combining feature-based data sources with functional interaction networks. The results are shown in Figure 5 for the mouse networks. Integrating feature-based data into LMGraph results in significant improvements in AUC for both asymmetric and symmetric 3Prop features when compared to LMGraph trained using only the network data. Furthermore, when compared to RLSR trained with 3Prop features extracted from a composite network combined with uniform weights, we found that LMGraph trained with the protein domain features results in significant performance gains for both asymmetric and symmetric 3Prop features. LMGraph with features also performs significantly better than

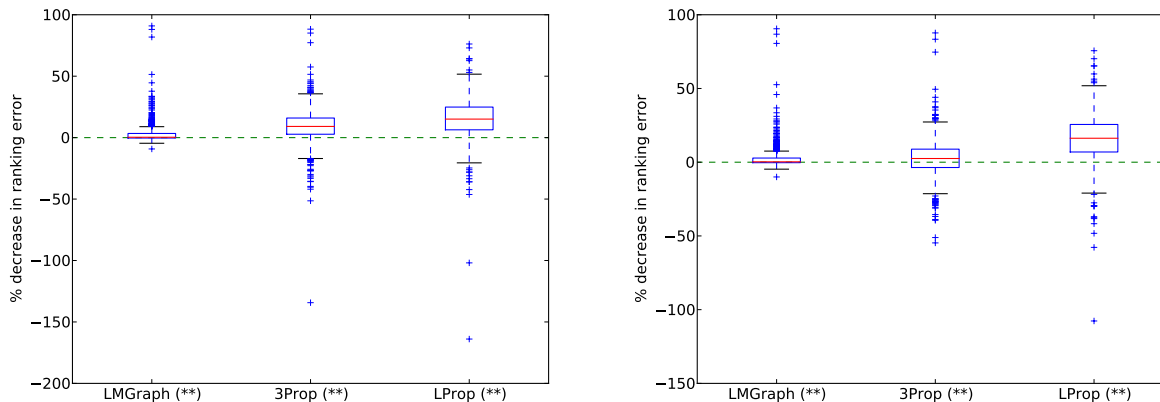


Fig. 5. Performance comparison of LMGraph trained using asymmetric (*left*) / symmetric (*right*) 3Prop and protein domain features with LMGraph, RLSR (3Prop) and label propagation (LProp) trained on a composite network combined with uniform weights on the mouse networks.

label propagation trained on a composite network combined with uniform weights; we note that it is not possible to combine features with label propagation using existing methods.<sup>12,20</sup>

#### 4. Conclusions

We proposed a computationally efficient machine learning algorithm, *LMGraph*, for gene function prediction from multiple functional interaction networks. The crux and novelty of our algorithmic contribution lies in the computationally efficient two-step procedure that allows us to combine multiple graph-based and feature-based data sources, where in the first step we extract features from the nodes of graphs and in the second step we combine these feature sets and train *linear* predictors. Our feature extraction method is based on the method proposed by Mostafavi *et al.*,<sup>19</sup> which is known to work well on functional interaction networks. We used a variant of the ensemble method proposed by Cortes *et al.*<sup>21</sup> to combine multiple data sources since (i) it has been shown to perform better than the traditional MKL methods,<sup>14</sup> (ii) it has been shown to outperform the strong uniform weighting baseline, and (iii) it is extremely easy to implement as a machine learning practitioner in bioinformatics. However, we would like to emphasize that the user is free to design and use other relevant feature extraction methods on graphs such as spectral embeddings and also other standard multiple kernel learning methods (with a linear kernel) in these steps.

We have shown experimentally that training linear predictors with symmetric and/or asymmetric graph-based 3Prop features is a viable alternative to label propagation. Furthermore, using these features in LMGraph resulted in significant performance gains when compared to propagating labels on composite networks combined using uniform weights which is known to be a hard baseline.<sup>12,20,21</sup> We have also demonstrated that our method, LMGraph, can be used to combine attribute data with functional interaction networks and that this combination can result in significant performance gains for gene function prediction tasks.

**Acknowledgments** SV was supported by a Natural Science and Engineering Research Council of Canada operating grant to QM. We would like to thank Khalid Zuberi and Sara Mostafavi for help compiling the data sets.

## References

1. A.-L. Barabási, N. Gulbahce and J. Loscalzo, *Nature Reviews Genetics* **12**, 56 (2011).
2. X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Molecular Systems Biology* **4** (2008).
3. S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan *et al.*, *Nature biotechnology* **24**, 537 (2006).
4. X. Zhu, Z. Ghahramani and J. D. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in *Proceedings of the International Conference on Machine Learning*, 2003.
5. D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, Learning with local and global consistency, in *Advances in Neural Information Processing Systems 16*, 2003.
6. S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios and Q. Morris, *Genome Biology* **9**, p. S4 (2008).
7. L. Peña-Castillo, M. Tasan, C. L. Myers, H. Lee, T. Joshi, C. Zhang, Y. Guan, M. Leone, A. Pagnani, W. K. Kim *et al.*, *Genome Biol* **9**, p. S2 (2008).
8. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates and D. Eisenberg, *Nature* **402**, 83 (1999).
9. P. Pavlidis, J. Weston, J. Cai and W. S. Noble, *Journal of Computational Biology* **9**, 401 (2002).
10. G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan and W. S. Noble, *Bioinformatics* **20**, 2626 (2004).
11. S. Mostafavi and Q. Morris, *Proteomics* **12**, 1687 (2012).
12. K. Tsuda, H. Shin and B. Schölkopf, Fast protein classification with multiple networks, in *Proceedings of the Fourth European Conference on Computational Biology and the Sixth Meeting of the Spanish Bioinformatics Network (Jornadas de BioInformática)*, 2005.
13. D. A. Spielman and S.-H. Teng, Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems, in *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2004.
14. G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui and M. I. Jordan, *Journal of Machine Learning Research* **5**, 27 (2004).
15. S. Sonnenburg, G. Rätsch, C. Schäfer and B. Schölkopf, *Journal of Machine Learning Research* **7**, 1531 (2006).
16. S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt and M. Varma, Multiple kernel learning and the SMO algorithm, in *Advances in Neural Information Processing Systems 23*, 2010.
17. A. J. Smola and R. I. Kondor, Kernels and regularization on graphs, in *Proceedings of the Annual Conference on Computational Learning Theory and the Seventh Kernel Workshop*, 2003.
18. S. Shalev-Shwartz, Y. Singer, N. Srebro and A. Cotter, *Mathematical Programming, Series B* **127**, 3 (2011).
19. S. Mostafavi, A. Goldenberg and Q. Morris, *PLoS ONE* **7**, p. e51947 (12 2012).
20. S. Mostafavi and Q. Morris, *Bioinformatics* **26**, 1759 (2010).
21. C. Cortes, M. Mohri and A. Rostamizadeh, Ensembles of kernel predictors, in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2011.
22. R. Herbrich, T. Graepel and K. Obermayer, Large margin rank boundaries for ordinal regression, in *Advances in Large Margin Classifiers*, eds. A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (MIT Press, Cambridge, MA, 2000).
23. Flicek *et al.*, *Nucleic Acids Research* **40**, D84 (2012).
24. Warde-Farley *et al.*, *Nucleic Acids Research* **38**, 214 (2010).
25. Ashburner *et al.*, *Nature genetics* **25**, 25 (2000).

# MATRIX FACTORIZATION-BASED DATA FUSION FOR GENE FUNCTION PREDICTION IN BAKER'S YEAST AND SLIME MOLD

MARINKA ŽITNIK

*Faculty of Computer and Information Science, University of Ljubljana,  
Tržaška 25, SI-1000, Slovenia  
E-mail: marinka.zitnik@fri.uni-lj.si*

BLAŽ ZUPAN

*Faculty of Computer and Information Science, University of Ljubljana,  
Tržaška 25, SI-1000, Slovenia  
Department of Molecular and Human Genetics, Baylor College of Medicine,  
Houston, TX-77030, USA  
E-mail: blaz.zupan@fri.uni-lj.si*

The development of effective methods for the characterization of gene functions that are able to combine diverse data sources in a sound and easily-extendible way is an important goal in computational biology. We have previously developed a general matrix factorization-based data fusion approach for gene function prediction. In this manuscript, we show that this data fusion approach can be applied to gene function prediction and that it can fuse various heterogeneous data sources, such as gene expression profiles, known protein annotations, interaction and literature data. The fusion is achieved by simultaneous matrix tri-factorization that shares matrix factors between sources. We demonstrate the effectiveness of the approach by evaluating its performance on predicting ontological annotations in slime mold *D. discoideum* and on recognizing proteins of baker's yeast *S. cerevisiae* that participate in the ribosome or are located in the cell membrane. Our approach achieves predictive performance comparable to that of the state-of-the-art kernel-based data fusion, but requires fewer data preprocessing steps.

*Keywords:* gene function prediction, data fusion, matrix factorization, Gene Ontology annotation, membrane protein, ribosomal protein

## 1. Introduction

Assigning functions to genes and proteins is a major challenge of biological research. Recent genome-scale data capture distinct but possibly noisy and incomplete views of cellular function. Collectively, these data provide valuable information for inference of gene and protein functions but require computational approaches capable of joint treatment of heterogeneous data sources.

Gene function prediction aims to provide a set of functional terms along with associated confidence for a given uncharacterized or partially characterized gene. In this work, we take a step towards improved gene function prediction through fusion of data sets that are either directly related to genes, such as genetic interactions, or are circumstantial, such as Medical Subject Headings (MeSH) terms assigned to the relevant biomedical literature. In our previous work, we proposed a matrix factorization-based data fusion<sup>1</sup> and demonstrated its utility in detection of drug-toxicity.<sup>2</sup> Its advantage over some well-known approaches that infer prediction models through integrative data analysis is its ability to directly consider data modality



and to retain the structure of data representation during fusion. Our algorithm can include any data source that can be represented in a matrix whereby the concrete selection of data sources depends on the given function prediction task.

Methods for gene function prediction often consider a metric space of genes, that is, a gene set equipped with a notion of distance or similarity between any pair of genes.<sup>3-6</sup> All available data has to be expressed through relations between genes and their functions, although for specific data sources that might not be natural in any sense. For instance, to include the semantic structure of the MeSH terms into the prediction model we should design a metric that would, for a pair of genes, measures the distance between the MeSH terms that are assigned to relevant gene-pair-associated literature. Such distance function is hard to construct, and for integration of many heterogeneous data sources, becomes a major obstacle in development of prediction system. Our approach can consider circumstantial evidence for gene function prediction directly even if expressed in a non-gene space. Its principal novelty is the ease of adding new data sources without requiring their substantial preprocessing or transformation. Data sources are simultaneously considered during data fusion and construction of predictive model.

