Pacific Symposium on Biocomputing 2015

Abstract Book

Poster Presenters: Poster space is assigned by abstract page number. Please find the page that your abstract is on and put your poster on the poster board with the corresponding number (e.g., if your abstract is on page 50, put your poster on board #50).

Proceedings papers with oral presentations #11-52 are not assigned poster space.

Papers are organized by session then the last name of the first author. Presenting authors' names are underlined.

TABLE OF CONTENTS

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

CELL INDEX DATABASE (CELLX): A WEB TOOL FOR CANCER PRECISION MEDICINE

Keith A. Ching[1], Kai Wang[1], Zhengyan Kan[1], Julio Fernandez[1], Wenyan Zhong[1], Jarek Kostrowicki[1], Tao Xie[1], Zhou Zhu[1], Jean-Francois Martini[2], Maria Koehler[2], Kim Arndt[1], Paul Rejto[1]

[1]Oncology Research Unit Pfizer Global Research & Development,
[2]Oncology Business Unit Pfizer Global Research & Development

The Cell Index Database, (CELLX) (http://cellx.sourceforge.net) provides a computational framework for integrating expression, copy number variation, mutation, compound activity, and meta data from cancer cells. CELLX provides the computational biologist a quick way to perform routine analyses as well as the means to rapidly integrate data for offline analysis. Data is accessible through a web interface which utilizes R to generate plots and perform clustering, correlations, and statistical tests for associations within and between data types for ~20,000 samples from TCGA, CCLE, Sanger, GSK, GEO, GTEx, and other public sources. We show how CELLX supports precision oncology through indications discovery, biomarker evaluation, and cell line screening analysis.

COMPARING NONPARAMETRIC BAYESIAN TREE PRIORS FOR CLONAL RECONSTRUCTION OF TUMORS

Amit G. Deshwar, Shankar Vembu, Quaid Morris

University of Toronto

Statistical machine learning methods, especially nonparametric Bayesian methods, have become increasingly popular to infer clonal population structure of tumors. Here we describe the treeCRP, an extension of the Chinese restaurant process (CRP), a popular construction used in nonparametric mixture models, to infer the phylogeny and genotype of major subclonal lineages represented in the population of cancer cells. We also propose new split-merge updates tailored to the subclonal reconstruction problem that improve the mixing time of Markov chains. In comparisons with the tree-structured stick breaking prior used in PhyloSub, we demonstrate superior mixing and running time using the treeCRP with our new split-merge procedures. We also show that given the same number of samples, TSSB and treeCRP have similar ability to recover the subclonal structure of a tumor.

STEPWISE GROUP SPARSE REGRESSION (SGSR): GENE-SET-BASED PHARMACOGENOMIC PREDICTIVE MODELS WITH STEPWISE SELECTION OF FUNCTIONAL PRIORS

In Sock Jang[1], Rodrigo Dienstmann[1], Adam A. Margolin[2], Justin Guinney[1]

[1]Sage Bionetworks, [2]Oregon Health & Science University

Complex mechanisms involving genomic aberrations in numerous proteins and pathways are believed to be a key cause of many diseases such as cancer. With recent advances in genomics, elucidating the molecular basis of cancer at a patient level is now feasible, and has led to personalized treatment strategies whereby a patient is treated according to his or her genomic profile. However, there is growing recognition that existing treatment modalities are overly simplistic, and do not fully account for the deep genomic complexity associated with sensitivity or resistance to cancer therapies. To overcome these limitations, large-scale pharmacogenomic screens of cancer cell lines – in conjunction with modern statistical learning approaches - have been used to explore the genetic underpinnings of drug response. While these analyses have demonstrated the ability to infer genetic predictors of compound sensitivity, to date most modeling approaches have been data-driven, i.e. they do not explicitly incorporate domain-specific knowledge (priors) in the process of learning a model. While a purely data-driven approach offers an unbiased perspective of the data – and may yield unexpected or novel insights - this strategy introduces challenges for both model interpretability and accuracy. In this study, we propose a novel prior-incorporated sparse regression model in which the choice of informative predictor sets is carried out by knowledge-driven priors (gene sets) in a stepwise fashion. Under regularization in a linear regression model, our algorithm is able to incorporate prior biological knowledge across the predictive variables thereby improving the interpretability of the final model with no loss – and often an improvement - in predictive performance. We evaluate the performance of our algorithm compared to well-known regularization methods such as LASSO, Ridge and Elastic net regression in the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (Sanger) pharmacogenomics datasets, demonstrating that incorporation of the biological priors selected by our model confers improved predictability and interpretability, despite much fewer predictors, over existing state-of-the-art methods.

INTEGRATIVE GENOME-WIDE ANALYSIS OF THE DETERMINANTS OF RNA SPLICING IN KIDNEY RENAL CLEAR CELL CARCINOMA

Kjong-Van Lehmann[1], Andre Kahles[1], Cyriac Kandoth[1],William Lee[1], Nikolaus Schultz[1], Oliver Stegle[2], Gunnar Rätsch[1]

[1]Memorial Kettering Cancer Center, [2]European Bioinformatics Institute

We present a genome-wide analysis of splicing patterns of 282 kidney renal clear cell carcinoma patients in which we integrate data from whole-exome sequencing of tumor and normal samples, RNA-seq and copy number variation. We proposed a scoring mechanism to compare splicing patterns in tumor samples to normal samples in order to rank and detect tumor-specific isoforms that have a potential for new biomarkers. We identified a subset of genes that show introns only observable in tumor but not in normal samples, ENCODE and GEUVADIS samples. In order to improve our understanding of the underlying genetic mechanisms of splicing variation we performed a large-scale association analysis to find links between somatic or germline variants with alternative splicing events. We identified 915 cis- and trans-splicing quantitative trait loci (sQTL) associated with changes in splicing patterns. Some of these sQTL have previously been associated with being susceptibility loci for cancer and other diseases. Our analysis also allowed us to identify the function of several COSMIC variants showing significant association with changes in alternative splicing. This demonstrates the potential significance of variants affecting alternative splicing events and yields insights into the mechanisms related to an array of disease phenotypes.

AN INTEGRATED FRAMEWORK FOR REPORTING CLINICALLY RELEVANT BIOMARKERS FROM PAIRED TUMOR/NORMAL GENOMIC AND TRANSCRIPTOMIC SEQUENCING DATA IN SUPPORT OF CLINICAL TRIALS IN PERSONALIZED MEDICINE

Sara Nasser, Ahmet A. Kurdolgu, Tyler Izatt, Jessica Aldrich, Megan L. Russell, Alexis Christoforides, Wiabhav Tembe, Jeffery A. Keifer, Jason J. Corneveaux, Sara A. Byron, Karen M. Forman, Clarice Zuccaro, Jonathan J. Keats, Patricia M. LoRusso, John D. Carpten, Jeffrey M. Trent, David W. Craig

Translational Genomics Research Institute, Yale School of Medicine, Barbara Ann Karmanos Cancer Institute

The ability to rapidly sequence the tumor and germline DNA of an individual holds the eventual promise of revolutionizing our ability to match targeted therapies to tumors harboring the associated genetic biomarkers. Analyzing high throughput genomic data consisting of millions of base pairs and discovering alterations in clinically actionable genes in a structured and real time manner is at the crux of personalized testing. This requires a computational architecture that can monitor and track a system within a regulated environment as terabytes of data are reduced to a small number of therapeutically relevant variants, delivered as a diagnostic laboratory developed test. These high complexity assays require data structures that enable real-time and retrospective ad-hoc analysis, with a capability of updating to keep up with the rapidly changing genomic and therapeutic options, all under a regulated environment that is relevant under both CMS and FDA depending on application. We describe a flexible computational framework that uses a paired tumor/normal sample allowing for complete analysis and reporting in approximately 24 hours, providing identification of single nucleotide changes, small insertions and deletions, chromosomal rearrangements, gene fusions and gene expression with positive predictive values over 90%. In this paper we present the challenges in integrating clinical, genomic and annotation databases to provide interpreted draft reports which we utilize within ongoing clinical research protocols. We demonstrate the need to retire from existing performance measurements of accuracy and specificity and measure metrics that are meaningful to a genomic diagnostic environment. This paper presents a three-tier infrastructure that is currently being used to analyze an individual genome and provide available therapeutic options via a clinical report. Our framework utilizes a non-relational variant-centric database that is scaleable to a large amount of data and addresses the challenges and limitations of a relational database system. Our system is continuously monitored via multiple trackers each catering differently to the diversity of users involved in this process. These trackers designed in analytics web-app framework provide status updates for an individual sample accurate to a few minutes. In this paper, we also present our outcome delivery process that is designed and delivered adhering to the standards defined by various regulation agencies involved in clinical genomic testing.

CHARACTERISTICS OF DRUG COMBINATION THERAPY IN ONCOLOGY BY ANALYZING CLINICAL TRIAL DATA ON CLINICALTRIALS.GOV

Menghua Wu[1,2], Marina Sirota[2], Atul J. Butte[2], Bin Chen[2]

[1]The Harker High School, [2]Stanford University School of Medicine

Within the past few decades, drug combination therapy has been intensively studied in oncology and other complex disease areas, especially during the early drug discovery stage, as drug combinations have the potential to improve treatment response, minimize development of resistance or minimize adverse events. In the present, designing combination trials relies mainly on clinical and empirical experience. While empirical experience has indeed crafted efficacious combination therapy clinical trials (combination trials), however, garnering experience with patients can take a lifetime. The preliminary step to eliminating this barrier of time, then, is to understand the current state of combination trials. Thus, we present the first large-scale study of clinical trials (2008-2013) from ClinicalTrials.gov to compare combination trials to non-combination trials, with a focus on oncology. In this work, we developed a classifier to identify combination trials and oncology trials through natural language processing techniques. After clustering trials, we categorized them based on selected characteristics and observed trends present. Among the characteristics studied were primary purpose, funding source, endpoint measurement, allocation, and trial phase. We observe a higher prevalence of combination therapy in oncology (25.6% use combination trials) in comparison to other disease trials (6.9%). However, surprisingly the prevalence of combinations does not increase over the years. In addition, the trials supported by the NIH are significantly more likely to use combinations of drugs than those supported by industry. Our preliminary study of current combination trials may facilitate future trial design and move more preclinical combination studies to the clinical trial stage.

CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE “BIG DATA” ERA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

IDENTIFYING MUTATION SPECIFIC CANCER PATHWAYS USING A STRUCTURALLY RESOLVED PROTEIN INTERACTION NETWORK

H. Billur Engin, Matan Hofree, Hannah Carter

University of California San Diego

Here we present a method for extracting candidate cancer pathways from tumor ‘omics data while explicitly accounting for diverse consequences of mutations for protein interactions. Disease-causing mutations are frequently observed at either core or interface residues mediating protein interactions. Mutations at core residues frequently destabilize protein structure while mutations at interface residues can specifically affect the binding energies of protein-protein interactions. As a result, mutations in a protein may result in distinct interaction profiles and thus have different phenotypic consequences. We describe a protein structure-guided pipeline for extracting interacting protein sets specific to a particular mutation. Of 59 cancer genes with 3D co-complexed structures in the Protein Data Bank, 43 showed evidence of mutations with different functional consequences. Literature survey reciprocated functional predictions specific to distinct mutations on APC, ATRX, BRCA1, CBL and HRAS. Our analysis suggests that accounting for mutation-specific perturbations to cancer pathways will be essential for personalized cancer therapy.

BINNING SOMATIC MUTATIONS BASED ON BIOLOGICAL KNOWLEDGE FOR PREDICTING SURVIVAL: AN APPLICATION IN RENAL CELL CARCINOMA

Dokyoon Kim, Ruowang Li, Scott M. Dudek, John R. Wallace, Marylyn D. Ritchie

Pennsylvania State University

Enormous efforts of whole exome and genome sequencing from hundreds to thousands of patients have provided the landscape of somatic genomic alterations in many cancer types to distinguish between driver mutations and passenger mutations. Driver mutations show strong associations with cancer clinical outcomes such as survival. However, due to the heterogeneity of tumors, somatic mutation profiles are exceptionally sparse whereas other types of genomic data such as miRNA or gene expression contain much more complete data for all genomic features with quantitative values measured in each patient. To overcome the extreme sparseness of somatic mutation profiles and allow for the discovery of combinations of somatic mutations that may predict cancer clinical outcomes, here we propose a new approach for binning somatic mutations based on existing biological knowledge. Through the analysis using renal cell carcinoma dataset from The Cancer Genome Atlas (TCGA), we identified combinations of somatic mutation burden based on pathways, protein families, evolutionary conversed regions, and regulatory regions associated with survival. Due to the nature of heterogeneity in cancer, using a binning strategy for somatic mutation profiles based on biological knowledge will be valuable for improved prognostic biomarkers and potentially for tailoring therapeutic strategies by identifying combinations of driver mutations.

TOPOLOGICAL FEATURES IN CANCER GENE EXPRESSION DATA

Svetlana Lockwood, Bala Krishnamoorthy

Washington State University

We present a new method for exploring cancer gene expression data based on tools from algebraic topology. Our method selects a small relevant subset from tens of thousands of genes while simultaneously identifying nontrivial higher order topological features, i.e., holes, in the data. We first circumvent the problem of high dimensionality by dualizing the data, i.e., by studying genes as points in the sample space. Then we select a small subset of the genes as landmarks to construct topological structures that capture persistent, i.e., topologically significant, features of the data set in its first homology group. Furthermore, we demonstrate that many members of these loops have been implicated for cancer biogenesis in scientific literature. We illustrate our method on five different data sets belonging to brain, breast, leukemia, and ovarian cancers.

DISTANT SUPERVISION FOR CANCER PATHWAY EXTRACTION FROM TEXT

Hoifung Poon, Kristina Toutanova, Chris Quirk

Microsoft Research

Biological pathways are central to understanding complex diseases such as cancer. The majority of this knowledge is scattered in the vast and rapidly growing research literature. To automate knowledge extraction, machine learning approaches typically require annotated examples, which are expensive and time-consuming to acquire. Recently, there has been increasing interest in leveraging databases for distant supervision in knowledge extraction, but existing applications focus almost exclusively on newswire domains. In this paper, we present the first attempt to formulate the distant supervision problem for pathway extraction and apply a state-of-the-art method to extracting pathway interactions from PubMed abstracts. Experiments show that distant supervision can effectively compensate for the lack of annotation, attaining an accuracy approaching supervised results. From 22 million PubMed abstracts, we extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of interactions are mentioned in the context of one or more cancer types, analysis of which yields interesting insights.

UNSUPERVISED FEATURE CONSTRUCTION AND KNOWLEDGE EXTRACTION FROM GENOME-WIDE ASSAYS OF BREAST CANCER WITH DENOISING AUTOENCODERS

Jie Tan, Matthew Ung, Chao Cheng, Casey S. Greene

Department of Genetics, Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, The Geisel School of Medicine at Dartmouth

Big data bring new opportunities for methods that efficiently summarize and automatically extract knowledge from such compendia. While both supervised learning algorithms and unsupervised clustering algorithms have been successfully applied to biological data, they are either dependent on known biology or limited to discerning the most significant signals in the data. Here we present denoising autoencoders (DAs), which employ a data-defined learning objective independent of known biology, as a method to identify and extract complex patterns from genomic data. We evaluate the performance of DAs by applying them to a large collection of breast cancer gene expression data. Results show that DAs successfully construct features that contain both clinical and molecular information. There are features that represent tumor or normal samples, estrogen receptor (ER) status, and molecular subtypes. Features constructed by the autoencoder generalize to an independent dataset collected using a distinct experimental platform. By integrating data from ENCODE for feature interpretation, we discover a feature representing ER status through association with key transcription factors in breast cancer. We also identify a feature highly predictive of patient survival and it is enriched by FOXM1 signaling pathway. The features constructed by DAs are often bimodally distributed with one peak near zero and another near one, which facilitates discretization. In summary, we demonstrate that DAs effectively extract key biological principles from gene expression data and summarize them into constructed features with convenient properties.

AUTOMATED GENE EXPRESSION PATTERN ANNOTATION IN THE MOUSE BRAIN

Tao Yang[1], Xinlin Zhao[1], Binbin Lin[1], Tao Zeng[2], Shuiwang Ji[2], Jieping Ye[1]

[1]Arizona State University, [2]Old Dominion University

Brain tumor is a fatal central nervous system disease that occurs in around 250,000 people each year globally and it is the second cause of cancer in children. It has been widely acknowledged that genetic factor is one of the significant risk factors for brain cancer. Thus, accurate descriptions of the locations of where the relative genes are active and how these genes express are critical for understanding the pathogenesis of brain tumor and for early detection. The Allen Developing Mouse Brain Atlas is a project on gene expression over the course of mouse brain development stages. Utilizing mouse models allows us to use a relatively homogeneous system to reveal the genetic risk factor of brain cancer. In the Allen atlas, about 435,000 high-resolution spatiotemporal in situ hybridization images have been generated for approximately 2,100 genes and currently the expression patterns over specific brain regions are manually annotated by experts, which does not scale with the continuously expanding collection of images. In this paper, we present an efficient computational approach to perform automated gene expression pattern annotation on brain images. First, the gene expression information in the brain images is captured by invariant features extracted from local image patches. Next, we adopt an augmented sparse coding method, called Stochastic Coordinate Coding, to construct high-level representations. Different pooling methods are then applied to generate gene-level features. To discriminate gene expression patterns at specific brain regions, we employ supervised learning methods to build accurate models for both binary-class and multi-class cases. Random undersampling and majority voting strategies are utilized to deal with the inherently imbalanced class distribution within each annotation task in order to further improve predictive performance. In addition, we propose a novel structure-based multi-label classification approach, which makes use of label hierarchy based on brain ontology during model learning. Extensive experiments have been conducted on the atlas and results show that the proposed approach produces higher annotation accuracy than several baseline methods. Our approach is shown to be robust on both binary-class and multi-class tasks and even with a relatively low training ratio. Our results also show that the use of label hierarchy can significantly improve the annotation accuracy at all brain ontology levels.