In the paper we outline our previously proposed data fusion algorithm<sup>2</sup> and then study it in computational experiments on three function prediction tasks for baker's yeast and slime mold's genome-wide data sets. We fuse eleven data sources to predict the Gene Ontology (GO)<sup>7</sup> annotations in slime mold *D. discoideum* and investigate the recognition of particular classes of proteins in baker's yeast *S. cerevisiae* by combining four data sources on cytoplasmic ribosomal class and four sources on membrane proteins. Our principal contribution in this work is a demonstration that matrix-based data fusion approach can be applied to gene function prediction problem and can successfully integrate a diverse set of data sources, thus raising the accuracy of predictions.

## 2. Related Work

Methods to predict gene annotations either follow approaches that transfer annotations from well-characterized to partially characterized genes,<sup>3,8</sup> or approaches that directly associate genes with functional classes using supervised learning.<sup>5,9-13</sup> Although annotation transfer is appealing at first sight, excessive transferring causes error propagation and is often outperformed by sophisticated classification algorithms.<sup>14</sup>

Recent methodological contributions to gene function prediction aim at extracting features from different biological data sets and use them to train classifiers for functional categories, such as GO terms or KEGG pathways.<sup>14</sup> They derive features from gene expression profiles, genetic interactions, protein-protein interaction networks, conserved protein domains, sequence similarity, physiochemical properties, co-expression and data on orthologs. For example, Vinayagam *et al.* (2004)<sup>9</sup> and Mitsakakis *et al.* (2013)<sup>13</sup> both applied support vector machines for the classification of GO terms from sequence data and microarray experiments, respectively, and Yan *et al.* (2010)<sup>11</sup> trained a random forest classifier for each functional category separately and tested their prediction model on data from fruit fly. The accuracy of developed methods for gene function prediction has been further improved by integrat-

ing data using multi-classifier approaches,<sup>12</sup> Bayesian reasoning,<sup>3,4,10,15</sup> network-based analysis<sup>5,16,17</sup> and kernel functions derived from different sources by multiple kernel learning.<sup>18,19</sup> Automated gene and protein function prediction methods are often trained to only one species, are not available for high-volume and heterogeneous data, or require the use of data derived by experiments, such as microarray analysis. The approach we proposed in this manuscript is organism-independent, it can be applied for various subsets of functional terms and it provides confidence estimates of predictions. Also, it does not impose any restrictions on the nature of underlying data.

Due to great potential of methods for computational prediction of gene function we recently witnessed several initiatives<sup>6,20,21</sup> for the critical assessment of their performance in different experimental settings. These evaluations concluded that although best methods perform well enough to guide the experiments, there is considerable need for improvement of currently available approaches one of which is efficient data integration.

### 3. Methods

Matrix factorization-based data fusion<sup>1</sup> can in principle consider an unlimited number of data sources. In the context of gene function prediction, these could either describe characteristics of genes and proteins directly (e.g., their physical interactions) or indirectly (e.g., through MeSH terms that are assigned to scientific publications, which in turn mention the genes of interest). Fig. 1 provides a toy example that combines five data sources on objects of three different types: genes, GO terms and experimental conditions. Given a multitude of data sources, we assume that each source describes relations between objects of two types. Data fusion by matrix factorization involves three main steps. First, every data source is represented as a matrix and together they are organized in a block-based matrix representation (Fig. 1, left; Sec. 3.2). *Constraint matrices*,  $\Theta_i$ , relate objects of type  $i$  and are placed on the main diagonal of block representation. The off-diagonal blocks, which relate objects of different types,  $i$  and  $j$  ( $i \neq j$ ), are called *relation matrices*,  $\mathbf{R}_{ij}$ . We expect that these matrices are sparse and that some are completely missing because associated data sources are not available. For example, a missing source from Fig. 1 would relate GO terms to experimental conditions. Second, we simultaneously factorize all relation matrices such that low-rank matrix factors are shared between decompositions of relation matrices that describe objects of common type (Fig. 1, middle; Sec. 3.3). Constraints indicate pairwise similarities or dissimilarities (it depends on signs of values) between the two objects. If constraints are violated, for instance, if two highly similar objects have very different low-rank profiles (i.e. corresponding rows in matrix factors), then current low-rank matrix approximations are penalized. Finally, we employ low-rank matrix factors to complete unobserved entries in relation matrices, to predict GO terms and to estimate confidence of predictions (Fig. 1, right; Sec. 3.4 and Sec. 3.5).

We apply data fusion to infer relations between genes or proteins and their functions. We observe target relation matrix in the context of all other data sources. We assume that it is encoded as a  $[0,1]$ -matrix that is only partially observed. Its entries indicate a degree of relation, 0 denoting that corresponding function is absent from the gene and 1 denoting the highest confidence that gene performs a specific function. We aim to predict its unobserved

entries by reconstructing them through matrix factorization.

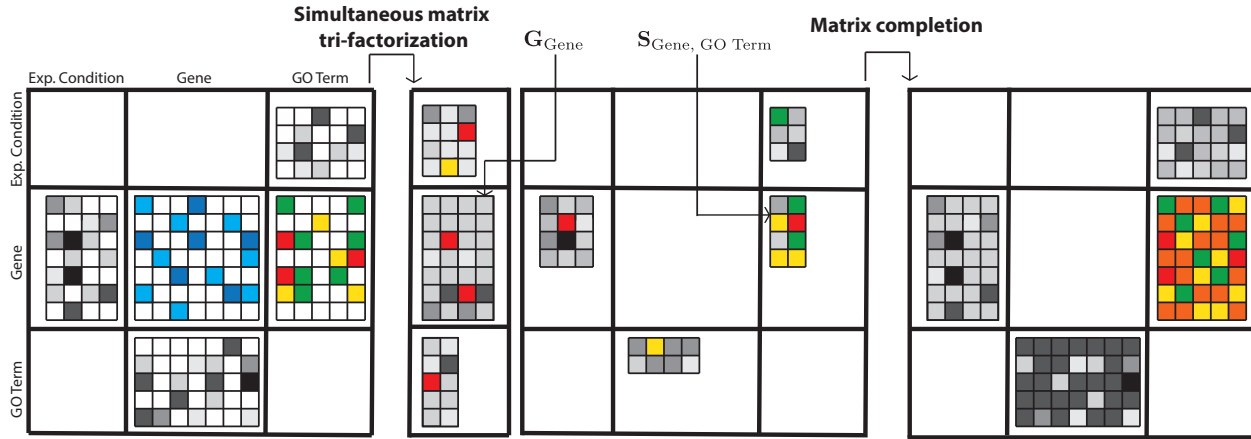


Fig. 1: An example of data fusion by matrix factorization that combines five data sources on objects of three different types: genes, Gene Ontology (GO) terms and experimental conditions. Target matrix relates genes to GO terms (matrix with colorful entries). Data is presented in a block-based system (left), then a compressed representation is inferred that shares low-rank matrix factors between decompositions of relation matrices (shown by matrices with grey entries), which relate objects of common type (middle). Constraint matrices (shown by matrix with blue entries) penalize violations of similarity constraints. Finally, original matrix of gene annotations is completed (right).

### 3.1. Data

#### 3.1.1. Gene Annotation Prediction in Slime Mold

In this study we observe objects of six different types: genes (type 1), GO terms (type 2), experimental conditions (type 3), publications from the PubMed database (PMID) (type 4), MeSH descriptors (type 5), and KEGG<sup>a</sup> pathways (type 6). The organization of object types and data sources is shown in Fig. 2a; fusion algorithm can integrate all available data if the underlying graph is connected. We include gene expression measurements at different time-points of a 24-hour development cycle<sup>22</sup> ( $\mathbf{R}_{13}$ , 14 experiments), gene annotations with experimental evidence code to 148 generic slim terms from the GO ( $\mathbf{R}_{12}$ ), associations of PMIDs and genes from dictyBase<sup>b</sup>, March, 2013 ( $\mathbf{R}_{14}$ ), genes participating in KEGG pathways ( $\mathbf{R}_{16}$ ), assignments of MeSH descriptors to publications from PubMed ( $\mathbf{R}_{45}$ ), references to published work associated with GO terms ( $\mathbf{R}_{42}$ ), and associations of enzymes involved in KEGG pathways and related to GO terms ( $\mathbf{R}_{62}$ ). To balance the target matrix  $\mathbf{R}_{12}$  for the purpose of performance evaluation we add an equal number of non-associations for which there is no evidence of any type in the GO.

We consider protein interaction scores from STRING v9.0<sup>c</sup> ( $\Theta_1$ ), the number of common ortholog groups between KEGG pathways ( $\Theta_6$ ) and slim term similarity scores ( $\Theta_2$ ) that are

<sup>a</sup><http://www.kegg.jp>

<sup>b</sup><http://dictybase.org/Downloads>

<sup>c</sup><http://string-db.org>

computed as  $-0.8^{\text{hops}}$ , where hops is the length of the shortest path between two terms in the GO graph. Similarly, MeSH descriptors are constrained with the average number of hops between each pair of descriptors in the MeSH hierarchy ( $\Theta_5$ ).

### 3.1.2. Yeast Ribosomal Protein Classification

We observe three object types: proteins (type 1), cellular complexes (type 2) and experimental conditions (type 3). Their relations are described by four data sources that correspond to arcs in Fig. 2b. We consider the MIPS Comprehensive Yeast Genome Database (CYGD)<sup>d</sup> assignments of 1150 yeast proteins to cellular complexes, of which 134 participate in the ribosome and the remaining  $\sim 5000$  yeast proteins are unlabeled.<sup>18</sup> We include gene expression measurements from the Stanford Microarray Database ( $\mathbf{R}_{13}$ , 441 experiments), protein interactions from STRING v9.0<sup>e</sup> ( $\Theta_1^{(1)}$ ) and Smith-Waterman pairwise sequence comparisons ( $\Theta_1^{(2)}$ ).

### 3.1.3. Yeast Membrane Protein Classification

We consider four data sources and three types of objects (Fig. 2c): proteins (type 1), subcellular locations (type 2) and Pfam<sup>e</sup> protein domain families. We consider subcellular location information of 2318 yeast proteins from the CYGD<sup>d</sup> database ( $\mathbf{R}_{12}$ ), of which 497 belong to various membrane protein classes and  $\sim 4000$  proteins have uncertain location.<sup>18</sup> We include the expectation values from the hidden Markov models in the Pfam database ( $\mathbf{R}_{13}$ ). Matrices  $\Theta_1^{(1)}$  and  $\Theta_1^{(2)}$  from Fig. 2c have the same meaning as for the ribosomal protein classification.

In both yeast experiments the target  $\mathbf{R}_{12}$  has a  $(6112 \times 2)$ -shape, where a row of  $[0, 1]$  denotes that the protein participates in ribosome or that it belongs to membrane protein class and a row of  $[1, 0]$  that the protein is not assigned to the ribosomal complex or that it does not belong to membrane protein class. Rows that correspond to unobserved proteins are set to  $[0.5, 0.5]$ .

## 3.2. Block-Based Data Representation

The data on slime mold from Sec. 3.1.1 can be represented in a block-based system:

$$\mathbf{R} = \begin{bmatrix} \mathbf{0} & \mathbf{R}_{12} & \mathbf{R}_{13} & \mathbf{R}_{14} & \mathbf{0} & \mathbf{R}_{16} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{42} & \mathbf{0} & \mathbf{0} & \mathbf{R}_{45} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{62} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \Theta^{(1)} = \text{Diag}(\Theta_1, \Theta_2, \mathbf{0}, \mathbf{0}, \Theta_5, \Theta_6). \quad (1)$$

The number of non-zero blocks corresponds to the number of included data sources. Such representation is then fed into fusion algorithm. The block-based schemes for yeast-related

<sup>d</sup><http://mips.helmholtz-muenchen.de/genre/proj/yeast>

<sup>e</sup><http://pfam.sanger.ac.uk>

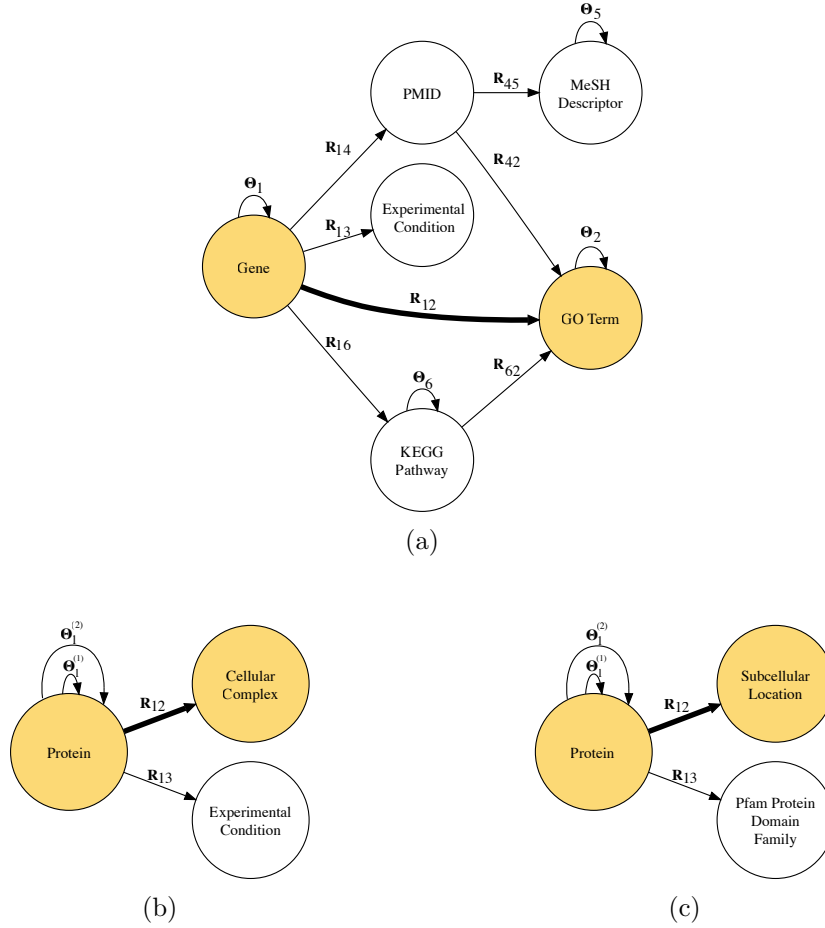


Fig. 2: Fusion configurations for the gene function prediction task in slime mold (a) and two yeast protein classification tasks to recognize cytoplasmic ribosomal proteins (b) and membrane proteins (c). Nodes represent types of objects and arcs correspond to relation and constraint matrices. The arcs that represent target matrices,  $\mathbf{R}_{12}$ , and their object types are highlighted.

data (Sec. 3.1.2 and Sec. 3.1.3) have the structure from Eq. (2), where the individual matrices are task-dependent:

$$\mathbf{R} = \begin{bmatrix} \mathbf{0} & \mathbf{R}_{12} & \mathbf{R}_{13} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \boldsymbol{\Theta}^{(t)} = \text{Diag}(\boldsymbol{\Theta}_1^{(t)}, \mathbf{0}, \mathbf{0}) \text{ for } t = 1, 2. \quad (2)$$

Our fusion approach is different from treating an entire system from Eq. (1) or Eq. (2) as a single large matrix. Factorization of such a matrix would disregard the structure from Eq. (1) and Eq. (2).<sup>1</sup>

### 3.3. Data Fusion by Matrix Factorization

Approximate matrix factorization estimates matrix  $\mathbf{R}_{ij}$  as a product of low-rank matrix factors that are found by solving an optimization problem, which maximizes some quality of approximation. A tri-factor decomposition, which we use in this study, decomposes  $\mathbf{R}_{ij}$  into

a product of three low-dimensional matrix factors such that  $\mathbf{R}_{ij} \approx \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T$  (Fig. 3).

$$\begin{array}{c} n_j \\ \boxed{\mathbf{R}_{ij}} \\ n_i \end{array} \approx \begin{array}{c} k_i \\ \boxed{\mathbf{G}_i} \\ n_i \end{array} \times \begin{array}{c} k_j \\ \boxed{\mathbf{S}_{ij}} \\ k_j \end{array} \times \begin{array}{c} n_j \\ \boxed{\mathbf{G}_j^T} \\ n_j \end{array}$$

Fig. 3: Matrix tri-factorization. Matrix  $\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$  relates objects of two types,  $i$  and  $j$ . For instance, we might relate genes to their expression profiles, publications to assigned MeSH terms or genes to themselves if they interact genetically.  $\mathbf{R}_{ij}$  is decomposed into a product of three matrix factors such that  $\mathbf{R}_{ij} \approx \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T$ , where  $\mathbf{G}_i \in \mathbb{R}^{n_i \times k_i}$ ,  $\mathbf{G}_j \in \mathbb{R}^{n_j \times k_j}$  and  $\mathbf{S}_{ij} \in \mathbb{R}^{k_i \times k_j}$ ,  $k_i \ll n_i$ ,  $k_j \ll n_j$ .

For data fusion we use simultaneous penalized tri-factorization to simultaneously decompose all blocks  $\mathbf{R}_{ij}$  while considering constraints in  $\Theta_i^{(t)}$  for  $t = 1, 2, \dots, t_i$ . The block matrix  $\mathbf{R}$  from Eq. (1) is tri-factorized into block matrices  $\mathbf{S}$  and  $\mathbf{G}$ :

$$\mathbf{S} = \begin{bmatrix} \mathbf{0} & \mathbf{S}_{12} & \mathbf{S}_{13} & \mathbf{S}_{14} & \mathbf{0} & \mathbf{S}_{16} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{42} & \mathbf{0} & \mathbf{0} & \mathbf{S}_{45} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{62} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{G} = \text{Diag}(\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3, \mathbf{G}_4, \mathbf{G}_5, \mathbf{G}_6). \quad (3)$$

Yeast data matrix in Eq. (2) is similarly decomposed into block matrix factors  $\mathbf{S}$  and  $\mathbf{G}$ , each having  $3 \times 3$  block-shape but we omit them here for brevity. Such factorization of block-based representation retains the block structure of our systems from Eq. (1) and Eq. (2). Matrix factors  $\mathbf{S}_{ij}$  in the resulting factorized system are specific to every data source and factors  $\mathbf{G}_i$  are specific to every object type. Factor  $\mathbf{G}_i$  is present in decompositions of all relation matrices that relate objects of type  $i$  to objects of some other type, whereas  $\mathbf{S}_{ij}$  is used only for decomposing  $\mathbf{R}_{ij}$ . Thus, they capture object type- and source-specific patterns, respectively. Sharing matrix factors between decompositions with common object type is the key idea of our data fusion approach.

The objective function minimized by simultaneous penalized matrix tri-factorization ensures good approximation of the input data and adherence to constraints, which are represented in constraint matrices:

$$\min_{\mathbf{G} \geq 0} \|\mathbf{R} - \mathbf{G} \mathbf{S} \mathbf{G}^T\| + \sum_{t=1}^{\max_i t_i} \text{tr}(\mathbf{G}^T \Theta^{(t)} \mathbf{G}), \quad (4)$$

where  $\|\cdot\|$  and  $\text{tr}(\cdot)$  denote the Frobenius norm and trace, respectively. Updating rules for decomposing relation matrices,<sup>1</sup> iteratively improve matrix factors  $\mathbf{G}$  and  $\mathbf{S}$ , which converge to a local minimum of the optimization problem in Eq. (4). The algorithm first initializes factors  $\mathbf{G}_i$  and then successively updates  $\mathbf{G}$  and  $\mathbf{S}$  until stopping criteria is met. See Žitnik *et al.* (2013)<sup>1</sup> for details about initialization algorithm, updating rules and stopping criteria.

### 3.4. Predicting Gene Functions from Matrix Factors

Our target  $\mathbf{R}_{12}$  is a partially observed  $[0, 1]$ -matrix, where 1 indicates that gene is assigned the corresponding function and 0 that it is not. We complete it as:  $\hat{\mathbf{R}}_{12} = \mathbf{G}_1 \mathbf{S}_{12} \mathbf{G}_2^T$ . When the fused model is requested to propose relations for a new gene  $g$  that was not included in the training data, we need to estimate its factorized representation and use the resulting factors for prediction. We formulate non-negative linear least-squares and solve them for  $\min_{\mathbf{h}_i \geq 0} \|\mathbf{G}_i \mathbf{S}_{1i}^T \mathbf{h}_i - \mathbf{g}_i\|_2$ , where  $\mathbf{g}_i \in \mathbb{R}^{n_i}$  is the original description of gene  $g$  in  $i$ -th data source and  $\mathbf{h}_i \in \mathbb{R}^{k_i}$  is its factorized representation. Here,  $i$  varies from 2 to the number of data sources used for fusion. A solution vector given by  $\sum_{i>1} \mathbf{h}_i^*$  is added as a new row to  $\mathbf{G}_1$  and new  $\hat{\mathbf{R}}_{12}$  is computed.