CHARACTERIZING THE IMPORTANCE OF ENVIRONMENTAL EXPOSURES, INTERACTIONS BETWEEN THE ENVIRONMENT AND GENETIC ARCHITECTURE, AND GENETIC INTERACTIONS: NEW METHODS FOR UNDERSTANDING THE ETIOLOGY OF COMPLEX TRAITS AND DISEASE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

MEASURES OF EXPOSURE IMPACT GENETIC ASSOCIATION STUDIES: AN EXAMPLE IN VITAMIN K LEVELS AND VKORC1

Dana C. Crawford[1], Kristin Brown-Gentry[2], Mark J. Rieder[3]

[1]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, [2]Center for Human Genetics Research, Vanderbilt University, [3]Adaptive Biotechnologies Corporation

Studies assessing the impact of gene-environment interactions on common human diseases and traits have been relatively few for many reasons. One often acknowledged reason is that it is difficult to accurately measure the environment or exposure. Indeed, most large-scale epidemiologic studies use questionnaires to assess and measure past and current exposure levels. While questionnaires may be cost-effective, the data may or may not accurately represent the exposure compared with more direct measurements (e.g., self-reported current smoking status versus direct measurement for cotinine levels). Much like phenotyping, the choice in how an exposure is measured may impact downstream tests of genetic association and gene-environment interaction studies. As a case study, we performed tests of association between five common VKORC1 SNPs and two different measurements of vitamin K levels, dietary (n=5,725) and serum (n=348), in the Third National Health and Nutrition Examination Studies (NHANES III). We did not replicate previously reported associations between VKORC1 and vitamin K levels using either measure. Furthermore, the suggestive associations and estimated genetic effect sizes identified in this study differed depending on the vitamin K measurement. This case study of VKORC1 and vitamin K levels serves as a cautionary example of the downstream consequences that the type of exposure measurement choices will have on genetic association and possibly gene-environment studies.

A BIPARTITE NETWORK APPROACH TO INFERRING INTERACTIONS BETWEEN ENVIRONMENTAL EXPOSURES AND HUMAN DISEASES

Christian Darabos, Emily D. Grussing, Maria E. Cricco, Kenzie A. Clark, Jason H. Moore

Dartmouth College

Environmental exposure is a key factor of understanding health and diseases. Beyond genetic propensities, many disorders are, in part, caused by human interaction with harmful substances in the water, the soil, or the air. Limited data is available on a disease or substance basis. However, we compile a global repository from literature surveys matching environmental chemical substances exposure with human disorders. We build a bipartite network linking 60 substances to over 150 disease phenotypes. We quantitatively and qualitatively analyze the network and its projections as simple networks. We identify mercury, lead and cadmium as associated with the largest number of disorders. Symmetrically, we show that breast cancer, harm to the fetus and non-Hodgkin's lymphoma are associated with the most environmental chemicals. We conduct statistical analysis of how vertices with similar characteristics form the network interactions. This dyadicity and heterophilicity measures the tendencies of vertices with similar properties to either connect to one-another. We study the dyadic distribution of the substance classes in the networks show that, for instance, tobacco smoke compounds, parabens and heavy metals tend to be connected, which hint at common disease causing factors, whereas fungicides and phytoestrogens do not. We build an exposure network at the systems level. The information gathered in this study is meant to be complementary to the genome and help us understand complex diseases, their commonalities, their causes, and how to prevent and treat them.

A SCREENING-TESTING APPROACH FOR DETECTING GENE-ENVIRONMENT INTERACTIONS USING SEQUENTIAL PENALIZED AND UNPENALIZED MULTIPLE LOGISTIC REGRESSION

H. Robert Frost, Angeline S. Andrew, Margaret R. Karagas, Jason H. Moore

Dartmouth College

Gene-environment (GxE) interactions are biologically important for a wide range of environmental exposures and clinical outcomes. Because of the large number of potential interactions in genome-wide association data, the standard approach fits one model per GxE interaction with multiple hypothesis correction (MHC) used to control the type I error rate. Although sometimes effective, using one model per candidate GxE interaction test has two important limitations: low power due to MHC and omitted variable bias. To avoid the coefficient estimation bias associated with independent models, researchers have used penalized regression methods to jointly test all main effects and interactions in a single regression model. Although penalized regression supports joint analysis of all interactions, can be used with hierarchical constraints, and offers excellent predictive performance, it cannot assess the statistical significance of GxE interactions or compute meaningful estimates of effect size. To address the challenge of low power, researchers have separately explored screening-testing, or two-stage, methods in which the set of potential GxE interactions is first filtered and then tested for interactions with MHC only applied to the tests actually performed in the second stage. Although two-stage methods are statistically valid and effective at improving power, they still test multiple separate models and so are impacted by MHC and biased coefficient estimation. To remedy the challenges of both poor power and omitted variable bias encountered with traditional GxE interaction detection methods, we propose a novel approach that combines elements of screening-testing and hierarchical penalized regression. Specifically, our proposed method uses, in the first stage, an elastic net-penalized multiple logistic regression model to jointly estimate either the marginal association filter statistic or the gene-environment correlation filter statistic for all candidate genetic markers. In the second stage, a single multiple logistic regression model is used to jointly assess marginal terms and GxE interactions for all genetic markers that pass the first stage filter. A single likelihood-ratio test is used to determine whether any of the interactions are statistically significant. We demonstrate the efficacy of our method relative to alternative GxE detection methods on a bladder cancer data set.

VARIABLE SELECTION METHOD FOR THE IDENTIFICATION OF EPISTATIC MODELS

Emily Rose Holzinger[1], Silke Szymczak[1], Abhijit Dasgupta[2], James Malley[3], Qing Li[1], Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch (NHGRI, NIH), [2]Clinical Trials and Outcomes Branch (NIAMS, NIH), [3]Center for Information Technology (NIH)

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, such as interactions between variables with small main effects. These types of effects likely contribute to the heritability of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no standardized method of selecting variables so that false positives are reduced while retaining adequate power. To this end, we have developed a novel variable selection method called relative recurrency variable importance metric (r2VIM). This method incorporates recurrency and variance estimation to assist in optimal threshold selection. For this study, we specifically address how this method performs in data with almost completely epistatic effects (i.e. no marginal effects). Our results show that with appropriate parameter settings, r2VIM can identify interaction effects when the marginal effects are virtually nonexistent. It also outperforms logistic regression, which has essentially no power under this type of model when the number of potential features (genetic variants) is large.

GENOME-WIDE GENETIC INTERACTION ANALYSIS OF GLAUCOMA USING EXPERT KNOWLEDGE DERIVED FROM HUMAN PHENOTYPE NETWORKS

Ting Hu, Christian Darabos, Maria E. Cricco, Emily Kong, Jason H. Moore

Dartmouth College

The large volume of GWAS data poses great computational challenges for analyzing genetic interactions associated with common human diseases. We propose a computational framework for characterizing epistatic interactions among large sets of genetic attributes in GWAS data. We build the human phenotype network (HPN) and focus around a disease of interest. In this study, we use the GLAUGEN glaucoma GWAS dataset and apply the HPN as a biological knowledge-based filter to prioritize genetic variants. Then, we use the statistical epistasis network (SEN) to identify a significant connected network of pairwise epistatic interactions among the prioritized SNPs. These clearly highlight the complex genetic basis of glaucoma. Furthermore, we identify key SNPs by quantifying structural network characteristics. Through functional annotation of these key SNPs using Biofilter, a software accessing multiple publicly available human genetic data sources, we find supporting biomedical evidence linking glaucoma to an array of genetic diseases, proving our concept. We conclude by suggesting hypotheses for a better understanding of the disease.

IDENTIFICATION OF GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS WITHIN THE FIBRINOGEN GENE CLUSTER FOR FIBRINOGEN LEVELS IN THREE ETHNICALLY DIVERSE POPULATIONS

Janina M. Jeff, Kristin Brown-Gentry, Dana C. Crawford

Icahn School of Medicine at Mount Sinai, Cigna-Health Spring, Case Western Reserve University

Elevated levels of plasma fibrinogen are associated with clot formation in the absence of inflammation or injury and is a biomarker for arterial clotting, the leading cause of cardiovascular disease. Fibrinogen levels are heritable with >50% attributed to genetic factors, however little is known about possible genetic modifiers that might explain the missing heritability. The fibrinogen gene cluster is comprised of three genes (FGA, FGB, and FGG) that make up the fibrinogen polypeptide essential for fibrinogen production in the blood. Given the known interaction with these genes, we tested 25 variants in the fibrinogen gene cluster for gene x gene and gene x environment interactions in 620 non-Hispanic blacks, 1,385 non-Hispanic whites, and 664 Mexican Americans from a cross-sectional dataset enriched with environmental data, the Third National Health and Nutrition Examination Survey (NHANES III). Using a multiplicative approach, we added cross product terms (gene x gene or gene x environment) to a linear regression model and declared significance at $p < 0.05$. We identified 19 unique gene x gene and 13 unique gene x environment interactions that impact fibrinogen levels in at least one population at $p <0.05$. Over 90% of the gene x gene interactions identified include a variant in the rate-limiting gene, FGB that is essential for the formation of the fibrinogen polypeptide. We also detected gene x environment interactions with fibrinogen variants and sex, smoking, and body mass index. These findings highlight the potential for the discovery of genetic modifiers for complex phenotypes in multiple populations and give a better understanding of the interaction between genes and/or the environment for fibrinogen levels. The need for more powerful and robust methods to identify genetic modifiers is still warranted.

DEVELOPMENT OF EXPOSOME CORRELATIONS GLOBES TO MAP OUT ENVIRONMENT-WIDE ASSOCIATIONS

Chirag J. Patel[1], Arjun K. Manrai[2]

[1]Center for Biomedical Informatics, Harvard Medical School, [2]Center for Biomedical Informatics, Harvard Medical School, Harvard-MIT Division of Health Sciences and Technology

The environment plays a major role in influencing diseases and health. The phenomenon of environmental exposure is complex and humans are not exposed to one or a handful factors but potentially hundreds factors throughout their lives. The exposome, the totality of exposures encountered from birth, is hypothesized to consist of multiple inter-dependencies, or correlations, between individual exposures. These correlations may reflect how individuals are exposed. Currently, we lack methods to comprehensively identify robust and replicated correlations between environmental exposures of the exposome. Further, we have not mapped how exposures associated with disease identified by environment-wide association studies (EWAS) are correlated with other exposures. To this end, we implement methods to describe a first "exposome globe", a comprehensive display of replicated correlations between individual exposures of the exposome. First, we describe overall characteristics of the dense correlations between exposures, showing that we are able to replicate 2,656 correlations between individual exposures of 81,937 total considered (3%). We document the correlation within and between broad a priori defined categories of exposures (e.g., pollutants and nutrient exposures). We also demonstrate utility of the exposome globe to contextualize exposures found through two EWASs in type 2 diabetes and all-cause mortality, such as exposure clusters putatively related to smoking behaviors and persistent pollutant exposure. The exposome globe construct is a useful tool for the display and communication of the complex relationships between exposure factors and between exposure factors related to disease status.

MITOCHONDRIAL VARIATION AND THE RISK OF AGE-RELATED MACULAR DEGENERATION ACROSS DIVERSE POPULATIONS

Nicole A. Restrepo[1], Sabrina L. Mitchell[1], Robert J. Goodloe[1], Deborah G. Murdock[1], Jonathan L. Haines[2], Dana C. Crawford[2]

[1]Center for Human Genetics Research, Vanderbilt University, [2]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University

Substantial progress has been made in identifying susceptibility variants for age-related macular degeneration (AMD). The majority of research to identify genetic variants associated with AMD has focused on nuclear genetic variation. While there is some evidence that mitochondrial genetic variation contributes to AMD susceptibility, to date, these studies have been limited to populations of European descent resulting in a lack of data in diverse populations. A major goal of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study is to describe the underlying genetic architecture of common, complex diseases across diverse populations. This present study sought to determine if mitochondrial genetic variation influences risk of AMD across diverse populations. We performed a genetic association study to investigate the contribution of mitochondrial DNA variation to AMD risk. We accessed samples from the National Health and Nutrition Examination Surveys, a U.S population-based, cross-sectional survey collected without regard to health status. AMD cases and controls were selected from the Third NHANES and NHANES 2007-2008 datasets which include non-Hispanic whites, non-Hispanic blacks, and Mexican Americans. AMD cases were defined as those > 60 years of age with early/late AMD, as determined by fundus photography. Targeted genotyping was performed for 63 mitochondrial SNPs and participants were then classified into mitochondrial haplogroups. We used logistic regression assuming a dominant genetic model adjusting for age, sex, body mass index, and smoking status (ever vs. never). Regressions and meta-analyses were performed for individual SNPs and mitochondrial haplogroups J, T, and U. We identified five SNPs associated with AMD in Mexican Americans at $p < 0.05$, including three located in the control region (mt16111, mt16362, and mt16319), one in MT-RNR2 (mt1736), and one in MT-ND4 (mt12007). No mitochondrial variant or haplogroup was significantly associated in non-Hispanic blacks or non-Hispanic whites in the final meta-analysis. This study provides further evidence that mitochondrial variation plays a role in susceptibility to AMD and contributes to the knowledge of the genetic architecture of AMD in Mexican Americans.

iPINBPA: AN INTEGRATIVE NETWORK-BASED FUNCTIONAL MODULE DISCOVERY TOOL FOR GENOME-WIDE ASSOCIATION STUDIES

Lili Wang[1], Parvin Mousavi[1], Sergio E. Baranzini[2]

[1]Queen's University, [2]University of California San Francisco

We introduce the integrative protein-interaction-network-based pathway analysis (iPINBPA) for genome-wide association studies (GWAS), a method to identify and prioritize genetic associations by merging statistical evidence of association with physical evidence of interaction at the protein level. First, the strongest associations are used to weight all nodes in the PPI network using a guilt-by-association approach. Second, the gene-wise converted p-values from a GWAS are integrated with node weights using the Liptak-Stouffer method. Finally, a greedy search is performed to find enriched modules, i.e., sub-networks with nodes that have low p-values and high weights. The performance of iPINBPA and other state-of-the-art methods is assessed by computing the concentrated receiver operating characteristic (CROC) curves using two independent multiple sclerosis (MS) GWAS studies and one recent ImmunoChip study. Our results showed that iPINBPA identified sub-networks with smaller sizes and higher enrichments than other methods. iPINBPA offers a novel strategy to integrate topological connectivity and association signals from GWAS, making this an attractive tool to use in other large GWAS datasets.

CROWDSOURCING AND MINING CROWD DATA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

REPUTATION-BASED COLLABORATIVE NETWORK BIOLOGY

Jean Binder[1], Stephanie Boue[1], Anselmo Di Fabio[2], R. Brett Fields[3], William Hayes[3], Julia Hoeng[1], Jennifer S. Park[3], Manuel C. Peitsch[1]

[1]Philip Morris International R&D, Philip Morris Products S.A., [2]Applied Dynamic Solutions, LLC, [3]Selventa

A pilot reputation-based collaborative network biology platform, Bionet, was developed for use in the sbv IMPROVER Network Verification Challenge to verify and enhance previously developed networks describing key aspects of lung biology. Bionet was successful in capturing a more comprehensive view of the biology associated with each network using the collective intelligence and knowledge of the crowd. One key learning point from the pilot was that using a standardized biological knowledge representation language such as BEL is critical to the success of a collaborative network biology platform. Overall, Bionet demonstrated that this approach to collaborative network biology is highly viable. Improving this platform for de novo creation of biological networks and network curation with the suggested enhancements for scalability will serve both academic and industry systems biology communities.

MICROTASK CROWDSOURCING FOR DISEASE MENTION ANNOTATION IN PUBMED ABSTRACTS

Benjamin M. Good, Max Nanis, Chunlei Wu, Andrew I. Su

Molecular and Experimental Medicine, The Scripps Research Institute

Identifying concepts and relationships in biomedical text enables knowledge to be applied in computational analyses. Many biological natural language processing (BioNLP) projects attempt to address this challenge, but the state of the art still leaves much room for improvement. Progress in BioNLP research depends on large, annotated corpora for evaluating information extraction systems and training machine learning models. Traditionally, such corpora are created by small numbers of expert annotators often working over extended periods of time. Recent studies have shown that workers on microtask crowdsourcing platforms such as Amazon's Mechanical Turk (AMT) can, in aggregate, generate high-quality annotations of biomedical text. Here, we investigated the use of the AMT in capturing disease mentions in PubMed abstracts. We used the NCBI Disease corpus as a gold standard for refining and benchmarking our crowdsourcing protocol. After several iterations, we arrived at a protocol that reproduced the annotations of the 593 documents in the 'training set' of this gold standard with an overall F measure of 0.872 (precision 0.862, recall 0.883). The output can also be tuned to optimize for precision (max = 0.984 when recall = 0.269) or recall (max = 0.980 when precision = 0.436). Each document was completed by 15 workers, and their annotations were merged based on a simple voting method. In total 145 workers combined to complete all 593 documents in the span of 9 days at a cost of $.066 per abstract per worker. The quality of the annotations, as judged with the F measure, increases with the number of workers assigned to each task; however minimal performance gains were observed beyond 8 workers per task. These results add further evidence that microtask crowdsourcing can be a valuable tool for generating well-annotated corpora in BioNLP. Data produced for this analysis are available at http://figshare.com/articles/Disease_Mention_Annotation_with_Mechanical_Turk/1126402.

CROWDSOURCING IMAGE ANNOTATION FOR NUCLEUS DETECTION AND SEGMENTATION IN COMPUTATIONAL PATHOLOGY: EVALUATING EXPERTS, AUTOMATED METHODS, AND THE CROWD

Humayun Irshad, Laleh Montaser-Kouhsari, Gail Waltz, Octavian Bucur, Jonathan A. Nowak, Fei Dong, Nicholas W. Knoblauch, Andrew H. Beck

Beth Israel Deaconess Medical Center and Harvard Medical School

The development of tools in computational pathology to assist physicians and biomedical scientists in the diagnosis of disease requires access to high-quality annotated images for algorithm learning and evaluation. Generating high-quality expert-derived annotations is time-consuming and expensive. We explore the use of crowdsourcing for rapidly obtaining annotations for two core tasks in computational pathology: nucleus detection and nucleus segmentation. We designed and implemented crowdsourcing experiments using the CrowdFlower platform, which provides access to a large set of labor channel partners that accesses and manages millions of contributors worldwide. We obtained annotations from four types of annotators and compared concordance across these groups. We obtained: crowdsourced annotations for nucleus detection and segmentation on a total of 810 images; annotations using automated methods on 810 images; annotations from research fellows for detection and segmentation on 477 and 455 images, respectively; and expert pathologist-derived annotations for detection and segmentation on 80 and 63 images, respectively. For the crowdsourced annotations, we evaluated performance across a range of contributor skill levels (1, 2, or 3). The crowdsourced annotations (4,860 images in total) were completed in only a fraction of the time and cost required for obtaining annotations using traditional methods. For the nucleus detection task, the research fellow-derived annotations showed the strongest concordance with the expert pathologist-derived annotations (F-M =93.68%), followed by the crowd-sourced contributor levels 1,2, and 3 and the automated method, which showed relatively similar performance (F-M = 87.84%, 88.49%, 87.26%, and 86.99%, respectively). For the nucleus segmentation task, the crowdsourced contributor level 3-derived annotations, research fellow-derived annotations, and automated method showed the strongest concordance with the expert pathologist-derived annotations (F-M = 66.41%, 65.93%, and 65.36%, respectively), followed by the contributor levels 2 and 1 (60.89% and 60.87%, respectively). When the research fellows were used as a gold-standard for the segmentation task, all three contributor levels of the crowdsourced annotations significantly outperformed the automated method (F-M = 62.21%, 62.47%, and 65.15% vs. 51.92%). Aggregating multiple annotations from the crowd to obtain a consensus annotation resulted in the strongest performance for the crowd-sourced segmentation. For both detection and segmentation, crowd-sourced performance is strongest with small images (400 x 400 pixels) and degrades significantly with the use of larger images (600 x 600 and 800 x 800 pixels). We conclude that crowdsourcing to non-experts can be used for large-scale labeling microtasks in computational pathology and offers a new approach for the rapid generation of labeled images for algorithm development and evaluation.