We then identify gene-function pairs  $(g, f^*)$  for which the predicted degree of relation  $\hat{\mathbf{R}}_{12}(g, f^*)$  is unusually high. Candidate functions for gene  $g$  have greater estimated association score than the mean estimated score of all known annotations of gene  $g$ :

$$\hat{\mathbf{R}}_{12}(g, f^*) > \frac{1}{|\mathcal{A}(g)|} \sum_{f \in \mathcal{A}(g)} \hat{\mathbf{R}}_{12}(g, f), \quad (5)$$

where  $\mathcal{A}(g)$  contains functions annotated to  $g$ . Eq. (5) is a gene-centric rule. Given a test gene, it identifies functional terms to which it might be assigned. If the gene does not have any known annotations we use the function-centric rule to identify gene-function candidate pairs.

### 3.5. Assessing Strength of Predictions

We combine the gene- and function-centric rules such that, if possible, the gene-centric rule is applied to identify gene-function candidate pairs and then the function-centric rule is used to assess the strength of the candidate pair  $(g, f^*)$ . We estimate the strength of association of gene  $g$  to function  $f^*$  by reporting an inverse percentile of association score in the distribution of scores for all true annotations to function  $f^*$ , that is, by considering the scores in the  $f^*$ -th column of  $\hat{\mathbf{R}}_{12}$  (Fig. 4). Higher value indicates higher confidence of prediction.

## 4. Performance Evaluation

We estimated the performance by ten-fold cross-validation. In each fold, we split the gene set to a train and test set. The data on genes from the test set were entirely omitted from the training data. We developed prediction models from the training data and tested them on the genes from the test set. The performance was evaluated using an  $F_1$  score, a harmonic mean of precision and recall, which was averaged across cross-validation runs. We selected the parameters of our data fusion algorithm, factorization ranks for each type of objects ( $k_i$ ), by observing the quality of  $\hat{\mathbf{R}}_{12}$  in internal cross-validation.<sup>1</sup> The parameters for kernel-based fusion, such as width of an RBF kernel and regularization weight, were also selected through internal cross-validation.

## 5. Kernel-Based Fusion Setup

We compared our data fusion algorithm to state-of-the-art integration by multiple kernel learning (MKL; Yu *et al.* (2010)<sup>19</sup>) that follows a multi-label classification approach. Kernel-

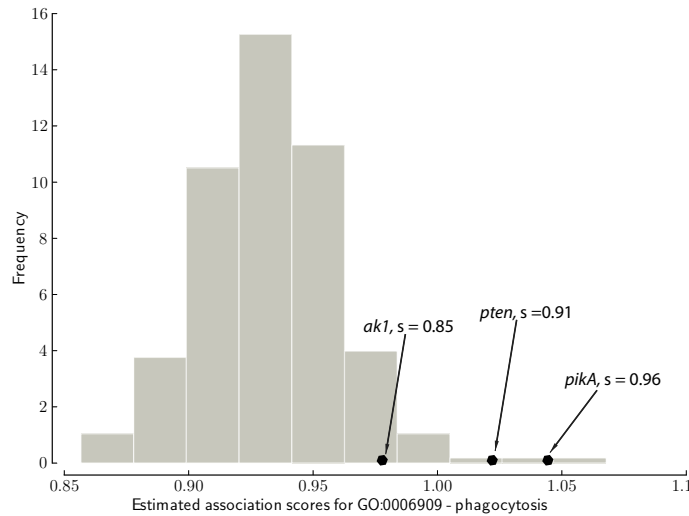


Fig. 4: An example of estimating strength of candidate slime mold genes for “phagocytosis” term. Association scores from  $\hat{\mathbf{R}}_{12}$  for all genes that are annotated with term “phagocytosis” are shown in grey. Strength of the candidate pair (*pikA*, “phagocytosis”),  $s = 0.96$ , is assessed by reporting its inverse percentile in the distribution of scores for true annotations (in grey). That is, the percentage of estimated association scores that are smaller or equal to the score of (*pikA*, “phagocytosis”).

based fusion used a multi-class  $L_2$  norm MKL with Vapnik’s SVM. The MKL was formulated as a second order cone program and solved using the conic optimization solver SeDuMi<sup>f</sup>. We generated the kernel matrices for yeast experiments using the kernels proposed by Lanckriet *et al.* (2004).<sup>18</sup> In slime mold study, we applied an RBF kernel to gene expression measurements and three linear kernels to protein interactions, genes that participate in KEGG pathways and to associations of genes to PMIDs. Data sources that describe relations between object types other than genes had to be transformed to explicitly relate them to genes. We represented the hierarchical structure of MeSH descriptors, semantic structure of the GO graph and KEGG ortholog groups as separate weighted graphs on genes (for instance, we counted common KEGG ortholog groups and calculated the similarities of sets of GO terms associated with genes) and constructed kernel matrices using diffusion kernel.

## 6. Results and Discussion

We evaluated our algorithm from the perspective of genes and functional terms. Thus, we addressed two related questions: “What is the function of a particular gene or protein?” and “What are the genes or proteins associated with a particular functional term?”.

### 6.1. Performance on Groups of Target Genes

We divided the *D. discoideum* gene set into three categories to compare predictive performance in each category. In Table 1 we present the cross-validated  $F_1$  scores when selecting the 100 or 1000 most GO-annotated genes and the accuracy obtained when considering whole slime

<sup>f</sup><http://sedumi.ie.lehigh.edu>



mold genome. The task was to provide a set of terms from the slim subset of GO terms for every gene. We used the slim subset of GO terms to limit the optimization complexity of the kernel-based approach<sup>18</sup> with which we compare our performance. These categories were selected to study the effects of data sparseness. Genes with many GO annotations tend to be better characterized and more data is available about them. Thus, functional terms of such genes would be considered easier to predict than those of genes with only few annotations. The accuracy of our matrix factorization-based data fusion is comparable to that of kernel-based approach. The performance of both approaches improved when we included more genes and hence more data. Also, our approach performed well when we added genes with sparser profiles although that increased the overall data sparsity.

Table 1: Cross-validated  $F_1$  scores for fusion by matrix factorization (MF) and kernel-based method (MKL).

Slime mold task	MF	MKL
100 genes	0.799	0.781
1000 genes	0.826	0.787
Whole genome	0.831	0.800

## 6.2. Performance on Functional Terms

We assessed the ability of our approach to predict individual GO terms when fusing whole genome data from Fig. 2a. Table 2 shows the  $F_1$  scores for nine selected GO terms that belong to “Biological Process” and “Molecular Function” categories from GO and which contain variable number of annotated genes. These GO terms are of high relevance in *Dictyostelium* community and were selected upon consultations. Predictions were examined in the context of a complete set of GO terms rather than using a generic slim subset of terms. The resulting data set had  $\sim 2000$  GO terms, each had on average 9.64 direct gene annotations.

Our approach achieved consistently higher accuracy than the kernel-based approach. With the exception of “actin binding” and “lysozyme activity” terms,  $F_1$  scores are rather high. We also found that prediction of less specific terms such as “chemotaxis” and “response to bacterium” showed high performance. That was not expected because genes annotated with less specific terms tend to have their profiles in data sets less similar. High performance is important as all nine gene functions and processes are of interest in the current research of *D. discoideum* where data fusion may propose new candidate genes for down-stream experimental studies.

## 6.3. Ribosomal and Membrane Protein Classification

Table 3 shows the results of training a factorization-based fusion model and a kernel-based method to recognize membrane and cytoplasmic ribosomal proteins in yeast. Our approach yielded better accuracy than kernel-based method on the membrane proteins but worse on the cytoplasmic ribosomal class. However, fused data sources were those whose kernels gave best individual performance in kernel learning.<sup>18</sup> Thus, the selection of data sources was

Table 2: Gene ontology term-specific cross-validated  $F_1$  scores for fusion by matrix factorization (MF) and kernel-based method (MKL). Terms in Gene ontology belong to one of three categories, “Biological Process” (BP), “Molecular Function” (MF<sub>n</sub>) or “Cellular Component”.

GO term name	Term identifier	Namespace	Size	MF	MKL
Activation of adenylate cyclase activity	0007190	BP	11	0.834	0.770
Chemotaxis	0006935	BP	58	0.981	0.794
Chemotaxis to cAM	0043327	BP	21	0.922	0.835
Phagocytosis	0006909	BP	33	0.956	0.892
Response to bacterium	0009617	BP	51	0.899	0.788
Cell-cell adhesion	0016337	BP	14	0.883	0.867
Actin binding	0003779	MF <sub>n</sub>	43	0.676	0.664
Lysozyme activity	0003796	MF <sub>n</sub>	4	0.782	0.774
Sequence-specific DNA binding TFA	0003700	MF <sub>n</sub>	79	0.956	0.894

biased toward kernel-based method. The approach using factorization circumvents tedious work of transforming different objects (e.g., strings, vectors, graphs) into kernel matrices. These transformations depend on the choice of the kernels and may affect MKL’s performance.

Results in this and previous sections suggest that factorization-based data fusion might be useful not only to identify proteins that share the same molecular function but also to recognize proteins that participate in the same biological processes or are located in the same subcellular region.

Table 3: Cross-validated  $F_1$  scores for yeast membrane and cytoplasmic ribosomal proteins using matrix factorization-based fusion (MF) and kernel-based method (MKL).

Yeast recognition task	MF	MKL
Membrane proteins	0.843	0.835
Ribosomal proteins	0.901	0.921

## 7. Conclusion

We have examined the applicability of our recently proposed matrix factorization-based data fusion approach<sup>1</sup> on the problem of gene function prediction. We studied three fusion scenarios to demonstrate high accuracy of our approach when learning from disparate, incomplete and noisy data. The studies were successfully carried out for two different organisms, where, for example, the protein-protein interaction network for yeast is nearly complete but it is noisy, whereas the sets of available interactions for slime mold are rather sparse and only about one-tenth of its genes have experimentally derived annotations.

Our approach can model any number of data sources that can be expressed in a matrix, and, unlike most current data fusion approaches, does not require transformation of data into gene-function space. This flexibility allows us to fuse the data derived from possibly very diverse data sources without substantial preprocessing and loss of information. Described method is applicable to problems such as prediction of regulatory, metabolic and other functional classes, prediction of protein subcellular location and their interactions.

## Acknowledgements

We thank Gad Shaulsky from Baylor College of Medicine, Houston, TX, for selecting functional terms from Table 2. This work was supported by the Slovenian Research Agency (P2-0209, J2-9699, L2-1112), National Institute of Health (P01-HD39691) and European Commission (Health-F5-2010-242038).

## References

1. M. Žitnik and B. Zupan, (*submitted*) Available at *Arxiv:1307.0803*. (2013).
2. M. Žitnik and B. Zupan, Matrix factorization-based data fusion for drug-induced liver injury prediction, in *Proc. of the 12th Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), ISMB/ECCB*, 2013.
3. Y. Chen and D. Xu, *Nucleic Acids Research* **32**, 6414 (2004).
4. H. Wu, Z. Su, F. Mao, V. Olman and Y. Xu, *Nucleic Acids Research* **33**, 2822 (2005).
5. S. Mostafavi and Q. Morris, *Proteomics* **12**, 1687 (2012).
6. P. Radivojac *et al.*, *Nature Methods* **10**, 221 (2013).
7. 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 *et al.*, *Nature Genetics* **25**, 25 (2000).
8. M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, E. Cilia, R. Velasco and P. Fontana, *BMC Bioinformatics* **13 Suppl 4**, p. S14 (2012).
9. A. Vinayagam, R. König, J. Moormann, F. Schubert, R. Eils, K.-H. Glatting and S. Suhai, *BMC Bioinformatics* **5**, p. 116 (2004).
10. Z. Barutcuoglu, R. E. Schapire and O. G. Troyanskaya, *Bioinformatics* **22**, 830 (2006).
11. H. Yan, K. Venkatesan, J. E. Beaver, N. Klitgord, M. A. Yildirim, T. Hao, D. E. Hill, M. E. Cusick, N. Perrimon, F. P. Roth and M. Vidal, *PLoS One* **5**, p. e12139 (2010).
12. J. Jung, G. Yi, S. Sukno and M. Thon, *BMC Bioinformatics* **11**, p. 215 (2010).
13. N. Mitsakakis, Z. Razak, M. Escobar and J. T. Westwood, *BioData Mining* **6**, p. 8 (2013).
14. Ö. S. Saraç, V. Atalay and R. Cetin-Atalay, *PLoS One* **5**, p. e12382 (2010).
15. O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman and D. Botstein, *Proceedings of the National Academy of Sciences* **100**, 8348 (2003).
16. S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios and Q. Morris, *Genome Biology* **9**, p. S4 (2008).
17. S. Mostafavi and Q. Morris, *Bioinformatics* **26**, 1759 (2010).
18. G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan and W. S. Noble, *Bioinformatics* **20**, 2626 (2004).
19. S. Yu, T. Falck, A. Daemen, L.-C. Tranchevent, J. A. Suykens, B. De Moor and Y. Moreau, *BMC Bioinformatics* **11**, p. 309 (2010).
20. L. Peña-Castillo, M. Tasan, C. L. Myers, H. Lee, T. Joshi, C. Zhang, Y. Guan, M. Leone, A. Pagnani, W. K. Kim *et al.*, *Genome Biology* **9**, p. S2 (2008).
21. W. T. Clark and P. Radivojac, *Bioinformatics* **29**, i53 (2013).
22. A. Parikh, E. R. Miranda, M. Katoh-Kurasawa, D. Fuller, G. Rot, L. Zagar, T. Curk, R. Sugang, R. Chen, B. Zupan, W. F. Loomis, A. Kuspa and G. Shaulsky, *Genome Biology* **11**, p. R35 (2010).

## APPLICATIONS OF BIOINFORMATICS TO NON-CODING RNAs IN THE ERA OF NEXT-GENERATION SEQUENCING

CHAO CHENG

*Department of Genetics, Institute for Quantitative Biomedical Sciences,  
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College  
Hanover, NH 03755, USA  
Email: chao.cheng@dartmouth.edu*

JASON MOORE

*Department of Genetics, Institute for Quantitative Biomedical Sciences,  
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College  
Hanover, NH 03755, USA  
Email: jason.moore@dartmouth.edu*

CASEY GREENE

*Department of Genetics, Institute for Quantitative Biomedical Sciences,  
Norris Cotton Cancer Center, Geisel School of Medicine, Dartmouth College  
Hanover, NH 03755, USA  
Email: casey.greene@dartmouth.edu*

The human genome encodes a large number of non-coding RNAs, which employ a new and crucial layer of biological regulation in addition to proteins. Technical advancement in recent years, particularly, the wide application of next generation sequencing analysis, provide an unprecedented opportunity to identify new non-coding RNAs and investigate their functions and regulatory mechanisms. The aim of this workshop is to bring together experimental and computational biologist to exchange ideas on non-coding RNA studies.

### 1. Background

Non-coding RNAs (ncRNAs) are RNA molecules encoded by genes in the genome that are transcribed and functional but not translated into proteins. Recent studies have shown that more than 90% human genome is transcribed but coding sequences occupy only a small fraction of the genome (<2%) [1]. This suggests the existence of a large number of non-coding RNAs [2]. In fact, the FANTOM3 (Functional Annotation of Mammalian cDNA) project has identified ~35,000 non-coding transcripts with similar processing as mRNAs, including 5' capping, splicing, and poly-adenylation, but with little or no open reading frame (ORF) [3]. Given the large number of non-coding RNAs, it is reasonable to assume that these molecules are critical players in biological processes. At present, we are just starting to understand the functions of non-coding RNA.

#### 1.1. Classifications of non-coding RNAs

Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, and piRNAs among other types.

Based on the size of the mature version of non-coding RNAs, we can divide them into long non-coding RNAs (lncRNAs) and small non-coding RNAs. The cutoff value for size is arbitrarily determined with non-coding RNAs longer than 200 nucleotides categorized as lncRNAs and the rest as small. Compared to the small non-coding RNAs, existing knowledge about lncRNAs is even more limited.

According to their genomic locations, lncRNAs can be grouped into stand-alone lncRNAs, natural antisense transcripts, long intronic RNAs, transcribed pseudogenes and other lncRNAs (e.g. promoter associated RNAs, enhancer RNAs). Importantly, stand-alone lncRNAs are transcription units that do not overlap protein-coding genes. Some of these are referred to as lincRNAs (large intergenic noncoding RNAs). A recent study indicates that the human genome produce tens of thousands of lincRNAs [4].

## **1.2. *Functions of non-coding RNAs***

The functions of certain non-coding RNA types such as microRNAs have been intensively studied under a variety of biological contexts. However the functions of most of the lncRNAs including lincRNAs remain elusive or unclear. Despite of this, the functionality of lncRNAs is suggested by (1) the conservation of their promoters, splice junctions, exons, predicted structures, genomic; (2) their association with particular chromatin signatures that are indicative of active transcription; (3) their regulation by key molecular signals and transcription factors; (4) their dynamic expression and alternative splicing during differentiation; (5) their tissue- and cell-specific expression patterns and subcellular localization; (6) their altered expression or splicing patterns in cancer and other diseases [5].

In fact, lncRNAs are known to be able to exert regulatory functions at the transcriptional, post-transcriptional and epigenetic levels by different mechanisms. At the transcriptional level lncRNAs target transcriptional activators or repressors, different components of the transcription reaction including RNA polymerase II and the DNA duplex to regulate gene transcription and expression [6]. At the post-transcriptional level they participate in pre-mRNA processing, splicing, transport, translation, and degradation. At the epigenetic level they are involved in gene imprinting, X-chromosome inactivation and many other biological processes.

Several regulatory mechanisms of lncRNAs have been elucidated [7]. First, some lncRNAs can serve as decoys to prevent regulatory proteins from binding to DNA. For example, the lncRNA Gas5 contains a hairpin sequence motif in its secondary structure that resembles the DNA-binding site of the glucocorticoid receptor (GR) and decoy GR to inhibit the transcription of its target genes [8]. Second, some lncRNAs can serve as adaptors to bring two or more proteins into complexes. Third, some lncRNAs are required for guiding the proper localization of specific protein complexes; Finally, some lncRNAs can compete with miRNAs for miRNA-binding sites or serve as “sponges” to sequester miRNAs away from their mRNA targets [9].

## 2. Major directions and challenges

The goal of this workshop is to encourage the development of advanced methods for identification and functional characterization of ncRNAs through a combination of experimental and bioinformatics approaches.

### 2.1. *Application of bioinformatics to studies of non-coding RNAs*

Computational and bioinformatics techniques have been applied to study non-coding RNA mainly in the following directions: (1) prediction and identification of new non-coding RNAs from genome sequence analysis or by combining computational analysis with experimental data (e.g. tiling array, RNA-seq data); (2) prediction of miRNA target genes; (3) prediction the secondary and tertiary structures of RNAs; (4) investigation on the conservation and evolution of non-coding RNA genes or miRNA target genes; (5) non-coding RNA function prediction by computational analysis such as “guilt by association”; (6) construction of integrated regulatory networks that include non-coding RNA regulatory layers; (7) construction of databases and webserver to facilitate non-coding RNA studies.

### 2.2. *Main challenges in computational analysis*

Compared to protein studies, application of computational methods to non-coding RNA field is still in its infancy. There are several challenges that limit its application. First, non-coding RNAs represent heterogeneous classes of molecules; each has their specific characteristics and regulatory mechanisms. Second, non-coding RNA genes are non-conserved or less than conserved than protein coding genes; many of them have low expression levels and no obvious knockout phenotypes. Third, the knowledge about non-coding RNAs is still limited, can consequently there is no training data large enough for implementing machine learning techniques or statistical models. Fourth, the quality of non-coding RNAs gene annotation is relatively low. With the technical advancement and accumulation of data, we expect these challenges would be overcome in a short future.

### 2.3. *Main topics of this workshop*

#### 2.3.1 *Identification, annotation, classification and the evolution of lncRNAs.*

Computational and experimental methods have been proposed to annotate lincRNAs with special consideration to their lower expression profile. Phylogenetic analysis of lincRNAs in mammalian has demonstrated an interesting evolutionary history of them.

#### 2.3.2 *Prediction RNA Secondary Structure*

Secondary structure is highly important to the correct processing and function of many non-coding RNAs. Many computational methods have been proposed for modeling and understanding RNA structure.

### 2.3.3 Expression analysis of lncRNAs

To gain insight into the potential cellular functions of lncRNAs, systematic gene expression profiling has been performed by RNA-seq or tiling array. In particular, disease associated lncRNAs have been predicted by integrative analysis.

### 2.3.4 Complexity of RNA regulatory mechanism

The regulatory mechanism of non-coding RNAs is very diverse and complicated. With the advancement of non-coding RNA studies, we would expect the discovery of more regulatory mechanisms.

## 3. Workshop contributions

The workshop includes six invited speakers.