ANALYZING SEARCH BEHAVIOR OF HEALTHCARE PROFESSIONALS FOR DRUG SAFETY SURVEILLANCE

David J. Odgers[1], Rave Harpaz[1], Alison Callahan[1], Gregor Stiglic[2], Nigam H. Shah[1]

[1]Stanford University, [2]University of Maribor

Post-market drug safety surveillance is hugely important and is a significant challenge despite the existence of adverse event (AE) reporting systems. Here we describe a preliminary analysis of search logs from healthcare professionals as a source for detecting adverse drug events. We annotate search log query terms with biomedical terminologies for drugs and events, and then perform a statistical analysis to identify associations among drugs and events within search sessions. We evaluate our approach using two different types of reference standards consisting of known adverse drug events (ADEs) and negative controls. Our approach achieves a discrimination accuracy of 0.85 in terms of the area under the receiver operator curve (AUC) for the reference set of well-established ADEs and an AUC of 0.68 for the reference set of recently labeled ADEs. We also find that the majority of associations in the reference sets have support in the search log data. Despite these promising results additional research is required to better understand users' search behavior, biasing factors, and the overall utility of analyzing healthcare professional search logs for drug safety surveillance.

REFINING LITERATURE CURATED PROTEIN INTERACTIONS USING EXPERT OPINIONS

Oznur Tastan[1], Yanjun Qi[2], Jaime G. Carbonell[3], Judith Klein-Seetharaman[4]

[1]Bilkent University, [2]University of Virginia, [3]Carnegie Mellon University, [4]University of Warwick

The availability of high-quality physical interaction datasets is a prerequisite for system-level analysis of interactomes and supervised models to predict protein-protein interactions (PPIs). One source is literature-curated PPI databases in which pairwise associations of proteins published in the scientific literature are deposited. However, PPIs may not be clearly labelled as physical interactions affecting the quality of the entire dataset. In order to obtain a high-quality gold standard dataset for PPIs between human immunodeficiency virus (HIV-1) and its human host, we adopted a crowd-sourcing approach. We collected expert opinions and utilized an expectation-maximization based approach to estimate expert labeling quality. These estimates are used to infer the probability of a reported PPI actually being a direct physical interaction given the set of expert opinions. The effectiveness of our approach is demonstrated through synthetic data experiments and a high quality physical interaction network between HIV and human proteins is obtained. Since many literature-curated databases suffer from similar challenges, the framework described herein could be utilized in refining other databases. The curated data is available at http://www.cs.bilkent.edu.tr/~oznur.tastan/supp/psb2015/

CROWDSOURCING RNA STRUCTURAL ALIGNMENTS WITH AN ONLINE COMPUTER GAME

Jérôme Waldispühl[1], Arthur Kam[1], Paul P. Gardner[2]

[1]McGill University, [2]University of Canterbury

The annotation and classification of ncRNAs is essential to decipher molecular mechanisms of gene regulation in normal and disease states. A database such as Rfam maintains alignments, consensus secondary structures, and corresponding annotations for RNA families. Its primary purpose is the automated, accurate annotation of non-coding RNAs in genomic sequences. However, the alignment of RNAs is computationally challenging, and the data stored in this database are often subject to improvements. Here, we design and evaluate Ribo, a human-computing game that aims to improve the accuracy of RNA alignments already stored in Rfam. We demonstrate the potential of our techniques and discuss the feasibility of large scale collaborative annotation and classification of RNA families.

PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARD IMPROVED MEDICINE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

KLEAT: CLEAVAGE SITE ANALYSIS OF TRANSCRIPTOMES

Inanç Birol, Anthony Raymond, Readman Chiu, Ka Ming Nip, Shaun D. Jackman, Maayan Kreitzman, T. Roderick Docking, Catherine A. Ennis, A. Gordon Robertson, Aly Karsan

British Columbia Cancer Agency

In eukaryotic cells, alternative cleavage of 3’ untranslated regions (UTRs) can affect transcript stability, transport and translation. For polyadenylated (poly(A)) transcripts, cleavage sites can be characterized with short-read sequencing using specialized library construction methods. However, for large-scale cohort studies as well as for clinical sequencing applications, it is desirable to characterize such events using RNA-seq data, as the latter are already widely applied to identify other relevant information, such as mutations, alternative splicing and chimeric transcripts. Here we describe KLEAT, an analysis tool that uses de novo assembly of RNA-seq data to characterize cleavage sites on 3’ UTRs. We demonstrate the performance of KLEAT on three cell line RNA-seq libraries constructed and sequenced by the ENCODE project, and assembled using Trans-ABySS. Validating the KLEAT predictions with matched ENCODE RNA-seq and RNA-PET libraries, we show that the tool has over 90% positive predictive value when there are at least three RNA-seq reads supporting a poly(A) tail and requiring at least three RNA-PET reads mapping within 100 nucleotides as validation. We also compare the performance of KLEAT with other popular RNA-seq analysis pipelines that reconstruct 3’ UTR ends, and show that it performs favourably, based on an ROC-like curve.

CAUSAL INFERENCE IN BIOLOGY NETWORKS WITH INTEGRATED BELIEF PROPAGATION

Rui Chang, Jonathan R. Karr, Eric E. Schadt

Department of Genetics and Genomic Sciences, Ichan School of Medicine at Mount Sinai

Inferring causal relationships among molecular and higher order phenotypes is a critical step in elucidating the complexity of living systems. Here we propose a novel method for inferring causality that is no longer constrained by the conditional dependency arguments that limit the ability of statistical causal inference methods to resolve causal relationships within sets of graphical models that are Markov equivalent. Our method utilizes Bayesian belief propagation to infer the responses of perturbation events on molecular traits given a hypothesized graph structure. A distance measure between the inferred response distribution and the observed data is defined to assess the 'fitness' of the hypothesized causal relationships. To test our algorithm, we infer causal relationships within equivalence classes of gene networks in which the form of the functional interactions that are possible are assumed to be nonlinear, given synthetic microarray and RNA sequencing data. We also apply our method to infer causality in real metabolic network with v-structure and feedback loop. We show that our method can recapitulate the causal structure and recover the feedback loop only from steady-state data which conventional method cannot.

MACHINE LEARNING FROM CONCEPT TO CLINIC: RELIABLE DETECTION OF BRAF V600E DNA MUTATIONS IN THYROID NODULES USING HIGH-DIMENSIONAL RNA EXPRESSION DATA

James Diggans[1], Su Yeon Kim[1], Zhanzhi Hu[1], Daniel Pankratz[1], Mei Wong[1], Jessica Reynolds[1], Ed Tom[1], Moraima Pagan[1], Robert Monroe[1], Juan Rosai[2], Virginia A. Livolsi[3], Richard B. Lanman[1], Richard T. Kloos[1], P. Sean Walsh[1], Giulia C. Kennedy[1]

[1]Veracyte Inc., [2]Centro Diagnostico Italiano, [3]University of Pennsylvania

The promise of personalized medicine will require rigorously validated molecular diagnostics developed on minimally invasive, clinically relevant samples. Measurement of DNA mutations is increasingly common in clinical settings but only higher-prevalence mutations are cost-effective. Patients with rare variants are at best ignored or, at worst, misdiagnosed. Mutations result in downstream impacts on transcription, offering the possibility of broader diagnosis for patients with rare variants causing similar downstream changes. Use of such signatures in clinical settings is rare as these algorithms are difficult to validate for commercial use. Validation on a test set (against a clinical gold standard) is necessary but not sufficient: accuracy must be maintained amidst interfering substances, across reagent lots and across operators. Here we report the development, clinical validation, and diagnostic accuracy of a pre-operative molecular test (Afirma BRAF) to identify BRAF V600E mutations using mRNA expression in thyroid fine needle aspirate biopsies (FNABs). FNABs were obtained prospectively from 716 nodules and more than 3,000 features measured using microarrays. BRAF V600E labels for training (n=181) and independent test (n=535) sets were established using a sensitive quantitative PCR (qPCR) assay. The resulting 128-gene linear support vector machine was compared to qPCR in the independent test set. Clinical sensitivity and specificity for malignancy were evaluated in a subset of test set samples (n=213) with expert-derived histopathology. We observed high positive- (PPA, 90.4%) and negative (NPA, 99.0%) percent agreement with qPCR on the test set. Clinical sensitivity for malignancy was 43.8% (consistent with published prevalence of BRAF V600E in this neoplasm) and specificity was 100%, identical to qPCR on the same samples. Classification was accurate in up to 60% blood. A double-mutant still resulting in the V600E amino acid change was negative by qPCR but correctly positive by Afirma BRAF. Non-diagnostic rates were lower (7.6%) for Afirma BRAF than for qPCR (24.5%), a further advantage of using RNA in small sample biopsies. Afirma BRAF accurately determined the presence or absence of the BRAF V600E DNA mutation in FNABs, a collection method directly relevant to solid tumor assessment, with performance equal to that of an established, highly sensitive DNA-based assay and with a lower non-diagnostic rate. This is the first such test in thyroid cancer to undergo sufficient analytical and clinical validation for real-world use in a personalized medicine context to frame individual patient risk and inform surgical choice.

A SYSTEMATIC ASSESSMENT OF LINKING GENE EXPRESSION WITH GENETIC VARIANTS FOR PRIORITIZING CANDIDATE TARGETS

Hua Fan-Minogue, Bin Chen, Weronika Sikora-Wohlfeld, Marina Sirota, Atul J. Butte

Division of System Medicine, Department of Pediatrics, Stanford University

Gene expression and disease-associated variants are often used to prioritize candidate genes for target validation. However, the success of these gene features alone or in combination in the discovery of therapeutic targets is uncertain. Here we evaluated the effectiveness of the differential expression (DE), the disease-associated single nucleotide polymorphisms (SNPs) and the combination of the two in recovering and predicting known therapeutic targets across 56 human diseases. We demonstrate that the performance of each feature varies across diseases and generally the features have more recovery power than predictive power. The combination of the two features, however, has significantly higher predictive power than each feature alone. Our study provides a systematic evaluation of two common gene features, DE and SNPs, for prioritization of candidate targets and identified an improved predictive power of coupling these two features.

DRUG-INDUCED mRNA SIGNATURES ARE ENRICHED FOR THE MINORITY OF GENES THAT ARE HIGHLY HERITABLE

Tianxiang Gao[1], Petter Brodin[2], Mark M. Davis[3], Vladimir Jojic[1]

[1]University of North Carolina at Chapel Hill, [2]Solna Karolinska Institutet, [3]Stanford University School of Medicine

The blood gene expression signatures are used as biomarkers for immunological and non-immunological diseases. Therefore, it is important to understand the variation in blood gene expression patterns and the factors (heritable/non-heritable) that underlie this variation. In this paper, we study the relationship between drug effects on the one hand, and heritable and non-heritable factors influencing gene expression on the other. Understanding of this relationship can help select appropriate targets for drugs aimed at reverting disease phenotypes to healthy states. In order to estimate heritable and non-heritable effects on gene expression, we use Twin-ACE model on a gene expression dataset MuTHER, measured on blood samples from monozygotic and dizygotic twins. In order to associate gene expression with drug effects, we use CMap database. We show that, even though the expressions of most genes are driven by non-heritable factors, drugs are more likely to influence expression of genes, driven by heritable rather than non-heritable factors. We further study this finding in the context of a gene regulatory network. We investigate the relationship between the drug effects on gene expression and propagation of heritable and non-heritable factors through regulatory networks. We find that the decisive factor in determining whether a gene will be influenced by a drug is the flow of heritable effects supplied to the gene through regulatory network.

AN INTEGRATIVE PIPELINE FOR MULTI-MODAL DISCOVERY OF DISEASE RELATIONSHIPS

Benjamin S. Glicksberg[1,2], Li Li[1], Wei-Yi Cheng[1], Khader Shameer[1], Jörg Hakenberg[1], Rafael Castellanos[1], Meng Ma[1], Lisong Shi[1], Hardik Shah[1], Joel T. Dudley[1,2], Rong Chen[1]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, [2]Department of Neuroscience, Icahn School of Medicine at Mount Sinai

In the past decade there has been an explosion in genetic research that has resulted in the generation of enormous quantities of disease-related data. In the current study, we have compiled disease risk gene variant information and Electronic Medical Record (EMR) classification codes from various repositories for 305 diseases. Using such data, we developed a pipeline to test for clinical prevalence, gene-variant overlap, and literature presence for all 46,360 unique diseases pairs. To determine whether disease pairs were enriched we systematically employed both Fishers' Exact (medical and literature) and Term Frequency-Inverse Document Frequency (genetics) methodologies to test for enrichment, defining statistical significance at a Bonferonni adjusted threshold of ($p < 1x10-6$) and weighted $q<0.05$ accordingly. We hypothesize that disease pairs that are statistically enriched in medical and genetic spheres, but not so in the literature have the potential to reveal non-obvious connections between clinically disparate phenotypes. Using this pipeline, we identified 2,316 disease pairs that were significantly enriched within an EMR and 213 enriched genetically. Of these, 65 disease pairs were statistically enriched in both, 19 of which are believed to be novel. These identified non-obvious relationships between disease pairs are suggestive of a shared underlying etiology with clinical presentation. Further investigation of uncovered disease-pair relationships has the potential to provide insights into the architecture of complex diseases, and update existing knowledge of risk factors.

PEAX: INTERACTIVE VISUAL ANALYSIS AND EXPLORATION OF COMPLEX CLINICAL PHENOTYPE AND GENE EXPRESSION ASSOCIATION

Michael Hinterberg, David P. Kao, Michael R. Bristow, Lawrence E. Hunter, J. David Port, Carsten Görg

University of Colorado School of Medicine

Increasing availability of high-dimensional clinical data, which improves the ability to define more specific phenotypes, as well as molecular data, which can elucidate disease mechanisms, is a driving force and at the same time a major challenge for translational and personalized medicine. Successful research in this field requires an approach that ties together specific disease and health expertise with understanding of molecular data through statistical methods. We present PEAX (Phenotype Expression Association eXplorer), built upon open-source software, which integrates visual phenotype model definition with statistical testing of expression data presented concurrently in a web-browser. The integration data and analysis tasks in a single tool allows clinical domain experts to obtain new insights directly through exploration of relationships between multivariate phenotype models and gene expression data, showing the effects of model definition and modification while also exploiting potential meaningful associations between phenotype and miRNA-mRNA regulatory relationships. We combine the web visualization capabilities of Shiny and D3 with the power and speed of R for backend statistical analysis, in order to abstract the scripting required for repetitive analysis of sub phenotype association. We describe the motivation for PEAX, demonstrate its utility through a use case involving heart failure research, and discuss computational challenges and observations. We show that our visual web-based representations are well-suited for rapid exploration of phenotype and gene expression association, facilitating insight and discovery by domain experts.

T-ReCS: STABLE SELECTION OF DYNAMICALLY FORMED GROUPS OF FEATURES WITH APPLICATION TO PREDICTION OF CLINICAL OUTCOMES

Grace T. Huang[1], Ioannis Tsamardinos[2], Vineet Raghu[1], Naftali Kaminski[3], Panayiotis V. Benos[1]

[1]University of Pittsburgh, [2]University of Crete, [3]Yale University

Feature selection is used extensively in biomedical research for biomarker identification and patient classification, both of which are essential steps in developing personalized medicine strategies. However, the structured nature of the biological datasets and high correlation of variables frequently yield multiple equally optimal signatures, thus making traditional feature selection methods unstable. Features selected based on one cohort of patients, may not work as well in another cohort. In addition, biologically important features may be missed due to selection of other co-clustered features We propose a new method, Tree-guided Recursive Cluster Selection (T-ReCS), for efficient selection of grouped features. T-ReCS significantly improves predictive stability while maintains the same level of accuracy. T-ReCS does not require an a priori knowledge of the clusters like group-lasso and also can handle “orphan” features (not belonging to a cluster). T-ReCS can be used with categorical or survival target variables. Tested on simulated and real expression data from breast cancer and lung diseases and survival data, T-ReCS selected stable cluster features without significant loss in classification accuracy.

META-ANALYSIS OF DIFFERENTIAL GENE CO-EXPRESSION: APPLICATION TO LUPUS

Sumit B. Makashir[1], Leah C. Kottyan[2], Matthew T. Weirauch[2]

[1]University of Cincinnati, [2]Cincinnati Children's Hospital

We present a novel statistical framework for meta-analysis of differential gene co-expression. In contrast to standard methods, which identify genes that are over or under expressed in disease vs controls, differential co-expression identifies gene pairs with correlated expression profiles specific to one state. We apply our differential co-expression meta-analysis method to identify genes specifically mis-expressed in blood-derived cells of systemic lupus erythematosus (SLE) patients. The resulting network is strongly enriched for genes genetically associated with SLE, and effectively identifies gene modules known to play important roles in SLE etiology, such as increased type 1 interferon response and response to wounding. Our results also strongly support previous preliminary studies suggesting a role for dysregulation of neutrophil extracellular trap formation in SLE. Strikingly, two of the gene modules we identify contain SLE-associated transcription factors that have binding sites significantly enriched in the promoter regions of their respective gene modules, suggesting a possible mechanism underlying the mis-expression of the modules. Thus, our general method is capable of identifying specific dysregulated gene expression programs, as opposed to large global responses. We anticipate that methods such as ours will be more and more useful as gene expression monitoring becomes increasingly common in clinical settings.