**Dr. Runsheng Chen** is a Professor in the Institute of Biophysics of Chinese Academy of Sciences. He is a member of Chinese Academy of Sciences. His research focuses on the identification of non-coding RNA genes in multiple organisms, function prediction and annotation of long non-coding RNAs, and the construction of non-coding RNA annotation databases. His lab has developed computational methods and tools for predicting, annotating and classifying non-coding RNAs.

**Dr. Yiwen Chen** received his PhD in physics from the University of North Carolina at Chapel Hill and is currently a Postdoctoral Fellow in Dr. Shirley Liu's Lab at the Dana-Farber Cancer Institute at Harvard School of Public Health. Dr. Liu's research focuses on developing bioinformatics methods and tools for analyzing high throughput data, using the dynamics of histone mark ChIP-seq and DNase-seq to infer in vivo transcription factor binding and regulation, employing genome wide approaches to understand the specificity and mechanism of epigenetic enzymes and lncRNAs, as well as integrating publicly available high throughput data to better understand cancer mechanisms.

**Dr. David Corey** is a Professor in the Department of Pharmacology at University of Texas Southwestern Medical Center. He received his PhD in Chemistry from the University of California, Berkeley. His research focuses on the mechanism of promoter-targeted antigene RNAs, the function of Argonaute and small RNA-dependent pathways in mammalian cell nuclei, the allele-selective inhibition of Huntington protein expression as well as the recognition of RNA and DNA by chemically modified nucleic acids and locked nucleic acids.

**Dr. Manuel Garber** is an Associate Professor in the Program in Bioinformatics and Integrative Biology, and the Director of the Bioinformatics core at University of Massachusetts Medical School. He received his PhD in Mathematics from Brandeis University. His research focuses on the evolutionary history of non-coding genes as well as the systematic dissection of the transcriptional regulation of the immune response. His lab has also been developing the tools to analyze, integrate and fully leverage the advancements in genome wide experimental technologies.

**Dr. John Hogenesch** is an Associate Professor of Pharmacology and the Associate Director of the Penn Genome Frontiers Institute at the University of Pennsylvania. He

received his PhD in Neuroscience from Northwestern University. His research focuses on the study of the mammalian circadian clock using genomic and computational tools. His lab has a longstanding interest in understanding noncoding RNA function through global gene expression analysis, and functional screening to gain insight into the potential cellular functions of lincRNAs and microRNAs.

**Dr. David Mathews** is an Associate Professor in the Department of Biochemistry and Biophysics at University of Rochester Medical Center. He received his PhD in Chemistry and MD in Medicine from University of Rochester. His research focuses on predicting RNA structure and developing computational tools for targeting RNA with pharmaceuticals and for using RNA as a pharmaceutical. His lab has developed software for predicting secondary structure of RNAs, software for predicting base pairing probabilities using a partition function and methods for predicting a secondary structure common to multiple sequences.

#### 4. Acknowledgements

We would like to thank all of the speakers for kindly accepting our invitation and generously supporting our workshop, as well as the PSB organizers for their assistance arranging this workshop and providing a venue for these discussions. J.M was supported by the NIH grants LM009012 and LM010098. C.C. and C.G. were supported by the NIH COBRE (Center of Biomedical Research Excellence) grant GM103534.

#### References

1. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57-74.
2. Kapranov P, Cheng J, Dike S, Nix DA, Dutttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484-1488.
3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**:1559-1563.
4. Hangauer MJ, Vaughn IW, McManus MT: **Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs.** *PLoS Genet* 2013, **9**:e1003569.
5. Mattick JS: **The genetic signatures of noncoding RNAs.** *PLoS Genet* 2009, **5**:e1000459.
6. Goodrich JA, Kugel JF: **Non-coding-RNA regulators of RNA polymerase II transcription.** *Nat Rev Mol Cell Biol* 2006, **7**:612-616.
7. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81**:145-166.
8. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP: **Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.** *Sci Signal* 2010, **3**:ra8.
9. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J: **Natural RNA circles function as efficient microRNA sponges.** *Nature* 2013, **495**:384-388.



## BUILDING THE NEXT GENERATION OF QUANTITATIVE BIOLOGISTS

KRISTINE A. PATTIN

*Institute for Quantitative Biomedical Sciences, Dartmouth College  
Hanover, NH 03755, USA  
Email: Kristine.A.Pattin@Dartmouth.edu*

ANNA C. GREENE

*Institute for Quantitative Biomedical Sciences, Dartmouth College  
Hanover, NH 03755, USA  
Email: Anna.C.Greene@Dartmouth.edu*

RUSS B. ALTMAN

*Department of Genetics, Stanford University  
Stanford, CA 94305, USA  
Email: russ.altman@stanford.edu*

LAWRENCE E. HUNTER

*Computational Bioscience Program, University of Colorado School of Medicine  
Aurora, CO 80045, USA  
Email: larry.hunter@ucdenver.edu*

DAVID A. ROSS

*Celera  
Alameda, CA 94502, USA  
Email: David.Ross@celera.com*

JAMES A. FOSTER

*Department of Biological Sciences, University of Idaho  
Moscow, ID 83844, USA  
Email: foster@uidaho.edu*

JASON H. MOORE

*Institute for Quantitative Biomedical Sciences, Dartmouth College  
Hanover, NH 03755, USA  
Email: Jason.H.Moore@Dartmouth.edu*

Many colleges and universities across the globe now offer bachelors, masters, and doctoral degrees, along with certificate programs in bioinformatics. While there is some consensus surrounding curricula competencies, programs vary greatly in their core foci, with some leaning heavily toward the biological sciences and others toward quantitative areas. This allows prospective students to choose a program that best fits their interests and career goals. In the digital age, most scientific fields are facing an enormous growth of data, and as a consequence, the goals and challenges of bioinformatics are rapidly changing; this requires that bioinformatics education also change. In this workshop, we seek to ascertain current trends in bioinformatics education by asking the question, “What are the core competencies all bioinformaticians should have at the end of their training, and how successful have programs been in placing students in desired careers?”

## 1. Background

Bioinformatics is an intrinsically multidisciplinary field, which makes educating students across this educational continuum difficult. Therefore bioinformatics education has been dubbed an NP-hard problem.<sup>1</sup> In 1998, Professor Russ Altman led the charge to formalize bioinformatics education at the graduate level. He specified five competency areas for bioinformatics training: biology, computer science, statistics, ethics, and core bioinformatics yet cautioned against defining the curriculum too narrowly.<sup>2</sup> Previous workshops, RECOMB-BE and ISMB-WEB, have discussed bioinformatics education at the graduate and undergraduate levels<sup>1</sup>, and the AMIA has suggested a list of biomedical informatics core competencies.<sup>3</sup> Defining bioinformatics curricula is a global effort.<sup>4-8</sup>

### 1.1. Challenges

There are many challenges associated with bioinformatics education, including faculty/instructor knowledge, computing resources, and breadth of knowledge required for bioinformatics training.<sup>9</sup> One of the major challenges graduate programs face is a lack of widespread adoption of these courses at the high school and undergraduate levels. Professors Ned Wingreen and David Botstein state, “The problem begins early in undergraduate education, and by the doctoral level there are severe interdisciplinary communication difficulties that are encountered by even the most motivated of collaborators.”<sup>10</sup> Many have examined tactics to bring bioinformatics into the high school classroom in order to make the connection between biology and computation.<sup>11-13</sup> Others have incorporated bioinformatics into biology, chemistry, and computer science undergraduate courses.<sup>14-23</sup> Summer programs have been shown to be efficacious in helping students step into bioinformatics careers.<sup>24</sup> Strategies to mitigate problems surrounding this educational divide are paramount to having an efficacious program that trains successful students.

Another challenge for graduate training programs is the depth of multidisciplinary understanding needed to produce bioinformatics scientists instead of technicians.<sup>9, 25</sup> Ranganathan et al. have proposed a “minimum skillset” for bioinformaticians<sup>26</sup>, and we would like to further address this to ask, “What are the key concepts and skills that one must graduate with in order to be successful today?” The modern biomedical researcher must be able to speak more than one language to successfully collaborate in a highly interdisciplinary environment; therefore it is beneficial to train a new generation of researchers who are well versed in the quantitative biomedical sciences and thus crucial to understand how such training programs succeed or fail. It would be valuable to faculty and directors of current programs or for those interested in bringing multidisciplinary programs to their institution to learn approaches and strategies for program development from existing and nascent programs. A few of the presentation and discussion topics include: curriculum design, effectively integrating multiple disciplines into a training program, getting first year students up to speed, the impact of a bioinformatics training program on faculty research, effectiveness (what works & what does not?), online curricula, student success, and the skills and tools students should have at program completion.

## 2. Workshop Contributions

**Russ B. Altman** is the Director of the Biomedical Informatics Training Program (BMI) at Stanford University. BMI is an interdisciplinary program that focuses on learning to develop and apply quantitative and computational methods to various biological and medical problems. BMI students represent a spectrum of educational backgrounds such as biology, research and clinical medicine, computer science, statistics, engineering, and a number of other fields. The faculty members come from a broad range of departments, including Bioengineering, Computer Science, Genetics, Medicine, Pediatrics, Radiology, and Statistics to provide research training and coursework in a number of related fields.

Ph.D. candidates are required to take core BMI classes, electives in computer science, statistics, mathematics, engineering, and allied informatics-related disciplines, coursework in social, legal, and ethical issues, and have access to unrestricted electives. BMI also aims to ensure students develop the ability to communicate their ideas and research effectively by having requirements and opportunities to present in journal clubs, colloquia, and lab meetings. BMI has successfully produced over 100 graduates who have pursued a number of different careers and can be found as distinguished faculty at top universities and medical schools as well as industry leaders at major corporations and startups.

**James A. Foster** is a Professor of Biological Sciences and Founding Member of the Institute for Bioinformatics and Evolutionary Studies (IBEST) and the Bioinformatics and Computational Biology (BCB) graduate program at the University of Idaho. Students in the BCB graduate program participate in research that encompasses the disciplines of computer science, biology, mathematics, and statistics and are affiliated with IBEST. BCB Ph.D. students may choose to focus their training in either the computer/mathematical sciences or the biological sciences.

Faculty members include physicists, chemists, molecular biologists, organismal biologists, ecologists, behavioral biologists, mathematicians, statisticians, and computer scientists. With access to such resources and technology in numerous scientific disciplines, BCB graduate students can no doubt become versatile members of the scientific community in whatever career path they decide to follow.