MELANCHOLIC DEPRESSION PREDICTION BY IDENTIFYING REPRESENTATIVE FEATURES IN METABOLIC AND MICROARRAY PROFILES WITH MISSING VALUES

Zhi Nie[1], Tao Yang[1], Yashu Liu[1], Binbin Lin[1], Qingyang Li[1], Vaibhav A. Narayan[2], Gayle Wittenberg[2], Jieping Ye[1]

[1]Arizona State University, [2]Johnson & Johnson Pharmaceutical Research & Development

Recent studies have revealed that melancholic depression, one major subtype of depression, is closely associated with the concentration of some metabolites and biological functions of certain genes and pathways. Meanwhile, recent advances in biotechnologies have allowed us to collect a large amount of genomic data, e.g., metabolites and microarray gene expression. With such a huge amount of information available, one approach that can give us new insights into the understanding of the fundamental biology underlying melancholic depression is to build disease status prediction models using classification or regression methods. However, the existence of strong empirical correlations, e.g., those exhibited by genes sharing the same biological pathway in microarray profiles, tremendously limits the performance of these methods. Furthermore, the occurrence of missing values which are ubiquitous in biomedical applications further complicates the problem. In this paper, we hypothesize that the problem of missing values might in some way benefit from the correlation between the variables and propose a method to learn a compressed set of representative features through an adapted version of sparse coding which is capable of identifying correlated variables and addressing the issue of missing values simultaneously. An efficient algorithm is also developed to solve the proposed formulation. We apply the proposed method on metabolic and microarray profiles collected from a group of subjects consisting of both patients with melancholic depression and healthy controls. Results show that the proposed method can not only produce meaningful clusters of variables but also generate a set of representative features that achieve superior classification performance over those generated by traditional clustering and data imputation techniques. In particular, on both datasets, we found that in comparison with the competing algorithms, the representative features learned by the proposed method give rise to significantly improved sensitivity scores, suggesting that the learned features allow prediction with high accuracy of disease status in those who are diagnosed with melancholic depression. To our best knowledge, this is the first work that applies sparse coding to deal with high feature correlations and missing values, which are common challenges in many biomedical applications. The proposed method can be readily adapted to other biomedical applications involving incomplete and high-dimensional data.

BAYCLONE: BAYESIAN NONPARAMETRIC INFERENCE OF TUMOR SUBCLONES USING NGS DATA

Subhajit Sengupta[1], Jin Wang[2], Juhee Lee[3], Peter Mueller[4], Kamalakar Gulukota[5], Arunava Banerjee[6], Yuan Ji[1,7]

[1]Center for Biomedical Research Informatics, NorthShore University HealthSystem; [2]Department of Statistics, University of Illinois at Urbana-Champaign; [3]Department of Applied Mathematics and Statistics, University of California Santa Cruz; [4]Department of Mathematics, University of Texas Austin; [5]Center for Molecular Medicine, NorthShore University HealthSystem; [6]Department of Computer & Information Science & Engineering, University of Florida; [7]Center for Biomedical Research Informatics, NorthShore University HealthSystem, Department of Health Studies, The University of Chicago

In this paper, we present a novel feature allocation model to describe tumor heterogeneity (TH) using next-generation sequencing (NGS) data. Taking a Bayesian approach, we extend the Indian buffet process (IBP) to define a class of nonparametric models, the categorical IBP (cIBP). A cIBP takes categorical values to denote homozygous or heterozygous genotypes at each SNV. We define a subclone as a vector of these categorical values, each corresponding to an SNV. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, we took a different approach using feature allocation. Importantly, we do not assume somatic mutations with similar cellular prevalence must be from the same subclone and allow overlapping mutations shared across subclones. We argue that this is closer to the underlying theory of phylogenetic clonal expansion, as somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Bayesian inference yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. We report results on both simulated and real data. BayClone is available at http://health.bsd.uchicago.edu/yji/soft.html.

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

PROCEEDINGS PAPER WITH POSTER PRESENTATION

STEPWISE GROUP SPARSE REGRESSION (SGSR): GENE-SET-BASED PHARMACOGENOMIC PREDICTIVE MODELS WITH STEPWISE SELECTION OF FUNCTIONAL PRIORS

In Sock Jang[1], Rodrigo Dienstmann[1], Adam A. Margolin[2], Justin Guinney[1]

[1]Sage Bionetworks, [2]Oregon Health & Science University

Complex mechanisms involving genomic aberrations in numerous proteins and pathways are believed to be a key cause of many diseases such as cancer. With recent advances in genomics, elucidating the molecular basis of cancer at a patient level is now feasible, and has led to personalized treatment strategies whereby a patient is treated according to his or her genomic profile. However, there is growing recognition that existing treatment modalities are overly simplistic, and do not fully account for the deep genomic complexity associated with sensitivity or resistance to cancer therapies. To overcome these limitations, large-scale pharmacogenomic screens of cancer cell lines – in conjunction with modern statistical learning approaches - have been used to explore the genetic underpinnings of drug response. While these analyses have demonstrated the ability to infer genetic predictors of compound sensitivity, to date most modeling approaches have been data-driven, i.e. they do not explicitly incorporate domain-specific knowledge (priors) in the process of learning a model. While a purely data-driven approach offers an unbiased perspective of the data – and may yield unexpected or novel insights - this strategy introduces challenges for both model interpretability and accuracy. In this study, we propose a novel prior-incorporated sparse regression model in which the choice of informative predictor sets is carried out by knowledge-driven priors (gene sets) in a stepwise fashion. Under regularization in a linear regression model, our algorithm is able to incorporate prior biological knowledge across the predictive variables thereby improving the interpretability of the final model with no loss – and often an improvement - in predictive performance. We evaluate the performance of our algorithm compared to well-known regularization methods such as LASSO, Ridge and Elastic net regression in the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (Sanger) pharmacogenomics datasets, demonstrating that incorporation of the biological priors selected by our model confers improved predictability and interpretability, despite much fewer predictors, over existing state-of-the-art methods.

CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE “BIG DATA” ERA

PROCEEDINGS PAPER WITH POSTER PRESENTATION

AUTOMATED GENE EXPRESSION PATTERN ANNOTATION IN THE MOUSE BRAIN

Tao Yang[1], Xinlin Zhao[1], Binbin Lin[1], Tao Zeng[2], Shuiwang Ji[2], Jieping Ye[1]

[1]Arizona State University, [2]Old Dominion University

Brain tumor is a fatal central nervous system disease that occurs in around 250,000 people each year globally and it is the second cause of cancer in children. It has been widely acknowledged that genetic factor is one of the significant risk factors for brain cancer. Thus, accurate descriptions of the locations of where the relative genes are active and how these genes express are critical for understanding the pathogenesis of brain tumor and for early detection. The Allen Developing Mouse Brain Atlas is a project on gene expression over the course of mouse brain development stages. Utilizing mouse models allows us to use a relatively homogeneous system to reveal the genetic risk factor of brain cancer. In the Allen atlas, about 435,000 high-resolution spatiotemporal in situ hybridization images have been generated for approximately 2,100 genes and currently the expression patterns over specific brain regions are manually annotated by experts, which does not scale with the continuously expanding collection of images. In this paper, we present an efficient computational approach to perform automated gene expression pattern annotation on brain images. First, the gene expression information in the brain images is captured by invariant features extracted from local image patches. Next, we adopt an augmented sparse coding method, called Stochastic Coordinate Coding, to construct high-level representations. Different pooling methods are then applied to generate gene-level features. To discriminate gene expression patterns at specific brain regions, we employ supervised learning methods to build accurate models for both binary-class and multi-class cases. Random undersampling and majority voting strategies are utilized to deal with the inherently imbalanced class distribution within each annotation task in order to further improve predictive performance. In addition, we propose a novel structure-based multi-label classification approach, which makes use of label hierarchy based on brain ontology during model learning. Extensive experiments have been conducted on the atlas and results show that the proposed approach produces higher annotation accuracy than several baseline methods. Our approach is shown to be robust on both binary-class and multi-class tasks and even with a relatively low training ratio. Our results also show that the use of label hierarchy can significantly improve the annotation accuracy at all brain ontology levels.

CROWDSOURCING AND MINING CROWD DATA

INVITED WORKSHOP WITH POSTER PRESENTATION

CROWDFUNDING WITH THE AMERICAN GUT

Daniel McDonald[1, 2], Justine Debelius[3], Rob Knight[1,3,4], Jeff Leach[5], Adam Robbins-Pianka[1,3], Antonio Gonzalez[3], Amnon Amir[3], Luke Thompson[3], Emily Teravest[3], Greg Humphrey[3]

[1]BioFrontiers Institute, [2]University of Colorado Dept. of Computer Science, [3]University of Colorado Dept. of Chem and Biochem, [4]HHMI, [5]Unaffiliated

American Gut is the largest academic crowdfunded science project, seeking to bring the technologies developed for the Human Microbiome Project to the general public. Previous studies have focused on well-defined cohorts, but there is still much to discover about the kinds of microbiomes that exist “out there in the wild”, which may be missed by projects focused on the ultra-healthy, or on individual diseases. Through the use of crowdfunding, we have successfully raised over $950,000 USD from over 8,000 members of the general public and key donations from industry partners. In exchange for a donation, individuals can choose to receive a sample kit and in return, they receive an analysis of their sample, how they compare to the rest of the population, how they compare to the Human Microbiome Project and a taxonomy summary. Individuals who choose to receive a sample are consented and presented a voluntary survey according to protocols approved by the Institutional Review Board of the University of Colorado at Boulder. All samples in the American Gut Project are processed using the standard Earth Microbiome Project 16S protocol, sequencing the V4 region on the Illumina MiSeq platform. Participant results are generated using a semi-automated pipeline built on top of the Quantitative Insights into Microbial Ecology platform using open-access IPython Notebooks that describe the processing pipeline (https://github.com/biocore/American-Gut/). As of September 2014, we have processed 4,417 samples from 3,624 participants generating over 100 million sequences. The age of the American Gut population covers the age spectrum from infant to elderly. Observed BMI covers underweight to obese, and a wide range of diets, activity levels, medications, and other factors are also covered. Here we present some conclusions from the project to date, including discovery of new kinds of microbes and new kinds of microbiomes in the population, correlations with public-supplied data, and challenges associated with successful large-scale crowdfunded projects.

PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARD IMPROVED MEDICINE

PROCEEDINGS PAPER WITH POSTER PRESENTATION

BAYCLONE: BAYESIAN NONPARAMETRIC INFERENCE OF TUMOR SUBCLONES USING NGS DATA

Subhajit Sengupta[1], Jin Wang[2], Juhee Lee[3], Peter Mueller[4], Kamalakar Gulukota[5], Arunava Banerjee[6], Yuan Ji[1,7]

[1]Center for Biomedical Research Informatics, NorthShore University HealthSystem; [2]Department of Statistics, University of Illinois at Urbana-Champaign; [3]Department of Applied Mathematics and Statistics, University of California Santa Cruz; [4]Department of Mathematics, University of Texas Austin; [5]Center for Molecular Medicine, NorthShore University HealthSystem; [6]Department of Computer & Information Science & Engineering, University of Florida; [7]Center for Biomedical Research Informatics, NorthShore University HealthSystem, Department of Health Studies, The University of Chicago

In this paper, we present a novel feature allocation model to describe tumor heterogeneity (TH) using next-generation sequencing (NGS) data. Taking a Bayesian approach, we extend the Indian buffet process (IBP) to define a class of nonparametric models, the categorical IBP (cIBP). A cIBP takes categorical values to denote homozygous or heterozygous genotypes at each SNV. We define a subclone as a vector of these categorical values, each corresponding to an SNV. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, we took a different approach using feature allocation. Importantly, we do not assume somatic mutations with similar cellular prevalence must be from the same subclone and allow overlapping mutations shared across subclones. We argue that this is closer to the underlying theory of phylogenetic clonal expansion, as somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Bayesian inference yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. We report results on both simulated and real data. BayClone is available at http://health.bsd.uchicago.edu/yji/soft.html.

SPECIAL OCCASION POSTER PRESENTATION

A TWENTIETH ANNIVERSARY TRIBUTE TO PSB

Darla Hewett[1], Michelle Whirl-Carrillo[1], Lawrence E. Hunter[2], Russ B. Altman[1], Teri E. Klein[1]

[1]Stanford University, [2]University of Colorado School of Medicine

PSB brings together top researchers from around the world to exchange research results and address open issues in all aspects of computational biology. PSB 2015 marks the twentieth anniversary of PSB. Reaching a milestone year is an accomplishment well worth celebrating. It is long enough to have seen big changes occur, but recent enough to be relevant for today. As PSB celebrates twenty years of service, we would like to take this opportunity to congratulate the PSB community for your success. We would also like the community to join us in a time of celebration and reflection on this accomplishment.

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA

POSTER PRESENTATIONS

FUSION GENE DETECTION PIPELINE FOR MULTIPLE SAMPLE ON K COMPUTER

Satoshi Ito[1], Yuichi Shiraishi[1], Teppei Shimamura[1,2], Satoru Miyano[1]

[1]Human Genome Center, the Institute of Medical Science, the University of Tokyo;
[2]Division of Systems Biology, Nagoya University Graduate School of Medicine

Fusion genes have an important role in cancer development. Although, recent advances in high-throughput sequencing technologies enabled us to potentially obtain genome-wide landscape of fusion genes, sensitive and accurate identification still remain challenging task because of many artifacts caused by the misaligned reads arising from the ambiguity of genomic sequences, generating numerous false positive detections. From the practical point of view, database of sequenced DNA increase dramatically so that parallel computation is inevitable and its efficiency is one of the most important problems. Shiraishi et al develops Genomon-fusion, which is a fusion gene detection pipeline. Their approach characterizes each gene fusion with a single base resolution by effectively utilizing soft-clipping short reads, reducing false positives by applying a number of filters. In this study, we ported Geomon-fusion onto K computer (GFK), which is the fastest supercomputer in Japan installed in Advanced Institute for Computer Science, RIKEN. Such kind of large-scale supercomputer enable us to analyze very large number of samples in a large databases such as CCLE, TCGA, etc. We performed fusion detection for 780 samples, which is all RNA-seq data in CCLE. Results and calculation summary will be demonstrate.

LANDSCAPE OF REGULATORY MUTATIONS IN CANCER

Collin Melton, Jason Reuter, Damek Spacek, Michael Snyder

Stanford University Department of Genetics

Aberrant regulation of gene expression is a common mechanism underlying the survival and proliferation of cancer cells. Here we integrate TCGA whole genome sequencing data with ENCODE and other regulatory annotations to identify point mutations in regulatory regions in many cancer subtypes. Overall, enrichment analyses demonstrate that mutations are underrepresented in regulatory regions indicating that they are under purifying selection. However, an analysis of the predicted functional effects of observed transcription factor binding site mutations shows selection for several specific mutations that destroy binding sites. Furthermore, using a novel method that adjusts for sample- and genomic locus-specific mutation rate, we identify specific sites that are repeatedly mutated in cancer. Mutated regulatory sites include the known telomerase reverse transcriptase (TERT) promoter and several novel regulatory regions of which a subset are in proximity to known cancer genes. Functional reporter assays demonstrate that two of these novel regions display decreased regulatory activity upon mutation. These data demonstrate that similar to coding regions, regulatory regions contain mutations under selective pressure and suggest a far greater role for regulatory region mutations in cancer than previously appreciated.

MULTI-CANCER MOLECULAR SIGNATURES AND THEIR INTERRELATIONSHIPS

Tai-Hsien Ou Yang, Dimitris Anastassiou

Department of Systems Biology and Department of Electrical Engineering, Columbia University

Multi-cancer molecular signatures and their interrelationships Tai-Hsien Ou Yang and Dimitris Anastassiou Department of Systems Biology and Department of Electrical Engineering, Columbia University We have identified several molecular signatures present in multiple cancer types [1]. These signatures resulted from mining rich data sets provided by The Cancer Genome Atlas (TCGA) Pan-Cancer analysis project, containing values from mRNA expression, microRNA expression, DNA methylation, and protein expression from twelve different cancer types. The membership of these signatures points to particular biological mechanisms related to cancer progression, suggesting that they represent important attributes of cancer in need of being elucidated for potential applications in diagnostic, prognostic and therapeutic products applicable to multiple malignancies. Our data mining approach [2] uses an iterative algorithm to identify patterns that manifest themselves as distinct molecular signatures, several of which were found in nearly identical form following separate analysis of data sets from multiple cancer types, more so than any other computational technique that we tried. The algorithm is designed to converge, in an unconstrained manner, to the core of gene coexpression patterns. Despite the unsupervised nature of the algorithm, we hypothesized that these signatures represent important biomolecular events of cancer in general, and therefore that they would be associated with cancer phenotypes. Consistently, we have demonstrated that some of these signatures, called attractor metagenes, are highly prognostic in breast cancer having being used in the winning model of the Sage Bionetworks/DREAM breast cancer prognosis challenge [3], resulting in a novel biomarker [4]. Examples of such multi-cancer signatures are attractor metagenes associated with mitotic chromosomal instability (CIN), mesenchymal transition (MES), lymphocyte infiltration (LYM), endothelial markers (END), as well as the DLK1-DIO3 RNA cluster. Two DNA methylation signatures are strongly associated with the LYM signature, providing opportunities to decipher the composition of protective infiltration by particular immune cell populations. [1] W.Y. Cheng, T.H. Ou Yang, H. Shen, P.W. Laird, D. Anastassiou and The Cancer Genome Atlas Research Network, "Multi-cancer molecular signatures and their interrelationships," arXiv: 1306.2584v2, 2013 [2] W.Y. Cheng, T.H. Ou Yang and D. Anastassiou, "Biomolecular events in cancer revealed by attractor metagenes," PLoS Computational Biology, Vol. 9, Issue 2, 2013. [3] W.Y. Cheng, T.H. Ou Yang and D. Anastassiou, "Development of a prognostic model for breast cancer survival in an open challenge environment," Science Translational Medicine, Vol. 5, Issue 181, p. 181ra50, 2013 [4] T.S. Ou Yang, W.Y. Cheng, T. Zheng, M.A. Maurer and D. Anastassiou, "Breast Cancer Prognostic Biomarker Using Attractor Metagenes and the FGD3-SUSD3 Metagene," Cancer Epidemiology, Biomarkers & Prevention, 2014, in Press.

USING TOTAL CORRELATION EXPLANATION TO IDENTIFY CANCER THERAPEUTIC TARGETS

Shirley Pepke[1], Greg Ver Steeg[2]

[1]Lyrid LLC, [2]University of Southern California

Ovarian cancer is the most deadly gynecologic malignancy. While new treatments have improved survival rates for other cancers, ovarian cancer is currently treated with the same regimen that has been in place for twenty years. Thus there is an urgent need to identify more effective therapies. Next generation sequence-based assays in principle allow for systematic characterization of individual tumor vulnerabilities, but have yet to be exploited in the context of ovarian cancer. The method of Total Correlation Explanation (CorEx) is a recently developed technique based upon concepts from information theory that is effective in identifying clusters and latent factors in complex data. We apply Total Correlation Explanation to TCGA mRNA-seq data from 420 ovarian tumor samples. Unlike most clustering algorithms, CorEx allows genes to appear in multiple groups, accommodating the multifunctional capacity of proteins in networks. We use CorEx to identify 100 cohorts of genes whose gene expression patterns across ovarian cancer patients exhibit high group mutual information. The resulting fine granularity of the learned subsets allows for relatively specific functional groupings and these are found to be associated with highly significant ontology and pathway annotations. For each discovered group of genes, CorEx constructs a latent factor that attempts to capture a likely common cause. We use conditional factor probabilities to assign each patient a score relative to each gene group. We rank the group scores according to influence on survival using Cox proportional hazard regression analysis. We then use the ovarian cancer factor model to assign group scores and assess prognostic significance based upon mRNA-seq data for basal-like breast tumors and squamous lung cancers. This analysis uncovers one group of genes whose expression pattern has prognostic value for all three cancers. Based upon gene mutation characteristics previously observed for the three cancers and the CorEx gene group annotations, we suggest the common prognostic group is related to compensatory pathway activation that alters response to agents commonly used in treating the three types of cancer. This finding supports the idea that other groupings found by the CorEx training may be useful for the identification of new therapeutic targets specific to ovarian cancer. Finally, we explore ways to incorporate genetic mutation information in addition to the mRNA-seq data while learning the ovarian cancer CorEx model.

ADVANCED ALGORITHMS FOR NEXT GENERATION DATA PROCESSING

Victor Solovyev, Igor Seledtsov, Vladimir Molodsov, Oleg Fokin

King Abdullah University of Science and Technology, Softberry Inc.

Dozens new algorithms have been developed for next-generation sequencing (NGS) data processing and many of them actively applied in the frontier cancer genome research. The collaborative competitions (such as Assemblathon, Alignathon and RGASP) assessed the state of the art in genome assembling, read mapping, discovery alternative transcripts. They demonstrated the lack of consistency between software tools in terms of comparisons obtained on different data sets as well as relative to various metrics evaluating the quality of results suggesting that there is still much room for improvement. We present here advanced suit of robust bioinformatics tools for efficient analysis of large-scale NGS data. It includes 1) OligoZip - de novo NGS reads assembler pipeline; 2) ReadsMap – a tool for RnaSeq spliced and non-spliced reads mapping and SNP identification; 3) TransSeq - program for de novo assembling alternative transcripts from short reads and gene expression quantification; 3) GenomeMatch – a tool to compare genomic sequences. The OligoZip pipeline includes adapters trimming module for PE and MP reads. The new important iterative procedure ReadsClean is developed to remove errors in the read set as well as separate "clean" from "dirty" reads that can not be corrected due to various reasons. The assembling algorithm uses only the set of clean reads. The algorithm consists of creating seed sequences, contig extension and iterative contig joining (scaffolds building) and "holes" patching modules. Using only PE reads the algorithm assembles ~2GB bacterial reads in 30-40 contigs with the average contig length 130K. ReadsMap has been significantly improved in speed of data processing and its accuracy to align RNASeq reads to the reference genome reached Sensitivity 0.99 and Specificity 0.96. TransSeq program that assembles the RNA-Seq data into unique sequences of transcripts and generates full-length transcripts for a set of alternatively spliced isoforms. The program demonstrates Sensitivity 0.97 and Specificity 0.96 in identifying known RNA transcripts on C. elegans test data. We incorporate this program into our Fgenesh++ gene prediction pipeline and demonstrated the benefits of using publicly available RNAseq data of Rice transcriptome to re-annotate the Rice genome chromosome sequences. We produce a a new annotation that includes alternatively spliced isoforms as well as non-coding 5'- and 3'-transcript regions that were absent in ab initio predicted genes. The Rice genome annotation can be download from KAUST Bioinformatics Web server: http://www.molquest.kaust.edu.sa and some of programs available at the Softberry Web server: http://linux1.softberry.com/berry.phtml?topic=fdp.htm.

INTEGRATING OMICS AND HISTOPATHOLOGY PROFILES TO IDENTIFY NOVEL SUBTYPES OF LUNG ADENOCARCINOMA

Kun-Hsing Yu[1,2], Daniel L. Rubin[1,3], Mark F. Berry[4], Michael Snyder[2]

[1]Biomedical Informatics Program, Stanford University; [2]Department of Genetics, Stanford University; [3]Department of Radiology, Stanford University; [4]Department of Cardiothoracic Surgery, Stanford University

Lung cancer causes more than 1.4 million deaths per year, and adenocarcinoma accounts for 40% of lung cancer. Patients with lung adenocarcinoma have diverse clinical outcome, and how omics abnormalities contribute to histopathology findings, which define tumor subtypes, remain largely unknown. In this study, we analyzed transcriptomic and proteomic profiles of lung adenocarcinoma patients (n=480) from The Cancer Genome Atlas (TCGA), distilled qualitative findings from pathology reports through an fully automated text mining pipeline, and extracted quantitative image features from the whole slide images. We divided TCGA patients into distinct training and test sets (70% cases in the training set, and 30% cases in the test set) and selected the top features by forward feature selection from the training set. We utilized supervised machine learning methods, including least absolute shrinkage and selection operator (LASSO), support vector machine, and random forest, to predict the histopathology phenotypes and survival outcomes. Leveraging transcriptomic and proteomic information, we successfully predicted histology grade of tumor with area under receiver operating characteristic curve (AUC) more than 0.88 on the test set, and identified important biological pathways enriched in the most predictive features, such as apoptosis and proteolysis. Incorporating histopathology and omics features, we classified our patients into two distinct survival groups, with statistically significant difference in their survival time ($p<0.0008$). Our results demonstrate that the integration of histopathology and omics studies can reveal molecular mechanisms of pathology findings and discover novel subtypes with clinical relevance. Accurate prediction of patient outcome will contribute to establishing personalized cancer treatment plans, thereby increasing the quality of care and reducing the cost of cancer management. We envision that these analytical approaches described here can be extended to other tumor types and potentially to other medical disciplines as well.

CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE “BIG DATA” ERA

POSTER PRESENTATION

PATHWHIZ: A WEB SERVER FOR HIGH QUALITY PATHWAY GENERATION IN THE BIG DATA ERA

Allison Pon, Timothy Jewison, Michael Wilson, Craig Knox, Fatema M. Disfany, Adam Maciejewski, Yilu Su, David S. Wishart

University of Alberta

Background: Many of today's pathway drawing tools are limited in their ability to generate meaningful, colourful, or biologically rich pathway diagrams. The development of more sophisticated, user-friendly tools is needed to advance the field of pathway generation, visualization, and analysis. Methods: PathWhiz is a newly developed web server designed for the easy creation of colourful, visually pleasing, and biologically accurate pathway diagrams that are machine readable, interactive, and fully web compatible. PathWhiz differs from other pathway drawing tools in that it is a web server rather than a stand-alone program, making it accessible from almost any place and compatible with essentially any operating system. Web accessibility also allows its communal library of pathway components to be constantly expanded by its users. In comparison with most other pathway drawing tools, PathWhiz can support a higher level of biological detail, physiological context, and biological complexity. In particular, PathWhiz, through a specially designed drawing palette, allows the effortless rendering of metabolites (via automated generation of their structures), proteins (including quaternary structures), covalent modifications, cofactors, membranes, subcellular structures, cells, tissues, and organs. Furthermore, PathWhiz has been designed so that pathways can be constructed quickly and intuitively by combining processes such as reactions, interactions, and transports. Pathways can also be automatically propagated across organisms via a simple replication function for existing pathways. PathWhiz supports the generation of BioPAX, SBGN and SBML data, in addition to high resolution PNG, image map, and SVG images.   Results: PathWhiz has already been used to generate more than 700 pathway diagrams for a number of popular databases including HMDB, DrugBank and SMPDB. PathWhiz is available at http://www.smpdb.ca/pathwhiz

CHARACTERIZING THE IMPORTANCE OF ENVIRONMENTAL EXPOSURES, INTERACTIONS BETWEEN THE ENVIRONMENT AND GENETIC ARCHITECTURE, AND GENETIC INTERACTIONS: NEW METHODS FOR UNDERSTANDING THE ETIOLOGY OF COMPLEX TRAITS AND DISEASE

POSTER PRESENTATIONS

DETECTING INTERACTIONS OF GENETIC RISK FACTORS IN DISEASE: A MACHINE LEARNING APPROACH

Patricia Francis-Lyon, Fu Cheng, Chaman Kaur

University of San Francisco

It is well known that specific genes may enhance or suppress susceptibility to disease. When more than one gene influences disease, their effects may be either independent or epistatic. In the latter case, the combined effects of genes on the observed phenotype are not merely additive; this is indicative of biological interaction. These interactions are of particular interest, as their discovery may yield information on protein pathways that are involved in disease, possibly leading to new therapeutic treatments. Here we use a machine learning approaches to detect gene interactions in disease susceptibility. We utilize supervised learning, training the machine to detect disease. We then quantify the effect on prediction accuracy when alleles of two or more genes are perturbed to the unmutated state in patterns that reveal and gene interactions. Here we utilize this approach with a support vector machine. We test the versatility of our approach on data generated to simulate seven biological disease models, some which represent epistasis and some which represent biological independence. We simulate both 2-loci and 3-loci versions of each disease model. In every disease model we correctly detect the presence or absence of gene interactions. This provides evidence that this machine learning approach can be used to successfully detect and characterize gene interactions in disease. We then utilize our method with sporadic breast cancer case-control data and compare our results to those reported in applying the multifactor dimensionality reduction (MDR) method, a leading method for detecting gene interactions. We find our results with this first set of actual data to be promising.

IDENTIFICATION AND REPLICATION OF CONTEXT-SPECIFIC GLOBAL REGULATORY DIFFERENCES BETWEEN OBESE AND LEAN INDIVIDUALS

Arthur Ko, Rita M. Cantor, Bogdan Pasaniuc, Elina Nikkola, Marcus Alvarez, Karen L. Mohlke, Michael Boehnke, Francis S. Collins, Olli Raitakari, Terho Lehtimäki, Ilkka Seppälä, Johanna Kuusisto, Markku Laakso, Päivi Pajukanta

1.Department of Human Genetics, David Geffen School of Medicine at UCLA; 2.Molecular Biology Institute at UCLA; 3.Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA; 4.Bioinformatics Interdepartmental Program, UCLA; 5.Department of Genetics, University of North Carolina, Chapel Hill,; 6.Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor; 7.National Human Genome Research Institute, National Institutes of Health; 8.Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku; 9.Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital; 10.Department of Clinical chemistry, Fimlab Laboratories and University of Tampere School of Medicine; 11.Department of Medicine, University of Eastern Finland and Kuopio University Hospital

Obesity is a highly prevalent risk factor for heart disease, type 2 diabetes, and certain cancers. Although the heritability estimates of obesity are high (40-80%), the variants identified in genome-wide association studies (GWAS) explain only a small portion of the trait variance (<10%). Meanwhile, the constant increase in obesity prevalence during the last 2 decades pinpoints to gene and environment (GxE) interactions in its etiology. However, GxE effects are small and difficult to detect in humans. To address this limitation, we investigated an intermediate cellular phenotype, adipose mRNA expression, since expression profiles will likely reflect the molecular responses or consequences of environmental changes underlying the increase in obesity. Accordingly, the cellular environment may activate regulatory variants and thus alter gene expression differently between obese and lean individuals, resulting in environmental- or context-dependent expression quantitative trait loci (eQTL). To this end, we performed eQTL mapping using 7,932,277 imputed and genotyped SNPs (MAF>1%) with adipose RNA-sequence data on 17,210 expressed genes in ~600 men from the Finnish METSIM cohort. We subdivided the individuals into obese and lean based on the cohort BMI median to ensure equal sample sizes of ~300 in both groups to avoid bias related to sample size. We considered the eQTLs only observed in the obese group, but not in the lean or overall groups as obese-specific (OS), and vice versa for lean-specific (LS). After correcting for multiple testing using an FDR<0.01, we discovered 2,450 genes regulated by 28,267 cis (+/-1Mb) eQTL OS variants; and 1,455 genes regulated by 10,814 cis-eQTL LS variants, respectively. We observed that a tighter regional linkage disequilibrium (LD) in obese versus lean Finns contributes to the larger number of OS versus LS cis-eQTL variants, as 43 of the top 100 OS gene regions with the highest numbers of cis OS eQTLs (FDR<0.01) displayed a significant regional LD difference (P<0.0005 by permutation). Interestingly, many of the OS genes that display extended LD also exhibit a high connectivity in protein-protein networks, suggesting them as regulatory hub genes. For example, dead box helicase 17 (DDX17), a known master-regulator of estrogen and androgen pathways, and non-histone chromosome protein 2-like 1 (NHP2L1), a known regulator of splicing, emerged as hub genes in obese Finns. We replicated 35% of OS eQTLs, including the two master regulators DDX17 and NHP2L1, and 36% of LS eQTLs in the independent Young Finns Study (FDR<0.01), with blood microarray data (n=1,414), suggesting that more than a third of context-specific cis-eQTLs are consistent across tissues and platforms (RNA-seq versus microarray). Taken together, context-specific eQTLs identify global regulatory differences between obese and lean individuals and help reveal the molecular mechanisms underlying obesity.

CROWDSOURCING AND MINING CROWD DATA

POSTER PRESENTATION

VERIFICATION OF BIOLOGICAL NETWORK MODELS USING A COLLABORATIVE REPUTATION-BASED PLATFORM

Jean Binder, Stephanie Boue, Anselmo Di Fabio, R. Brett Fields, William Hayes, Julia Hoeng, Jennifer S. Park, Manuel C. Peitsch, Walter Schlage, Marja Talikka

Philip Morris International R&D, Philip Morris International R&D, Applied Dynamic Solutions, Selventa, Selventa, Philip Morris International R&D, Selventa, Philip Morris International R&D, Philip Morris International R&D, Philip Morris International R&D

Biological network models are an important tool in drug development and toxicity research to gain a mechanistic understanding of the effects of compounds on human health. The comprehensiveness of these network models is important to capture all possible key downstream effectors linking compounds to specific biological pathways. A current limitation in the use of biological network models is the lack of an easy way to update, collaborate and share these networks. We have created a web-based crowdsourcing platform to facilitate collaboration on 49 networks capturing a wide range of biological processes as part of the sbvIMPROVER Network Verification Challenge (NVC). These networks consist of causal statements in the Biological Expression Language (BEL) based on scientific publications, with their species, disease and tissue context captured in a computable format. The goal of the NVC is for the scientific community to vote, comment and add new biology to improve these networks. By implementing a reputation-based web application available to the entire scientific community to add biology and vote on supporting scientific evidence, the NVC has created a collaborative crowdsourcing platform to complement peer review of publications describing the networks and to ensure complete and accurate biological networks that can be used as a standard in the field. The first NVC resulted in the improvement of the networks with over 2000 votes cast and 800 literature evidences added to the website (http://bionet.sbvimprover.com). Networks continue to be improved during the second NVC (NVC2) with a long-term goal of making them the most up-to-date and comprehensive networks that can be used by the scientific community to download, utilize and continue to refine for use in toxicological and drug discovery applications.

PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARD IMPROVED MEDICINE

POSTER PRESENTATIONS

ASCERTAINMENT BIAS IN BREAST CANCER RISK ASSESSMENT BASED ON FAMILY HISTORY

Tinhinan Belaribi[1], Flora Alarcon[2], Antoine De Pauw[3], Nadine Andrieu[4,3], Marie-Gabrielle Dondon[4,3], Dominique Stoppa-Lyonnet[3,], Grégory Nuel[6,1]

[1]Department of Probability (LPMA, UMR 7599), Sorbonne Universités; [2]Department of Applied Mathematics (MAP5, UMR 8145), Sorbonne Paris Cité; [3]Institut Curie; [4]National Center for Medical Research (INSERM, U900); [5]Mines ParisTech, [6]Institute for Mathematics (INSMI), National Center for Scientific Research (CNRS)

Introduction: It is now well known that mutations in BRCA1 and BRCA2 genes are responsible for monogenic form of breast and ovarian cancer and genetic testing for this two genes is now a standard procedure for patients with severe family history (FH). However, mutation in BRCA genes don't explain all familial cases. For example, at the Institut Curie, it is estimated that only 3% of women with breast cancer and 10% of women with ovarian cancer before age 70 carry a BRCA mutation. From there (or since 1989), different models taking into account patients FH have been developed to assess their cancer risks which allow to establish the decision thresholds needed by clinicians to define prevention strategies. Among them BOADICEA [1], used at Institut Curie, seems to be the most attractive one because of an important number of risk factors taken into account, including a polygenic component. This model allows to calculate the risk for an unaffected patient to be diagnosed with breast cancer or ovarian cancer given his FH. Unfortunately, a recent retrospective study from Institute Curie's family cancer data [2] shows that BOADICEA has a poor predictive power (especially, it under-estimates the cancer risk in families without a BRCA1/2 mutation). The complex problem of ascertainment bias could probably explain a part of this poor predictive power. Indeed, Institute Curies family data are obtained not randomly in the population but through several criteria in order to increase the probability to have mutation carriers. Method: Our initial work in collaboration with the Institut Curie aims to study, through simula- tions, the bias introduced by the ascertainment on the predictive risk calculation, using standard estimation method. Then, we compare and discuss various standard ascertainment correction meth- ods, including that used on BOADICEA. Finally, we propose an ascertainment correction in order to have unbiased risk. Results: Results show an overestimation of the risk in absence of correction for the ascertainment process. On the other hand, the risk is underestimated when ascertainment bias is corrected like in BOADICEA. This underestimation could be explained by the fact that the likelihood is con- ditioned only by phenotypes. Indeed, this correction is not adjusted to the Institut Curies family data for which recruitment criterion may also depend on genotypes. Hence, we propose a Prospec- tive Likelihood correction [3], which consists in conditioning the likelihood by the ascertainment event. Further work includes clinical investigation on the exact nature of ascertainment process and statistical analysis of correction robustness to misspecified ascertainment. References [1] A. Antoniou et al., British journal of cancer 98, 1457 (2008). [2] A. de Pauw et al., Bulletin du cancer 96, 979 (2009). [3] F. Alarcon et al., Genetic epidemiology 33, 379 (2009).