**Lawrence E. Hunter** is the Director of the Center for Computational Pharmacology and the Computational Bioscience Program at the University of Colorado Denver. The Computational Bioscience Ph.D. program curriculum is designed to integrate training in both computation and biomedical sciences with a student's research and teaching activities. Therefore, graduate students are expected to emerge from this program as independent researchers with a solid foundation in computational methods and molecular biomedicine, the science and technology comprising the two, as well as the skills to collaborate, communicate, and develop computational approaches that can be applied to a wide variety of biological problems.

Faculty represent a breadth of scientific research with multiple appointments in the departments of Medicine, Pharmacology, Biometrics, Biochemistry & Molecular Genetics, Computer Science, as well as from National Jewish Health. Four key competencies: knowledge, communication, professionalism, and life-long learning skills structure the core of the program's teaching philosophy. The Computational Bioscience Program recognizes that bioinformatics is a

rapidly evolving field, and while these core objectives will remain steadfast, they are committed to continually reviewing, revising, and improving their curriculum to keep pace with this evolution.

**Jason H. Moore** is the Director of the Institute for Quantitative Biomedical Sciences (iQBS) at Dartmouth College. IQBS is based on the idea that biomedical research studies move through a series of related activities that require a specific skillset and a specific scientific language. In any given study, there are specialists who design the experiment, who collect and organize the data, who analyze the data and who interpret the data. The major areas of expertise are bioinformatics, biostatistics and epidemiology. While there is some overlap, there is no one discipline that incorporates understanding across the entire process.

To cover these three disciplines both in student research and coursework, the program is interdepartmental with faculty from the Departments of Biological Sciences, Community and Family Medicine, Computer Science, Genetics and Medicine at Dartmouth College and the Geisel School of Medicine. Numerous collaborations exist between QBS members and those in other Ph.D. programs at Dartmouth including the Molecular and Cellular Biology Program, the Program in Experimental and Molecular Medicine and the Graduate Programs at The Dartmouth Institute for Health Policy and Clinical Practice.

The overarching goal of QBS is to prepare students for productive careers as quantitative scientists in the biomedical sciences by cross-training Ph.D. students and providing in-depth collaborative experiences in specific applications. Successful students will be able to effectively lead biomedical research studies from start to finish or participate as interdisciplinary members of large collaborative groups. QBS is an innovative approach to graduate training that combines multiple disciplines to train the next generation of collaborative scientists.

**David A. Ross** is the Director of Computational Biology at Celera. Celera is well known for its original mission to sequence the human genome and subsequently provide clients with early access to this data. Since then Celera has made important contributions to scientific research by developing “shotgun” sequencing and commencing the Applera Genomics Initiative, an effort that identified over 40,000 novel SNPs which became the foundation for new genetic tests that Celera is developing. Today, they provide a number of services including diagnostic products used for personalized disease management. Celera represents the interdisciplinary nature of biotechnology, where a student with interdisciplinary training would be an asset in many ways. Dr. Ross’ perspective on the bioinformatics training needed to be successful in an industry-based career will be welcomed.

## References

1. S. Ranganathan, PLoS Computational Biology **1**, 6, 447–448 (2005).
2. R.B. Altman, Bioinformatics **14**, 7, 549–550 (1998).
3. C.A. Kulikowski, E.H. Shortliffe, L.M. Currie, P.L. Elkin, L.E. Hunter, T.R. Johnson, I.J. Kalet, L.A. Lenert, M.A. Musen, J.G. Ozbolt, J.W. Smith, P.Z. Tarczy-Hornoch, and J.J. Williamson, Journal of the American Medical Informatics Association : JAMIA **19**, 6, 931–8 (2012).
4. I. Koch and G. Fuellen, Briefings in Bioinformatics **9**, 3, 232–42 (2008).
5. D. Counsell, Briefings in Bioinformatics **4**, 1, 7–21 (2003).

6. M.S. Shamsir and Z.A.M. Hussein, AJTLHE: ASEAN Journal of Teaching and Learning in Higher Education **2**, 1, 30–40 (2010).
7. M. Gerstein, D. Greenbaum, K. Cheung, and P.L. Miller, Journal of Biomedical Informatics **40**, 1, 73–9 (2007).
8. R.B. Altman and T.E. Klein, Journal of Biomedical Informatics **40**, 1, 55–8 (2007).
9. M.P. Cummings and G.G. Temple, Briefings in Bioinformatics **11**, 6, 537–43 (2010).
10. N. Wingreen and D. Botstein, Nature Reviews. Molecular Cell Biology **7**, 11, 829–32 (2006).
11. S.R. Gallagher, W. Coon, K. Donley, A. Scott, and D.S. Goldberg, PLoS Computational Biology **7**, 10, e1002244 (2011).
12. S.H. Wefer and K. Sheppard, CBE Life Sciences Education **7**, 1, 155–62 (2008).
13. J. McQueen, J.J. Wright, and J.A. Fox, PLoS Computational Biology **8**, 8, e1002636 (2012).
14. A.E. Bednarski, S.C.R. Elgin, and H.B. Pakrasi, Cell Biology Education **4**, 3, 207–20 (2005).
15. N.B. Centeno, J. Villà-Freixa, and B. Oliva, Biochemistry and Molecular Biology Education **31**, 6, 386–391 (2003).
16. S. Cooper, Biochemistry and Molecular Biology Education **29**, 4, 167–168 (2001).
17. A.L. Feig and E. Jabri, Biochemistry and Molecular Biology Education **30**, 4, 224–231 (2002).
18. J.E. Honts, Cell Biology Education **2**, 4, 233–47 (2003).
19. B.S. Chapman, J.L. Christmann, and E.F. Thatcher, Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology **34**, 3, 180–6 (2006).
20. J.L. Ditty, C.A. Kvaal, B. Goodner, S.K. Freyermuth, C. Bailey, R.A. Britton, S.G. Gordon, S. Heinhorst, K. Reed, Z. Xu, E.R. Sanders-Lorenz, S. Axen, E. Kim, M. Johns, K. Scott, and C.A. Kerfeld, PLoS Biology **8**, 8, e1000448 (2010).
21. J.A. Boyle, Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology **32**, 4, 236–8 (2004).
22. D. Weisman, Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology **38**, 1, 4–9 (2010).
23. L.L. Furge, R. Stevens-Truss, D.B. Moore, and J.A. Langeland, Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology **37**, 1, 26–36 (2009).
24. B. Krilowicz, W. Johnston, S.B. Sharp, N. Warter-Perez, and J. Momand, CBE Life Sciences Education **6**, 1, 74–83 (2007).
25. P.A. Pevner, Bioinformatics **20**, 14, 2159–2161 (2004).
26. T.W. Tan, S.J. Lim, A.M. Khan, and S. Ranganathan, BMC Genomics **10**, Suppl 3, S36 (2009).

## **UNCOVERING THE ETIOLOGY OF AUTISM SPECTRUM DISORDERS: GENOMICS, BIOINFORMATICS, ENVIRONMENT, DATA COLLECTION AND EXPLORATION, AND FUTURE POSSIBILITIES**

SARAH PENDERGRASS

*Center for Systems Genomics, Department of Biochemistry & Molecular Biology, The Pennsylvania State University,  
University Park, PA, 16802, USA  
Email: [sap29@psu.edu](mailto:sap29@psu.edu)*

SANTHOSH GIRIRAJAN

*Center for Systems Genomics, Department of Biochemistry & Molecular Biology, The Pennsylvania State University,  
University Park, PA, 16802, USA  
Email: [sxg47@psu.edu](mailto:sxg47@psu.edu)*

SCOTT SELLECK

*Department of Biochemistry & Molecular Biology, 206D Life Sciences Building,  
University Park, PA 16802, USA  
Email: [sbs24@psu.edu](mailto:sbs24@psu.edu)*

A clear and predictive understanding of the etiology of autism spectrum disorders (ASD), a group of neurodevelopmental disorders characterized by varying deficits in social interaction and communication as well as repetitive behaviors, has not yet been achieved. There remains active debate about the origins of autism, and the degree to which genetic and environmental factors, and their interplay, produce the range and heterogeneity of cognitive, developmental, and behavioral features seen in children carrying a diagnosis of ASD. Unlocking the causes of these complex developmental disorders will require a collaboration of experts in many disciplines, including clinicians, environmental exposure experts, bioinformaticists, geneticists, and computer scientists. For this workshop we invited prominent researchers in the field of autism, covering a range of topics from genetic and environmental research to ethical considerations. The goal of this workshop: provide an introduction to the current state of autism research, highlighting the potential for multi-disciplinary collaborations that rigorously evaluate the many potential contributors to ASD. It is further anticipated that approaches that successfully advance the understanding of ASD can be applied to the study of other common, complex disorders. Herein we provide a short review of ASD and the work of the invited speakers.

## 1. Autism Spectrum Disorders a Brief Introduction

Autism spectrum disorders (ASD) are characterized by a range of clinical features that can vary from individual to individual in both the degree of severity and variability of the clinical presentation. This can include abnormalities in language, reciprocal social interactions, and/or other communication skills as well as repetitive behaviors<sup>1</sup>. Autism spectrum disorders are divided into three basic categories: autistic disorder (frequently referred to as autism), Asperger syndrome, and pervasive developmental disorder (PDD-NOS)<sup>1</sup>. These disorders, as of 2008, affect 1 in 88 children, and are more prevalent in males than females<sup>2</sup>. The prevalence estimates of ASD have increased, above increases due to changes in diagnostic criteria<sup>1,3</sup>. In addition, children with ASD often have intellectual disability, estimated as high as 68% of ASD cases<sup>4</sup>, and approximately 75% have lifelong disability requiring social/educational support<sup>5</sup>. The presence of ASD can have a significant impact on the quality of life of affected persons, but also for their family and/or other caregivers.

## 2. The Pacific Symposium on Biocomputing (PSB) ASD Workshop

Many studies have been investigating the connection between genetic variation and ASD. Twin studies have indicated that ASD are highly heritable<sup>6,7</sup>. Linkage studies have implicated a polygenic basis for autistic disorder<sup>8</sup>. However, genome-wide association studies (GWAS) for ASD have identified few potential loci associated with ASD<sup>9-11</sup>. Copy-number variation (CNV) studies, in contrast, have been more successful in identifying genomic regions associated with an increased risk for autism, and also other neurodevelopmental disabilities such as schizophrenia and epilepsy, with overlap of several genomic regions<sup>5,12,13</sup>. Copy number variations can be deletions, duplications, inversions, or translocations. While the location of CNVs may differ from individual to individual with ASD, these CNVs can still result in similar clinical features and outcomes<sup>12</sup>.