FINDINGS FROM THE THIRD CRITICAL ASSESSMENT OF GENOME INTERPRETATION (CAGI), A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Steven E. Brenner[1], John Moult[2], CAGI Participants

[1]University of California, Berkeley; [2]IBBR, University of Maryland

The Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In the experiment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings. The third CAGI experiment (concluded in July 2013) consisted of ten diverse challenges. CAGI deliberately extends challenges from previous years, with the continuity allowing measurement of progress. For example, in the second CAGI, in a challenge to predict Crohn's disease from exomes, one group was able to identify 80% of affected individuals before the first false positive healthy person. In the third CAGI experiment, this challenge used an improved dataset, and several groups performed remarkably well, with one group achieving a ROC AUC of 0.94. The experiment also revealed important population structure to Crohn's disease in Germany. For three years, CAGI has posed a challenge with Personal Genome Project (PGP) genome data. This year, two groups were able to successfully map a significant number of complete genomes to their corresponding trait profiles submitted by PGP participants. In the expanded challenge to predict benign versus deleterious variants in DNA double-strand break repair MRN genes—Rad50 (from last year), Mre11, and Nbs1—as determined by those that appear in a breast cancer case versus healthy control, predictions show how methods differ sharply in their effectiveness even amongst proteins in the same complex. A new challenge in 2013 year was to use exomes from families with lipid metabolism disorders. In the case of hypoalphalipoproteinemia (HA), a company made predictions which showed how understanding the problem structure and employing an extensive knowledgebase led to remarkably good results. Another related challenge revealed a twist wherein real-world data differed sharply from theoretical models. The other challenges were to predict which variants of BRCA1 and BRCA2 are associated with increased risk of breast cancer; to predict how variants in p53 gene exons affect mRNA splicing; to predict how well variants of a p16 tumor suppressor protein inhibit cell proliferation; and to identify potential causative SNPs in disease-associated loci. Overall, CAGI revealed that the phenotype prediction methods embody a rich and diverse representation of biological knowledge, and they are able to make predictions that are highly statistically significant. However, we also found the accuracy of prediction on the phenotypic impact of any specific variant was unsatisfactory and of questionable clinical utility. The most effective predictions came from methods honed to the precise challenge, including the specific genes of interest as well as the problem context. Prediction methods are clearly growing in sophistication, yet there are extensive opportunities for further progress. Complete information may be found at http://genomeinterpretation.org.

CLINICAL SEQUENCING IS JUST LIKE RESEARCH… OR MAYBE NOT SO MUCH

Stephen Lincoln[1], Allison Kurian[2], Andrea Desmond[3], Yuya Kobayashi[1], Michael Anderson[1], Geoffrey Nilsen[1], Shan Yang[1], Martin Powers[1], Swaroop Aradhya[1], Leif Ellisen[3], James Ford[2]

[1]Invitae, [2]Stanford University, [3]Massachusetts General Hospital

Many clinical laboratories now offer NGS panel and exome tests. While clearly successful, these tests raise serious questions that slow adoption. Thankfully these questions have answers with data behind them. We present a few surrounding hereditary cancer testing, a large market with clear near-term patient benefits. 1. Is more data better? To clinicians, the answer is often "no" unless they know how to act on it. We completed a study of over 1000 patients undergoing BRCA1/2 testing and showed that 4% of the BRCA-negative patients are positive for another gene which may explain their personal or familial cancer. As most of these findings are in moderate penetrance genes, the interventions for BRCA1/2 (prophylactic surgery and chemotherapy) are usually not appropriate, and indeed care guidelines do not yet exist for many of these genes. Nevertheless we find that 70% of these findings indeed would warrant consideration of a change in care even under current guidelines. 2. Is NGS accurate enough? Certain classes of variants are challenging (CNVs, large indels, complex alterations and variants in difficult genomic contexts). These are a tiny fraction (<0.1%) of the coding variants in any individual, allowing high sensitivity to be achieved in research NGS without "worrying about them". However, 14% of the clinically actionable variants are of these challenging classes, and we have seen examples of each type. Using a battery of algorithms we were able show 100% sensitivity for these by comparison with traditional technologies. 3. What about false positives? High sensitivity can come at a cost of additional false positives, and a standard practice in clinical NGS is to confirm actionable variants using an orthogonal technique. The impact on logistics, costs and turn around time is high. Techniques to improve this workflow exist, and our data also shows that such confirmation may not always be necessary. Some classes of variants ("pearly whites") have strong NGS data and a track record of correct calls and may not need confirmation. Importantly, read-depth is not the key parameter in identifying these. 4. What about Interpretation? Determining the pathogenicity of variants to clinical standards is critical – consider whether you yourself would have surgery based on a test result. Even after 2 years of testing, we still see about 0.7 new, uninterpreted, rare variants per patient in hereditary cancer genes alone. Interpretation is not fully automatable but certainly can be scaled with software. Genetic counseling practices are also adapting and these have been shown to be clinically feasible and appreciated by patients. In summary, challenges remain for clinical NGS but solutions exist to best benefit patients and their caregivers. If you're nice to me, I promise not to talk about regulatory issues or reimbursement.

A CAUTIONARY TALE FOR GENOMIC MEDICINE: POPULATION DIVERSITY AND THE GENETICS OF HYPERTROPHIC CARDIOMYOPATHY

Arjun K. Manrai[1], Birgit Funke[2], Heidi Rehm[2], Morten Olesen[3], Bradley Maron[4], Peter Szolovits[5], David Margulies[6], Joseph Loscalzo[4], Isaac Kohane[6]

[1]Harvard Medical School, [2]Partners HealthCare Laboratory for Molecular Medicine, [3]University of Copenhagan, [4]Brigham and Women's Hospital, [5]MIT, [6]Harvard Medical School

BACKGROUND—Risk stratification for hypertrophic cardiomyopathy (HCM) is an exemplar of the clinical gains attainable by targeted genetic testing. Using sequencing results, clinicians routinely assess risk for the patient's relatives and even tailor therapy for rare patients. However, the benefits of genetic testing come with the risk that variants may be misclassified. METHODS—Using publicly accessible exome data, we identified variants previously considered causal of HCM that were overrepresented in the general population. We studied these variants in diverse populations, and reevaluated their initial ascertainments in the medical literature. We reviewed patient records at a leading genetic testing laboratory for variant occurrences during the near decade-long history of the laboratory. RESULTS—Multiple patients, all of African or unspecified ancestry, received positive reports with variants initially classified as pathogenic and later changed to benign. All studied high-frequency variants were significantly more common in African Americans than European Americans ($P < 0.001$). If diverse control sequencing data had been available, these variants would likely have been classified earlier as benign, possibly avoiding multiple misclassifications in African-ancestry individuals. We identify methodological shortcomings that may have led to these errors in the medical literature. CONCLUSIONS—These findings highlight the value of diverse population sequencing data, which can prevent variant misclassifications by identifying ancestry informative yet clinically uninformative markers. These findings expand upon current guidelines, which recommend using ethnically matched controls to interpret variants. As diverse sequencing data become more widely available, we expect variant reclassifications to increase, particularly for ancestry groups that have historically been less well studied.

AUTOMATED MOLECULAR PHENOTYPING USING BAYESIL

Siamak Ravanbakhsh, Philip Liu, Trent C. Bjorndahl, Rupasri Mandal, Jason R. Grant, Michael Wilson, Roman Eisner, Igor Sinelnikov, Xiaoyu Hu, Claudio Luchinat, Russell Greiner, Tamara Lim, David S. Wishart

University of Alberta

Molecular phenotype measurements are essential to the characterization of complex diseases, notably, many diseases cause significant changes to the concentrations of small molecules (a.k.a. metabolites). Because of this, the field of metabolomics is suited for molecular phenotypic measurements. However metabolomics is a relatively low throughput, manually intensive, error prone process that requires trained experts to analyze and process the data. A system that can quickly, accurately and automatically produce a person's metabolic profile would enable efficient and reliable molecular phenotype characterization which could significantly improve the way personalized medicine is practiced. This poster presents such a system: Given a 1D 1H NMR spectrum of a complex biofluid such as serum or CSF, our Bayesil system can automatically determine this metabolic profile, and do so without any human guidance. This requires first performing all of the required spectral processing steps (i.e., Fourier transformation, phasing, solvent-removal, chemical shift referencing, baseline correction, and line shape convolution) then matching this resulting spectrum against a reference compound library, which contains the signatures of each relevant metabolite. Many of these processing steps are novel algorithms, and our matching step views spectral matching as an inference problem within a probabilistic graphical model that rapidly approximates the most probable metabolic profile. Our extensive studies on a diverse set of complex mixtures (real biological samples, defined mixtures and realistic computer generated spectra), show that Bayesil can perform complete automation of the spectral processing but also fully automated spectral deconvolution by fitting our growing “metabolic library”, which now exceeds 220 compounds. Bayesil consistently performs with sensitivity and specificity greater than 98% for compound identification in mixtures with over 60 different compounds. It also determines metabolite concentrations (down to 10 µM for spectra with reasonable signal to noise levels) within 10% of the known or expert-measured concentrations in fewer than 5 minutes on a single CPU processor. These results demonstrate that Bayesil is the first fully-automatic publicly-accessible system that provides quantitative NMR spectral profiling effectively – with an accuracy that meets or exceeds the performance of highly trained human experts. The high-throughput efficiency of Bayesil may open the door to a wealth of metabolomic applications of NMR for molecular phenotype measurements. In a clinical setting, Baysil could dramatically reduce costly analysis times helping to democratize metabolomics for personalized medicine. A detailed description of the software, its algorithms, its spectral databases and its testing performance will be presented. Users can access Bayesil at http://www.bayesil.ca. 82

FUNCTIONAL IMPACT SNP SET ENRICHMENT ANALYSIS (FISSEA)

Erick R. Scott, Adam Mark, Ali Torkamani, Chunlei Wu, Andrew I. Su

The Scripps Research Institute

The 20th century witnessed the practice and science of medicine organized into organ-focused disciplines, while the biological sciences has organized along cellular pathways and pathogenic mechanisms. With the advent of cost-efficient and high throughput molecular phenotyping assays (e.g. whole genome sequencing, shotgun proteomics, and metabolomics) calls to re-organize medicine by molecular signatures have proliferated. Ultimately, methods that bridge the divide between molecular signatures and organ-based medicine will be needed to facilitate precision and personalized medicine. We hypothesize that tissues with a greater burden of deleterious DNA variants will lead to impaired tissue function and ultimately disease. We test this hypothesis using functional impact SNP Set Enrichment (fiSSEA) which seeks to integrate genetic susceptibility with tissue-specific gene function. In contrast to traditional SNP Set Enrichment Analysis, which prioritizes candidate SNPs from genome-wide association studies using p-values, we leverage functional impact predictions. Specifically, for each gene in an individual's genome we transform nonsynonymous single nucleotide polymorphisms (SNPs) into functional impact prediction scores, e.g. CADD scores. We then sum the functional impact prediction scores for each protein-coding genotype and calculate the median genotype score for all coding variants in a given gene, yielding a gene-specific median functional impact score. In order to compare the burden of deleterious SNPs across tissues, we calculate unweighted Kolmogorov-Smirnov statistics using the Gene Set Enrichment Analysis (GSEA) package with tissue-specific gene sets (Enrichment Profiler). Using publicly available whole genome sequences we find preliminary evidence of an increased burden of variants with predicted deleterious functional impact in tissues of disease relevance. For example, a recent N-of-1 study identified genetic susceptibility to type-II diabetes prior to a clinical diagnosis. We find fiSSEA identifies an increased burden of predicted deleterious SNPs in two type-II diabetes associated tissue gene sets, adipose and skeletal muscle. We find a weaker enrichment of deleterious variants in adipose and skeletal muscle gene sets in two out of four additional fiSSEA analyzed genomes. We have released fiSSEA as an open source package (https://github.com/SuLab/fiSSEA).

PHARMGKB: DRUG LABEL CURATION TOOL

Michelle Whirl-Carrillo, Julia Barbarino, Ellen McDonagh, Katrin Sangkuhl, Ryan Whaley, Russ B. Altman, Teri E. Klein

Stanford University

The Pharmacogenomics Knowledge Base, PharmGKB, is a publically available online knowledge resource that contains associations between genetic variation and drug response. Information ranges from basic genotype-phenotype relationships to clinically implemented dosing guidelines. PharmGKB curates drug labels that have been identified to contain pharmacogenomics or gene marker information (gene, genetic variant, protein or chromosomal information) using a web-based annotation tool. The intention of annotated drug labels on PharmGKB is to provide users with a quick overview of the pharmacogenomic information found on the label, the genes and phenotypes involved, and how strongly the label recommends testing or not. Explanation of how labels are selected for annotation is found here: https://preview.pharmgkb.org/view/drug-labels.do. PharmGKB scientific curators write a short 1-2 sentence summary of the pharmacogenomics relevance of the label, along with a more extensive review of the information in the label including one or more supporting quotes which represent the PGx implications in the label. Drug labels are annotated with: • the agency which approved the label (US Food and Drug Administration [FDA], European Medicines Agency [EMA] or Pharmaceuticals and Medical Devices Agency, Japan [PDMA]). • the main genes/proteins discussed in the label. • whether the label is on the FDA's "Table of Pharmacogenomic Biomarkers in Drug Labels" (http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm) (FDA-approved labels only). • all genes and phenotypes found in the label, grouped by label heading. • the level of action recommended based on the information found in the label. Perhaps the most useful annotation on the label is the most subjective – the level of action recommended based on the label information. Vague wording on labels can leave practitioners wondering what to do with the provided genetic information when prescribing the drug. PharmGKB curators assess the label information and assign one of the following tags: genetic testing required, genetic testing recommended, actionable PGx and informative PGx. PharmGKB provides language to explain how the level of PGx information on labels in interpreted, with criteria for each category (see https://www.pharmgkb.org/page/drugLabelLegend). These definitions may change over time as new label wording comes up and definitions need to adjust for these new cases. (Work supported by NIH grant R24 GM61374.)

GENERAL POSTER PRESENTATIONS

METAGENERANKER – A WEB-BASED TOOL FOR GENE PRIORITIZATION USING GENE EXPRESSION PROFILES, SNP GENOTYOPES, AND EQTL DATA

Jingmin Che, Jinwoo Kim, Hyeonjung Lee, Miyoung Shin

School of Electronics Engineering/ Kyungpook National University

For the better understanding of disease pathogenesis, we often need to discover significant genes actively involved in human disease. Many earlier approaches have attempted to do so by prioritizing candidate genes based on gene expression profiles or SNP genotype data, but they suffer from producing many false-positives results. MetaGeneRanker is a web-based tool for gene prioritization that employs three genetic resources, including gene expression data, SNP genotype data, and expression quantitative trait loci (eQTL) data. In particular, this tool implements our meta-analysis strategy for gene prioritization in which an improved technique for the order of preference by similarity to ideal solution (TOPSIS) algorithm is utilized to combine gene significance scores from distinct resources. We evaluated this method with two datasets regarding prostate cancer and lung cancer for the identification of disease-related genes. As results, our strategy for gene prioritization showed its superiority to conventional methods in finding out significant disease-related genes by using several distinct genetic resources, while taking good advantage of potential complementarities among available resources. For the use of MetaGeneRanker, one should input both gene expression profiles and SNP genotype data. Here the input data can be the preprocessed microarray data or any score data obtained from the both resources. As the output of this tool, the users can have a list of prioritized genes along with their TOPSIS scores. This tool is freely available online. *Correspondence should be addressed to Miyoung Shin (email: shinmy@knu.ac.kr).

DSPLICETYPE: A GENERALIZED FRAMEWORK FOR DETECTING DIFFERENTIAL SPLICING AND DIFFERENTIAL EXPRESSION EVENTS USING RNA-SEQ

Nan Deng[1,2], Dongxiao Zhu[2], Changxin Bai[2]

[1]Cedars-Sinai Medical Center, [2]Wayne State University

Alternative splicing plays an important role in regulating gene expression and activities in higher eukaryotes. Aberrant differential splicing events with or without gene-level differential expression have been reported to link a number of human disease. The high-throughput RNA-Seq technology holds a better promise to interrogate transcriptomes systematically in terms of differential splicing and differential expression. We present a generalized framework to investigate the synergistic and antagonistic effects of differential splicing and differential expression. The proposed computational method, dSpliceType, detects most common types of differential splicing events with or without differential expression between two conditions using RNA-Seq. In particular, the multivariate dSpliceType is among the fist to utilize sequential dependency of normalized base-wise read coverage signals and capture biological variability among replicates using a multivariate statistical model. We applied the method to a public RNA-Seq data set and compared the transcriptomes between H1 (human embryonic stem cell) and H1 differentiated neuronal progenitor cultured cell lines. dSpliceType detected a large amount of differential splicing events in different types from the two cell lines, and we present several case studies from the analysis results.

PREDICTING INTERFACES IN RNA-PROTEIN COMPLEXES AND BINDING PARTNERS IN RNA-PROTEIN INTERACTION NETWORKS

Rasna R. Walia[1], Carla M. Mann[2], Li C. Xue[3], Usha K. Muppirala[1,2], Benjamin A. Lewis[2], Vasant G. Honavar[4], Drena Dobbs[1,2]

[1]Bioinformatics & Computational Biology Program, Iowa State University; [2]Dept. of Genetics, Development & Cell Biology, Iowa State University; [3]College of Information Sciences and Technology, The Huck Institutes of the Life Sciences, Pennsylvania State University; [4]Dept. of Computer Science, Genome Informatics Facility;

The importance of RNA-protein interactions in viral replication as well as in post-transcriptional and epigenetic regulation of cellular gene expression suggests they should be promising potential targets for intervening in both infectious and genetic diseases. We have used a combination of computational and experimental approaches to interrogate RNA-protein interactions, with two major goals: i) to identify determinants of recognition specificity in RNA-protein complexes; and ii) to understand how networks of RNA-protein interactions are regulated and integrated into cellular regulatory and signalling networks. Over the past decade, we have developed several databases and computational tools for analyzing RNA-protein complexes and interaction networks. These include two databases, PRIDB (1), a database of interfaces from all structurally characterized protein-RNA complexes, and RPIntDB (2), a database of experimentally-validated RNA-protein interactions; and two webservers, RNABindRPlus (3), for predicting interfaces in RNA-protein complexes and RPISeq (4), for predicting partners in RNA-protein interaction networks. In our studies on the "interface prediction" problem, we have demonstrated that machine learning classifiers that use PSSM-based encodings of protein sequences consistently out-perform classifiers that use other sequence-derived encodings. Surprisingly, RNABindRPlus, a sequence-based ensemble method that combines an optimized SVM classifier with a sequence homology-based classifier, also out-performs available structure-based methods. On an independent test dataset of 44 proteins, RNABindRPlus predicts interfacial residues with Specificity of 0.72, Sensitivity of 0.63, and MCC of 0.55. To address the "partner prediction" problem, we have developed a sequence-based method, RPISeq that uses a Random Forest (RF) classifier to predict whether a given pair of RNA and protein sequences interacts. On two non-redundant benchmark datasets extracted from PRIDB, RPISeq-RF classifiers achieved accuracies of 89.6% and 76.2%, with ROC AUCs of 0.96 and 0.92. More recently, we have exploited a compendium of bipartite RNA-protein interfacial motifs to further improve interface predictions and to dramatically reduce the number of "false positive" partner predictions in RNA-protein interaction networks. Taken together, our results indicate that computational tools can be reliable enough to identify key RNA-binding residues for targeted mutagenesis and to identify likely partners for specific RNAs or proteins of biomedical importance. We will present examples of results obtained using these approaches to analyse and predict interfacial residues in clinically important RNPs (e.g., Rev-RRE complexes in HIV-1 and other retroviruses) and to identify interaction partners in cellular signalling networks (e.g., ncRNA-protein interaction networks implicated in cancer). > References 1. PRIDB: http://pridb.gdcb.iastate.edu 2. RPIntDB: http://pridb.gdcb.iastate.edu/RPISeq/RPIntDB.html 3. RNABindRPlus: http://einstein.cs.iastate.edu/RNABindRPlus 4. RPISeq: http://pridb.gdcb.iastate.edu/RPISeq

INTRINSICALLY DISORDERED PROTEINS AND THE EVOLUTION OF MULTICELLULAR ORGANISMS

A. Keith Dunker[1], Sarah E. Bondos[2], Fei Huang[1], Christopher J. Oldfield[1]

[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine; [2]Department of Molecular and Cellular Medicine, Texas A & M Health Science Center

An intrinsically disordered protein (IDP)-based developmental toolkit, which allows functional diversification and environmental responsiveness for molecules that direct the development of complex metazoans, is proposed based on several lines of evidence. IDPs and IDP regions lack an intrinsic equilibrium tertiary, and instead have structures that vary over both time and the population. Despite this lack of structure, IDPs perform many molecular functions – e.g. protein binding, nucleic acid binding, and entropic functions such as providing flexible linkers for tethered searches. These molecular functions underlie numerous biological processes such as transcription regulation, DNA condensation, cell-cell communication, cell adhesion, cell division, and cellular differentiation. These are some of the same biological processes that are vital to multicellular organisms, and in fact the proteomes of multicellular eukaryotic organisms are substantially enriched in IDPs compared to the proteomes of Bacteria and Archaea. To investigate the relationship between intrinsic disorder and multicellularity, we examined the role of IDPs in example proteins from the five biological processes specific to multicellular organisms, namely: (1) cellular adhesion, (2) intercellular communication, (3) developmental pathways, (4) regulation of developmental programs, and (5) cell type-specific biochemistry. We find that in each case examined, intrinsic disorder plays a central role. In particular, the functional repertoire of IDPs is greatly expanded by modulation via post-translational modifications (PTMs), which modulate the binding properties of IDPs, and tissue-specific alternative splicing (AS) of the pre-mRNA coding for IDPs, which utilize splice variants within IDP regions to rewire the protein and gene interaction networks in a cell-type-specific manner. We propose that the combination of IDPs, PTMs, and AS – the IDP-PTM-AS developmental toolkit – has contributed to independent evolution of multicellularity in several different clades. Evidence in support of this hypothesis is particularly strong for the metazoans and land plants. Evidence is accumulating for the fungi, but is currently lacking for the red and brown algae multicellular clades. We further propose that a simplified IDP-PTM developmental toolkit was crucial for the evolution of the simple multicellular organisms that have arisen among the prokaryotes.

COMPUTATIONAL TECHNIQUES FOR STUDYING DIFFERENTIAL TRANSCRIPTION AND VARIANTS OF THE MEDIATOR COMPLEX

Suzanne Renick Gallagher, May Alhazzani, Leslie Seitz, Debra S. Goldberg

University of Colorado

While the general process of gene transcription is well understood, the mechanisms by which different genes are activated in different conditions or different cell types are not. Transcription must be precisely controlled for proper development and response to differing conditions. The Mediator protein complex is essential for most transcription in eukaryotes, and seems to have a role in differential transcription. CDK8 and CDK19 are homologous proteins that function similarly, alternatively occupying the same position in the CDK module of Mediator. We wish to identify the functional differences between CDK8 and CDK19 by considering how the presence of each one impacts the transcriptional program that Mediator helps regulate. Towards this end, we have studied differences in gene expression associated with the presence of CDK8 and CDK19 under various stress conditions. We have developed methods to predict the transcription factors (TFs) that play a role in this differential gene expression both directly and indirectly. Many TFs are implicated with both CDK8 and CDK19, so we have also identified TFs that are significantly more associated with one than the other. We have also looked at hypergraph techniques for studying affinity purification data protein interaction data that uses Mediator subunits as the bait proteins in order to find other proteins more closely associated with a particular subunit than its homologs. While our research has focused on the Mediator complex and the CDK8/CDK19 homolog pair, we believe that the techniques developed will be applicable more broadly to the problem of differential transcription.

ESTIMATING HERITABILITY FOR PATHWAYS USING MIXED LINEAR MODEL ANALYSIS

Jacob B. Hall[1], Jonathan L. Haines[2], William S. Bush[2]

[1]Center for Human Genetics Research, Vanderbilt University; [2]Institute for Computational Biology, Case Western Reserve University

Background: Many methods for genetic pathway analysis exist and typically use gene or SNP p-values to calculate a rank-based pathway statistic. Such methods fail to associate a pathway with a trait if little or no main effects exist within that pathway, so a method that incorporates cumulative effects of SNPs within a pathway, regardless of effect size, would provide a useful measure of a pathway's significance. In this study we propose a versatile and easy-to-implement pipeline for estimating heritability and significance of pathways by using a mixed linear model (MLM) approach. Methods: To demonstrate the validity of our method we applied the following steps to a case-control age-related macular degeneration (AMD) dataset. The first and most flexible step is to define pathways. In our example we manually selected a list of eight terms with plausible relevance to AMD, accessing the Gene Ontology (GO) to determine pathway gene lists. The second step is to partition regions around pathway genes to be analyzed. We assessed three ways of partitioning, taking gene-flanking regions into consideration and also using ENCODE data to select more distant, but plausibly relevant regulatory SNPs to include for analysis. The third step is to estimate the heritability for each pathway using Genome-wide Complex Trait Analysis (GCTA) by fitting genetic relationship matrices (GRMs) in a MLM via the restricted maximum likelihood (REML) method. The final step is to assess significance and perform secondary analyses, such as testing SNP overlap between pathways. Results: For an AMD dataset consisting of 1,145 cases and 668 controls we selected eight pathways to test for association to risk for AMD: angiogenesis, antioxidant activity, apoptosis, the complement system, inflammatory response, nicotine, oxidative phosphorylation, and the tricarboxylic acid cycle (TCA). Some pathways, such as complement, have well-known AMD association. Others, such as TCA only have plausible association based on biological mechanisms cited in existing literature. Only two pathways explained a statistically significant amount of risk for AMD—complement and inflammatory response. Secondary analyses revealed, however, that the majority of the risk explained by the inflammatory response pathway was due to overlap with complement genes, emphasizing the need to assess overlap between pathways when present. Conclusions: Here, we show that our proposed pipeline was able to effectively estimate heritability for AMD and was able to confirm known risk pathways, as well as rank pathways based on the amount of AMD risk explained. We also note that analysis of SNP overlap can clarify whether or not related pathways share a common trait-associated genetic component, thus potentially identifying a shared biological basis between pathways. The pipeline is flexible in how pathways are defined and can be easily applied to both quantitative traits as well as binary/case-control datasets.

NETRANKER: A NETWORK-BASED GENE RANKING TOOL USING PPI AND GENE EXPRESSION DATA

Erkhembayar Jadamba[1], Mingyu Park[2], Miyoung Shin[2]

[1]Bio-Intelligence & Data Mining Laboratory, Graduate School of Electrical Engineering and Computer Science, Kyungpook National University; [2]School of Electronics Engineering, Kyungpook National University

Over the past years, gene prioritization methods have been mainly established based on gene expression profiles and/or PPI (protein-protein interaction) data, occasionally with other information like functional annotations. More recently, biological networks (e.g., PPI networks) are often explored to discover disease-related genes via the guilt-by-association principle over the network topology, outperforming earlier approaches. Here we introduce a web-based tool, NetRanker, that allows users to prioritize candidate genes based on PPI networks and gene expression data. In particular, it provides three different network-based ranking algorithms, including PageRank, Heat Diffusion Rank, and HITs algorithms, each of which is applicable to prioritize genes over STRING or PINA. Gene expression profiles are used to initialize the weights of nodes in a PPI network, so NetRanker can investigate the differential expression of their neighborhoods over the PPI network for gene prioritization. For the use of NetRanker, one should input gene expression profiles in raw cel files or a list of genes (in gene symbols) to be prioritized, along with the specification of a PPI network to be employed. For comparisons with other existing tools, we performed some experiments to identify disease-related genes, and observed that NetRanker can show better performance in detecting true disease genes than existing approaches. NetRanker is freely available for all users interactively via a web application. Correspondence to shinmy@knu.ac.kr

GENOME LEVEL HETEROGENEITY ANALYSIS BY WHOLE EXOME AND TRANSCRIPTOME SEQUENCING OF KIDNEY AND LIVER CANCER

Je-Gun Joung, Joon Seol Bae, Park, Woong-Yang

Samsung Genome Institute

Intratumor heterogeneity has been well known that it makes difficult to make personalized treatments. Result from a single tumor-biopsy sample is not enough to find tumor-driving mutations and to quantify activities of molecular components. To examine intratumor heterogeneity, we performed exome sequencing and RNA sequencing on multiple separated samples obtained from kidney cancer and liver cancer. We characterized tumor heterogeneity through comparative analysis between single biopsy at four regions and pooled biopsies in somatic mutations and gene expression profiles. Six and eight known somatic mutation sites were detected from kidney and liver cancer, respectively. Mutation analysis by whole exome sequencing revealed that overall consistence was observed between single regional biopsies as well as pooled biopsies in mutation level, but apparent discrepancy was presented in particular mutations including CTNNB1 (c.G94C) and AR (c.T170A) sites of liver cancer. Transcriptome profiles between pooled samples have higher correlation than between single regional biopsies. It suggests that pooled samples could compensate for inconsistent measurements on individual biopsy due to the intratumor heterogeneity. The intratumor heterogeneity containing in both kidney and liver cancers indicates that a single biopsy might not be sufficient to determine individualized cancer therapy. Pooling biopsies from a single tumor may be served as a tool in better assessing to obtain sufficient diagnostic decision.

METABOLOMICS OPERATIONS WITHIN THE HUMANCYC METABOLIC DATABASE

Peter D. Karp, Richard Billington, Timothy A. Holland, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Suzanne Paley

SRI International

HumanCyc is a curated metabolic database for humans that describes 288 metabolic pathways and 2312 reactions. Metabolic pathways provide a valuable framework for analyzing metabolomics data. The HumanCyc.org website provides a number of computational operations for metabolomics data analysis. Users can search for metabolites using multiple criteria including monoisotopic molecular weight, chemical formula, and InChI string. The site provides translation of metabolite identifers across multiple metabolite databases. The site enables researchers to store and manipulate metabolite lists using a facility called SmartTables. SmartTables support a number of operations including transformations (such as transforming a metabolite list to a list of all pathways in which those metabolites are substrates), and metabolite enrichment analysis. That analysis operation identifies metabolite sets that are statistically over-represented for the substrates of specific metabolic pathways. HumanCyc.org also enables visualization of metabolomics data on individual pathway diagrams and on the full human metabolic map diagram. Most of these operations are available both interactively and as programmatic web services. These operations are also available for the other 5,500 organisms within BioCyc.

ANALYSIS OF RNA EXPRESSION CHANGE OF KAMPO-TREATED MICE: HOCHUEKKITO AND INFLUENZA

Kotoe Katayama[1], Kaori Munakata[2,3], Katsuaki Dan[3], Rui Yamaguchi[4], Seiya Imoto[4], Kenji Watanabe[3], Satoru Miyano[4]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo; [2]Keio Research Institute at SFC, Keio University; [3]Center for Kampo Medicine, Keio University, School of Medicine; [4]Human Genome Center, Institute of Medical Science, University of Tokyo

Influenza viruses are highly infective disease that spreads in humans around the world in seasonal epidemic. The virus infects about 10% to 20% of the world's total population annually, resulting in around 250,000 deaths. Pandemic influenza virus has been a public health threat. A vaccine builds the basically prevention of influenza infection. However, there is an essential fault. The strategy of the present prevention requires regular monitoring to check coincidence between a vaccine and viral strains. Kampo medicine (Japanese traditional medicine) became integrated into the Japanese health care system—The National Health Insurance Program—46 years ago alongside modern medicine, and the program has covered all citizens since 1961. Because there is only one type of medical license in Japan, physicians prescribe both Kampo medicine and Western biomedicine. At present, 148 Kampo extracts are approved as prescription drugs. A large amount of clinical and basic research on Kampo medicines has been performed, including more than 10 multicenter, placebo-controlled, double blind studies. Hochuekkito (TJ-41) is Kampo medicine; it is composed of several plant materials, including Angelicae radix, Astragali radix, Atractylodis rhizoma, Aurantii nobilis pericarpium, Bupleuri radix, Cimicifugae rhizoma, Ginseng radix, Glycyrrhizae radix, Zingiberis rhizoma, and Zizyphi fructus. TJ-41 is clinically effective against influenza infectious disease, however details of the immune response kinetics against influenza virus infection were not elucidated. We investigated the influence of TJ-41 on the gene expression profile in the lung and spleen by microarray analyses using mice of different conditions. TJ-41 (1 g/kg) was given to C57BL/6 mice orally once a day for 2 weeks. In addition, a water-only control group was used. We divided the case-control into 2 categories. One was then intranasally infected with influenza virus and the other wasn't. After 6 and 24 hours from infection, lung and spleen were extracted and microarray analyses were conducted. After normalizing microarray data by percentile normalization, we applied t-test and Gene Set Enrichment Analysis for extracting significantly differentially expressed genes and pathway to understand TJ-41 effects to the prevention for Influenza virus.

VARIFI - AUTOMATIC VARIANT IDENTIFICATION, FILTERING AND ANNOTATION OF CANCER SEQUENCING DATA

Milica Krunic[1], Niko Popitsch[1], Bettina Pichlhoefer[2], Leonhard Muellauer[2], Arndt von Haeseler[3]

1Center for Integrative Bioinformatics Vienna Max F. Perutz Laboratories University of Vienna Medical University of Vienna; [2]Institute of Pathology Laboratory of Molecular Pathology Medical University Vienna; [3]Center for Integrative Bioinformatics Vienna Max F. Perutz Laboratories University of Vienna Medical University of Vienna

Fast and affordable benchtop sequencers (e.g. PGM from Ion Torrent) are becoming more important in improving a personalized medical treatment. Still, distinguishing genetic variants between healthy and diseased individuals from sequencing errors remains a challenge. Here we present VARIFI, a pipeline for finding reliable variants (SNPs and INDELs). In contrast to current methods, VARIFI assigns a confidence score to every identified variant. This is based on a concordance between three mappers (bwa, bowtie2 and NextGenMap) and two state of the art variant callers (samtools and gatk). In addition, VARIFI includes methods to assess low abundant variants, which is necessary for an early stage cancer diagnostics. We trained the pipeline on more than 150 amplicon sequenced cancer samples. VARIFI applies variant filters for biases associated with Ion Torrent technology (e.g. homopolymer bias). VARIFI automatically extracts variant information from available databases (like dbSNP, COSMIC) and incorporates methods for variant effect prediction (e.g. PolyPhen-2, SIFT). Moreover, VARIFI allows manual inspection of each variant using IGV. VARIFI is currently successfully applied in the Vienna General Hospital as a diagnostic support for cancer patients.

MOLECULAR MODELING OF BIOLOGICAL SYSTEMS FOR NANOMEDICINE APPLICATIONS

Hwankyu Lee

Dankook University

Antimicrobial peptides and nanoparticles such as liposomes, dendrimers, and carbon nanotubes have great potential for antitumor therapeutics applications. To improve their efficiency in biomedical applications, those peptides and nanoparticles are often modified with peptides and polymers. Here, we present multiscale molecular dynamics simulations of the drug transporters modified with polypeptides and polyethylene glycols (PEGs), showing the reduced cytotoxicity as well as the increased water solubility and circulating lifetime. In particular, simulations of dendrimers and carbon nanotubes grafted with peptides and PEG show that the size and grafting density of conjugates significantly modulate the conformation and internal structure (dense core or dense shell) of the particles, implying important possible effects of the conjugate methodology on encapsulation efficiency and cytotoxicity, as observed or proposed in experiment. Also, PEGylated antimicrobial peptides were simulated, showing that PEGylated magainin 2 and tachyplesin I have the different extents of the membrane-permeabilizing activity on lipid bilayer surface. This work aids in the rational design of synthetic peptides, nanoparticles, and drug complexes for nanomedicine applications.

IPRO: AN INTEGRATED SUITE OF COMPUTATIONAL TOOLS FOR PROTEIN AND ANTIBODY DESIGN

Costas Maranas, Tong Li, Matt Grisewood

Penn State

Proteins are an important class of biomolecules with cross-cutting applications across biotechnology and medicine. In many cases, native proteins need to be redesigned by changing the amino acid usage pattern in response to improving various performance metrics. Algorithms can help sharpen the design of a protein library by focusing amino acid usage patterns to those that optimize the value of the computationally accessible proxies. The Iterative Protein Redesign & Optimization Suite of Programs (IPRO) offers an integrated environment for (1) altering protein binding affinity and specificity, (2) grafting a binding pocket into an existing protein scaffold, (3) predicting an antibody's tertiary structure based on its sequence, (4) enhancing enzymatic activity, and (5) assessing the structure and binding energetics for a specific mutant. This poster provides an overview of the methods involved in IPRO, input language terminology, algorithmic details, software implementation specifics and application highlights. IPRO can be downloaded at http://maranas.che.psu.edu. OptZyme is geared towards designing for improved enzymatic activity for an unnatural substrate. The key concept underlying OptZyme is that transition state analogues can be used as a substitute for their typically-unknown transition state counterparts. Using the E. coli β-glucuronidase benchmark system, were able to switch substrate specificity from a native substrate analog to a novel substrate. We showed a strong correlation between interaction energy with the substrate and KM, and between interaction energy with the transition state analogue and kcat/KM. A weighted difference of the transition state analogue and substrate interaction energies correlated well with kcat. OptMAVEn (Optimal Method for Antibody Variable region Engineering) to de novo design the entire antibody variable region targeting a given antigen epitope. OptMAVEn simulates in silico the in vivo steps of antibody generation and evolution, and is capable of capturing the critical structural features responsible for affinity maturation of antibodies along with a procedure to minimize potential immunogenicity. OptMAVEn was applied to design models of neutralizing antibodies targeting influenza hemagglutinin and HIV gp120. The observed rates of mutations and types of amino acid changes during in silico affinity maturation are consistent with what has been observed during in vivo affinity maturation.