During the workshop, Dr. Santhosh Girirajan and Dr. Evan Eichler will describe work investigating the genetic and phenotypic heterogeneity of neurodevelopmental disorders in the context of CNVs, particularly for ASD<sup>12,14-16</sup>. Dr. Girirajan's research has been focused on the discovery of genetic variants associated with the causation, diagnosis, and biological interpretation of ASD. A recent manuscript by Girirajan et al. showed evidence that individuals with autism have higher numbers of larger copy-number variants, and that these are more duplication based instead of deletion events<sup>17</sup>.

Dr. Eichler will be speaking about the successful application of exome sequencing for children with ASD and their parents, as well as work determining copy-number variant (CNV) burden differences across neurodevelopmental phenotypes. Dr. Eichler is a leader in study of the relationship between CNVs and human disease and has focused his research on building an understanding of the evolution, pathology and mechanism(s) of recent gene duplication and DNA transposition within the human genome<sup>18</sup>. This research has included discovery of these important genomic regions, development of methods to assess their variation, detection of rapid gene evolution, and identifying the correlation between discovered genetic variation and phenotypic differences, including autism spectrum disorders.

Dr. Neale will describe the impact of high-throughput sequencing on ASD gene discovery, highlighting the contribution of rare variation to ASD as well as the pleiotropic effects of ASD associated mutations. His talk will also review challenges that remain in this field for detection and interpretation of inherited and de-novo rare-variants in ASD. Dr. Neale has conducted analyses for genetic data focused on psychiatric illness, particularly ADHD and autism, but also Tourette's obsessive compulsive disorder, schizophrenia<sup>19</sup> and eating disorders.

Investigation of environmental risk factors for ASDs is a growing research field<sup>19,20</sup>. The wide heterogeneity of ASD symptoms, and how to best ascertain individuals for study, are challenging. Different clinical features and the range of severity across individuals may stem from varying genetic contributors, but could also be due, in part, to variations in environmental exposures. Further, increasing rates of ASD indicate the potential for environmental exposure playing an important role in the etiology and/or heterogeneity of ASD. There is also a need for the exploration of gene-environment interactions<sup>20</sup>. Identifying both genetic variants concomitant with environmental exposure may provide important insights into the etiology of ASD.

Dr. Heather Volk will be speaking at the workshop, describing her work exploring the relationship between environmental exposure and the etiology of autism. Her research focuses on the environmental and genetic epidemiology of autism and other neurodevelopmental disorders and on gene-environment interactions in complex disease. Oxidative stress and inflammation may play a key role in ASD, with adverse prenatal effects. Dr. Volk will describe the impact of exposure to traffic-related air pollution on prenatal development and risk of ASD. In two recent reports by Volk et al., children with autism were more likely to have the highest exposure to traffic-related air pollution during gestation and the first year of life, compared to non-autistic controls<sup>21</sup> and maternal residence at time of delivery was more likely to be close to a freeway for autism cases vs. controls<sup>22</sup>.

Dr. Pessah will be describing work from the UC Davis Medical Investigation of Neurodevelopmental Disorders (MIND) Institute. Researchers at the MIND Institute are among the world's experts on molecular and environmental contributors to ASD as well as the use of epidemiological data for testing the cellular and molecular mechanisms of ASD. These researchers have established the most comprehensive database in the world of the environmental exposures of children with confirmed ASD or atypical development, linked to an extensive archive of clinical samples, and Dr. Pessah will describe interdisciplinary approaches that leverage this unique set of resources.

Epigenetic changes are another potential contributor to the etiology of ASD. Epigenetics is the study of heritable changes in chromosomes, not encoded in the DNA sequence, including DNA methylation and chromatin organization. DNA methylation is an important link between genetic and environmental interaction, as DNA hypomethylation is known to lead to genome instability. "Environmental epigenetics" explore this connection, identifying important environmental influences on epigenetic change<sup>23</sup>. For example, arsenic, cadmium, benzene and other exposures have been associated with DNA methylation in genes, as well as dietary factors<sup>23</sup>. The invited speaker, Dr. LaSalle of the MIND Institute, has performed pioneering studies on the epigenetic etiologies of ASD. The clinical applications of her research include understanding the pathogenesis of the neurodevelopmental disorders autism, Rett syndrome, Prader-Willi syndrome,



Dup15q syndrome, and Angelman syndrome, through identifying epigenetic pathways disrupted in rare genetic disorders on the autism spectrum. Dr. LaSalle's recent research is on environmental exposures affecting the DNA methylome and employing novel bioinformatics methods for analysis and visualization of epigenomic data relevant to autism.

An important part of any research is any potential ethical implications of those discoveries, this is true for ASD research as well. Because of the large numbers of individuals affected with autism, and the impact on children and families, as well as the potential for environmental exposure during pregnancy/youth to play a role, publications and press outside of the research-manuscript realm are more likely to report research results from studies of environmental exposures and ASD. A well known example is a paper published in 1998 in *The Lancet*<sup>24</sup>, later retracted, linking measles, mumps, and rubella (MMR) vaccine and autism. A review by the Institute of Medicine clearly showed no link between the thimerosal-containing vaccines after review of over 200 studies<sup>25</sup>. However the controversy that emerged over vaccines due to the initial publication has had a lasting impact on parents choosing to vaccinate children, and public health, even while autism is no more common among vaccinated than unvaccinated children<sup>26</sup>.

With the increasing amount of ASD research and the recent extension of research in different complex directions, there are a range of important ethical considerations when reporting the results of ASD studies to families, clinicians and the research community. Dr. Newschaffer is an epidemiologist and his currently involved in large risk factor epidemiology studies, autism phenotyping studies, genomic and epigenomic research, and studies focused on the utilization and evaluation of health care and behavioral intervention services<sup>1,27-29</sup>. He will discuss a number of issues including the uncertainty, comprehension, inadvertent harm, as well as appropriate roles of clinicians, scientists, and the media, in ASD communication.

### 3. Conclusions

While a complete understanding of ASD is still growing, through comprehensive and collaborative efforts we may begin to identify additional pieces of the ASD puzzle that can be linked with our existing current knowledge to grow a clearer picture of these disorders. The collaboration of multiple domain-experts will be required to effectively analyze the growing genetic and epidemiological data being collected. To foster these cross-disciplinary interactions and research projects, we have developed this workshop for PSB, to share the current knowledge of the genetic and environmental contributions to ASD and to highlight methods for future research in this field, including important ethical considerations. The intention is to grow new ideas, collaborations, and possibilities for future research in this field, between current autism spectrum researchers and other scientists in attendance at PSB. In addition to improving the understanding of the etiology of ASD, methodologies developed for the ASD field have the potential for expanding and improving the study of other common, complex disorders.

### References

1. Rossi, J., Newschaffer, C. & Yudell, M. Autism spectrum disorders, risk communication, and the problem of inadvertent harm. *Kennedy Inst. Ethics J.* **23**, 105–138 (2013).

2. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators & Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *Morb. Mortal. Wkly. Rep. Surveill. Summ. Wash. Dc* 2002 **61**, 1–19 (2012).
3. Weintraub, K. The prevalence puzzle: Autism counts. *Nature* **479**, 22–24 (2011).
4. Yeargin-Allsopp, M. *et al.* Prevalence of autism in a US metropolitan area. *Jama J. Am. Med. Assoc.* **289**, 49–55 (2003).
5. Deth, R. C. Genomics, intellectual disability, and autism. *N. Engl. J. Med.* **366**, 2231–2232; author reply 2232 (2012).
6. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095–1102 (2011).
7. Posthuma, D. & Polderman, T. J. C. What have we learned from recent twin studies about the etiology of neurodevelopmental disorders?: *Curr. Opin. Neurol.* **26**, 111–121 (2013).
8. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493–507 (1999).
9. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
10. Weiss, L. A., Arking, D. E., Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, M. J. & Chakravarti, A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* **461**, 802–808 (2009).
11. Anney, R. *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072–4082 (2010).
12. Coe, B. P., Girirajan, S. & Eichler, E. E. A genetic model for neurodevelopmental disease. *Curr. Opin. Neurobiol.* **22**, 829–836 (2012).
13. Devlin, B. & Scherer, S. W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).
14. Coe, B. P., Girirajan, S. & Eichler, E. E. The genetic variability and commonality of neurodevelopmental disease. *Am. J. Med. Genet. C Semin. Med. Genet.* **160C**, 118–129 (2012).
15. Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum. Mol. Genet.* **19**, R176–187 (2010).
16. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* **45**, 203–226 (2011).
17. Girirajan, S. *et al.* Global increases in both common and rare copy number load associated with autism. *Hum. Mol. Genet.* **22**, 2870–2880 (2013).
18. Mefford, H. C. & Eichler, E. E. Duplication hotspots, rare genomic disorders, and common disease. *Curr. Opin. Genet. Dev.* **19**, 196–204 (2009).
19. Pessah, I. N. *et al.* Immunologic and neurodevelopmental susceptibilities of autism. *Neurotoxicology* **29**, 532–545 (2008).
20. Chaste, P. & Leboyer, M. Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin. Neurosci.* **14**, 281–292 (2012).
21. Volk, H. E., Lurmann, F., Penfold, B., Hertz-Picciotto, I. & McConnell, R. Traffic-related air pollution, particulate matter, and autism. *Jama Psychiatry Chic. Ill* **70**, 71–77 (2013).
22. Volk, H. E., Hertz-Picciotto, I., Delwiche, L., Lurmann, F. & McConnell, R. Residential proximity to freeways and autism in the CHARGE study. *Environ. Health Perspect.* **119**, 873–877 (2011).
23. LaSalle, J. M. A genomic point-of-view on environmental factors influencing the human brain methylome. *Epigenetics Off. J. Dna Methylation Soc.* **6**, 862–869 (2011).
24. Retraction--Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* **375**, 445 (2010).
25. *Immunization Safety Review: Vaccines and Autism.* at <[http://www.nap.edu/catalog.php?record\\_id=10997](http://www.nap.edu/catalog.php?record_id=10997)>
26. Silencing debate over autism. *Nat. Neurosci.* **10**, 531–531 (2007).
27. Newschaffer, C. J. & Curran, L. K. Autism: an emerging public health problem. *Public Heal. Reports Wash. Dc* 1974 **118**, 393–399 (2003).
28. Newschaffer, C. J. *et al.* The epidemiology of autism spectrum disorders. *Annu. Rev. Public Health* **28**, 235–258 (2007).
29. Newschaffer, C. J., Fallin, D. & Lee, N. L. Heritable and nonheritable risk factors for autism spectrum disorders. *Epidemiol. Rev.* **24**, 137–153 (2002).