SIMILARITY INDEX FOR IMAGE PATTERNS FROM IMAGING MASS SPECTROMETRY

Masaaki Matsuura[1], Masaru Ushijima[2], Shigeki Kajihara[3]

[1]Teikyo University Japan, [2]Japanese Foundation for Cancer Research Japan, [3]Shimadzu Corporation Japan

In this presentation, we discuss a statistical similarity index of image patterns from the imaging mass spectrometry (IMS) data to classify enormous image results. The IMS is one of the measurement technologies of biochemistry used in mass spectrometry with microscopic imaging capabilities to detect and visualize bio-molecules in tissue sections such as pathological samples. In traditional mass spectrometry, we obtain one spectrum for each sample. On the other hand, in IMS, we obtain one spectrum for each spot in which laser beam is shot in one tissue section. By using a recent apparatus, we can obtain the total 62500 (250 x 250) spots for 2.5 cm square region in the tissue. Therefore we have to treat huge amount of spectra data and enormous resulting image patterns of intensity in the region to analyze the IMS data. A high-speed method for evaluation of similarity of image patterns had been required to detect and classify important imaging patterns from many of them. Recently we have developed a similarity index for image patterns. For constructing the similarity index, we considered a two dimensional density estimation for each image pattern, and formulated a Kullback-Leibler (KL) divergence of the two density estimates to evaluate the distance of the two image patterns. Then, to reduce calculation time, we extended this formula to a high speed version of index of the distance, using a sub-regional simple index instead of using a KL-divergence value for each spot. Here, we assume that we treat dataset for the image pattern of peaks after some data processing for spectra data. In these data processing, a peak picking method is applied to each spectrum to detect molecular masses represented by m/z (mass-to-charge ratio), and then, a calibration method for different m/z values for the same peak among different spectra is applied to obtain common m/z value for the same peak. We show performance of some indexes and compare calculation time. Furthermore, we examine performance of similarity for proposed indexes by different sub-region sizes. Our proposed method showed good performance and we will discuss future problems.

COMPUTATIONAL IDENTIFICATION OF HIV ANTIBODY EPITOPES USING THE ENV STRUCTURE AND SEQUENCE CONTEXT.

Ben Murrell[1], Daniel Sheward[2], Kemal Eren[1], Hugh Murrell[3], Davey Smith[1], Sergei Kosakovsky Pond[1], Konrad Scheffler[1]

[1]University of California, San Diego; [2]University of Cape Town;[3]University of KwaZulu Natal

A subset of HIV infected individuals mount an antibody response that is capable of neutralizing a diverse variety of HIV isolates. Understanding such antibody responses against HIV's envelope protein is critical for rational vaccine design. A key part of this involves characterizing the epitope - the specific region of the envelope where the antibody binds. Identifying epitopes targeted by broadly neutralizing antibodies using experimental techniques alone is costly, time consuming, and may not identify all evolutionarily viable escape pathways. Computational methods for epitope identification are thus needed to supplement experimental and structural epitope mapping. The identification of epitopes from matched pairs of sequences and IC50 titers is riddled with statistical difficulties. Indeed, there are many more possible features (amino acid sites or putative glycans) than observations, phylogenetic relatedness introduces nontrivial correlational structure between features, and the effect of individual features frequently depends on the surrounding features. Successful solutions will thus need to maximally exploit the signal present in the data and employ strong regularization strategies to minimize the influence of noise. To this end, we present a novel Bayesian approach that takes advantage of two underexploited sources of information: 1) the context of the surrounding sequence, and 2) the recently published structure of the HIV envelope trimer. To illustrate why sequence context is important, consider a panel of env sequence/titer pairs that happens to contain two sequences with very different titers, but that are identical everywhere except at a single amino acid site. Context sensitive methods will notice that only the variable site can explain this titer difference, rightly extracting maximal evidence from just two sequences. Also, since sites comprising an epitope cluster together on the protein surface, the envelope trimer structure can be exploited for regularization, allowing the suppression of noise from features that happen to correlate with titers by chance alone, but exhibit no spatial colocation. We will describe the method (implemented in R and freely available), present results for a number of broadly neutralizing antibodies, and discuss directions for future research.

CONSTRUCTION OF THE INTEGRATED CO-EXPRESSION NETWORK INCLUDING DNA METHYLATION AND GENE-EXPRESSION

Masahiro Nakatochi,, Teppei Shimamura

Nagoya Univeristy Hospital, Nagoya University

Co-expression network analysis have been often performed to understand the mechanisms of complex diseases such as metabolic syndrome, cancer and mental disease. As a flow of the network analysis, at first, this analysis construct co-expression network based on the gene expression profiles measured by DNA microarray and RNAseq. Next, based on the constructed co-expression network identify network modules (clusters of interconnected nodes) which be significantly related to disease traits and expression levels of genes associated with diseases. However, gene expression is regulated by a variety of DNA modification such DNA methylation and histone modification. Constructions of co-expression networks based on expression levels are not enough. It is hoped that constructions of integrated co-expression networks based on not only gene-expression but also DNA methylation and other factors. In this study, we suggest a novel construction method of the integrated co-expression network based on gene expression levels and DNA methylation levels. This method is based on the weighted co-expression network analysis, and considering both gene expression levels and DNA methylation levels, summarizing these methylation levels in each CpG island. Furthermore, we applied this method to the public available dataset from adipose tissue in the Multiple Tissue Human Expression Resource (MuTHER) project. It was extracted that modules related to the expression levels of adipocytokine genes, which play a prominent role in metabolic syndrome and coronary artery disease. This method is expected to provide the new insights about the architecture of genetics and epigenetics.

COMPUTATIONAL ANALYSIS OF THE RELATIONSHIP BETWEEN HISTONE MODIFICATION AND CORE PROMOTER ELEMENT

Yayoi Natsume-Kitatani[1], Hiroshi Mamitsuka[2]

[1]Japan Science and Technology Agency, [2]Kyoto University

Two each of four different histones (H2A, H2B, H3 and H4) constitute a histone octamer, around which DNA wraps to form histone-DNA complex called nucleosome. It is widely known that amino acid residues in each histone are occasionally modified to exhibit several biological consequences including transcriptional regulation. The mode of transcription can be grouped into two types, dispersed transcription and focused transcription. The focused core promoters have many DNA motifs that are conserved among many species and function to regulate transcription. These DNA motifs are represented by the Initiator (Inr), the TATA box, DPE, and are collectively called "core promoter element (CPE)". We obtained lists of core promoters which have only Inr and no TATA or DPE (Inr group), core promoters which have a TATA but no DPE (TATA group), core promoters which have a DPE but no TATA (DPE group), and core promoters which have both a TATA and DPE (TATA-DPE group). We compared these groups by calculating histone modification ratio at core promoter region and transcribed region for each promoter. We found that the pattern of temporal change in histone modification differs based on whether the group has the TATA box or not. The promoters in TATA-less groups are static and are divided into three types: i) promoters which have active marks (H3K4me3 and H3K27ac) continuously, ii) promoters which have an inactive mark (H3K27me3) continuously and have active marks occasionally, iii) promoters which have histone modification only occasionally. On the other hand, the promoters in TATA-containing groups have less clear tendency and less histone modifications detected. We performed linear regression analysis to explain the RNA expression values with the histone modification ratio. In TATA-less groups, a positive correlation is observed between measured and predicted RNA expression values as we expected. However, in TATA-containing groups, the observations can be divided into two types: i) the promoters that have a positive correlation between measured and predicted RNA expression values, ii) the promoters whose measured values are independent of the regression equations. When the relative frequency of histone modifications is compared, DPE-containing groups have higher frequency of H3K27me3 in both core promoter region and transcript region. DPE-less groups have higher frequency of active marks.

AN ONTOLOGY APPROACH TO COMPARATIVE PHENOMICS IN PLANTS. ARABIDOPSIS, TOMATO, MAIZE, RICE, SOYBEAN AND MEDICAGO

Anika Oellrich[1], Ramona Walls[2], Ethalinda Cannon[3], Steven B. Cannon[4], Laurel Cooper[5], Jack Gardiner[3], Georgios Gkoutos[6], Lisa Harper[7], Mingze He[3], Robert Hoehndorf[8], Pankaj Jaiswal[5], Johnny Lloyd[9], Scott R. Kalberer[4], David Meinke[10], Naama Menda[11], Laura Moore[12], Rex Nelson[4], Carolyn J. Lawrence[3], Eva Huala[13]

[1]Wellcome Trust Sanger Institute; [2]The iPlant Collaborative; [3]Iowa State University; [4]USDA-ARS, Ames, IA; [5]Department of Botany & Plant Pathology, Oregon State University, Corvallis, OR; [6]Computer Science Department, Aberystwyth University; [7]USDA-ARS, Albany, CA; [8]King Abdullah University of Science & Technology; [9]Michigan State University; [10]Oklahoma State University; [11]Boyce Thompson Institute for Plant Research; [12]Oregon State University; [13]Phoenix Bioinformatics

Plant phenotypes are described using a variety of formats, but primarily using free text. While this enables some limited comparison of phenotype data across a single species, or within a knowledge domain such as leaf development or maize breeding, queries that span a broader set of species are not possible in the absence of a common template for describing phenotypes. A pilot project has been undertaken to formalize phenotype descriptors for six plant species, encompassing both crops and model species, and focusing on mutant phenotypes associated with known, sequenced genes. Mutant phenotypes of Arabidopsis, maize, rice, soybean, tomato, and Medicago were manually curated and the free text descriptions were converted into a common Entity-Quality format using taxonomically broad ontologies (Plant Ontology, Gene Ontology, ChEBI, PATO and EO). Cross-species and cross-domain phenotype comparisons and semantic similarity analyses are enabled by utilizing standardized ontology terms and associated relations. The ontology-based phenotypic descriptions are being compared to an existing classification of plant phenotypes. In addition, the semantic similarity dataset is being evaluated for its ability to enhance predictions of gene families, gene functions, and shared metabolic pathways across the six species. The use of ontologies, annotation standards, formats and best practices for plant phenotype data is a novel approach which can be expanded to other plant species, with less well-characterized genomes. These tools will enable us to explore the relationship between gene function, sequence similarity, and phenotypic similarity to make predictions useful for crop improvement.

HiTMAP: A HIGH-THROUGHPUT METHYLATION ANALYSIS PROGRAM FOR BISULFITE SEQUENCING

Benjamin Pullman, Yao Yang, Stuart A. Scott

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

DNA methylation plays an essential role in gene expression regulation, imprinting and X chromosome inactivation, and deregulation of the epigenetic machinery has been directly implicated in both Mendelian disorders and tumorigenesis. Many techniques have been developed to detect CpG methylation, one of the most common involving bisulfite conversion and subsequent DNA sequencing. The recent availability of high-throughput, next-generation sequencing has prompted the development of both whole-genome bisulfite sequencing as well as highly quantitative and multiplexed targeted bisulfite sequencing; however, these techniques require significant computational expertise to manage and analyze data. To address the need for a stand-alone program capable of analyzing data from targeted bisulfite sequencing technologies, we developed HiTMAP: a High Throughput Methylation Analysis Program for bisulfite sequencing data. HiTMAP is a comprehensive web tool that takes raw bisulfite sequence data and demultiplexes against sample barcodes, aligns sequencing reads to in silico converted genomic reference sequences, quantitates CpG methylation levels, and exports resulting methylation data for both individual CpG sites and amplicon regions. The user-facing side of HiTMAP provides an online interface for uploading raw sequence and reference files, setting alignment and methylation quantitation parameters, and for retrieving and saving analysis output data and result figures. The front end was written in JavaScript with AngularJS, and computation is accomplished on a server cluster running Scala code that makes use of the Spark and Spray libraries. HiTMAP is currently being validated with single-molecule real-time (SMRT; Pacific Biosciences) bisulfite sequencing data and other DNA methylation analysis packages. Coupled with an online raw data submission system, HiTMAP eliminates the need for manual data manipulation, local computational resources and expertise, and provides an efficient mechanism to measure CpG methylation from high-throughput next-generation bisulfite sequencing data.

DEVELOPING NEW METHODS FOR IDENTIFYING GENOMIC VARIANTS IN DIFFICULT PHARMACOGENOMIC INTEREST GENES

Xiang Qin, Mark Wang, Harsha Doddapaneni, Donna Muzny, Richard Gibbs, Steve Scherer

Human Genome Sequencing Center at Baylor College of Medicine

Genomic variation is one of the main sources of inter-patient variability in drug response and treatment outcomes. Genotyping of genes of pharmacogenomic interest has been an important tool in the study of the association of genomic variants with drug response for precision medicine. With the advent of next-generation sequencing technologies, ascertaining both common and rare genomic variants by capture sequencing of targeted “pharmaco” genes has become more common in both research and clinical settings. The Human Genome Sequencing Center (HGSC) at Baylor College of Medicine, in collaboration with the other PGRN deep sequencing resource centers, successfully designed and tested a capture panel specific for 84 genes of pharmacogenomic interest. Although genotyping in most of these genes is highly accurate compared to orthogonal platforms, variant calling in some genes or regions is still error prone due to high sequence homology among genes and neighboring pseudogenes, genomic rearrangement and gene copy number variation. In order to improve variant calling accuracy, we are developing new methods to more effectively call SNPs, indels and structural variation from Illumina capture data by incorporating variant identification with SNP phasing of known variants, known structural variant reference mapping and discordant read pair information. By combining haplotype phasing, structural variation and paired read information in the process, we aim to resolve variants from different haplotypes, thus improving overall variant calling. We also are testing PacBio capture data to evaluate the utility of longer PacBio reads in phasing and structural variation identification. Furthermore, we are testing PCR amplicon spike-in for long repetitive regions. These new approaches, both individually and in combination are demonstrating great promise in the resolution of variants within difficult genes or regions of pharmacological significance.

DIFFERENTIAL MODULE ANALYSIS FOR CANCER APPLICATIONS

Chandan K. Reddy

Department of Computer Science, Wayne State University

Identifying the genes that are differentially expressed between different phenotypes (e.g. various subtypes of a given disease, normal vs. cancer, treated vs. not treated, etc.) is a crucial task in understanding the mechanisms of various diseases. The classical approach aiming to identify individual differentially expressed (DE) genes may not be sufficient to fully understand the underlying mechanisms of a given disease because some genes may play an important role even if they do not meet the criteria for differential expression (e.g. very small changes in the expression of transcription factors can trigger huge downstream changes). In addition, the activities of the genes are not independent of each other. Thus, it becomes critical to be able to study groups of genes in the context of differential analysis, rather than analyzing single gene at a time. Existing approaches that aim to extract such co-expressed groups of genes fail to incorporate class-specific information, and hence, are not capable of identifying a distinguishing set of patterns that may be specific to a particular class only. These problems are exacerbated in the gene expression data analysis due to the presence of several other challenges such as noisy data, and presence of negative correlations. In addition, due to the heterogeneity of some diseases such as cancer, the set of genes can be co-expressed only in a subset of the samples. Hence, the lack of analysis methods that can take the class into consideration when looking for patterns of expression is an important obstacle that prevents a better understanding of such data. In this project, we address this challenge by developing robust algorithms using novel differential modeling approaches that can extract, in a class-aware manner, groups of genes that are correlated. We developed techniques, under a general framework of 'differential modeling', that can quantitatively and qualitatively characterize such differential patterns amongst multiple classes of conditions. Our methods can uniquely capture and analyze such class-specific differential subsets of genes through subspace co-expressions and sub-network modules. The proposed work is rigorously evaluated using both simulated and real data and the results show that our approach provides more robust results compared to the existing state-of-the-art methods. Specifically, our work has the following two components (1) Novel computational methods that can efficiently extract co-expression patterns that are strongly coherent with respect to a specific class of conditions but are not coherent in the other class(es). (2) Novel network-based differential analysis approach for systematically analyzing the topological differences between two gene networks built from gene expression data. The goal here is to extract genes and subnetworks of genes that have significantly different activities in terms of their co-expression values with other genes.

BAYESIAN INTEGRATED ANALYSIS OF CHROMATIN INTERACTION MAPS

Teppei Shimamura[1], Yasuharu Kanki[2], Youichiro Wada[2], Satoru Miyano[2]

[1]Nagoya University Graduate School of Medicine, [2]University of Tokyo

Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) enables the high-throughput detection of long-range genomic interactions bound by protein factors in complex genomes. Accurate identification of chromatin interactions is essential for understanding genome-wide gene regulatory mechanisms controlled by chromosomal interactions and searching for enhancers and regulators within active gene transcription units called transcription factories. Traditional approaches for analyzing ChIA-PET data rely on using only reads that map uniquely to a reference genome. This can lead to the omission of up to 50% of alignable reads and result in increasing false negative detection of genomic interactions. On the other hand, the naive approach of simply using multi-mapped reads increases ambiguity and noise, and results in increasing false positive detection of genomic interactions. Here, we present a probabilistic framework known as a hierarchical Bayes model that utilizes reads that map to multiple locations on the reference genome and integrates ChIA-PET data with chromatin accessibility information measured by DNase I sensitivity or FAIRE assays and genomic information such as sequence mappability and evolutionary conservation.

EXSTRACS 2.0: A SCALABLE MICHIGAN-STYLE LEARNING CLASSIFIER SYSTEM FOR DETECTING, MODELING, AND CHARACTERIZING HETEROGENEITY AND EPISTASIS

Ryan Urbanowicz, Jason H. Moore

Geisel School of Medicine at Dartmouth College

Algorithmic scalability is a major concern for any machine learning strategy in this age of `big data'. A large number of potentially predictive attributes is emblematic of problems in bioinformatics, genetic epidemiology, and many other fields. Previously, ExSTraCS was introduced as an extended Michigan-style supervised learning classifier system that combined a set of powerful heuristics to successfully tackle the challenges of classification, prediction, and knowledge discovery. ExSTraCS seeks to make no assumptions about the data, and is therefore model free and particularly well suited to complex problems that are multi-factorial, interacting (non-linear), heterogeneous, noisy, class imbalanced, or multi-class. While Michigan-style learning classifier systems are an advantageous, powerful, and flexible class of algorithms, they are not considered to be particularly scalable. This work introduces an effective strategy to dramatically improve learning classifier system scalability. The new ExSTraCS 2.0 addresses scalability with (1) a rule specificity limit, (2) new approaches to expert knowledge guided covering and mutation mechanisms, and (3) the implementation and utilization of the TuRF algorithm for improving the quality of expert knowledge discovery in larger datasets. Performance over a complex spectrum of simulated genetic datasets demonstrated that these new mechanisms dramatically improve nearly every performance metric on datasets with 20 attributes and made it possible for ExSTraCS to reliably scale up to perform on related 200 and 2000-attribute datasets. ExSTraCS 2.0 was also able to reliably solve the 6, 11, 20, 37, and 70-bit multiplexer problems (with heterogeneous 3, 4, 5, 6, and 7-way underlying interactions) and did so in fewer learning iterations than previously reported, with smaller finite training sets. Additionally, ExSTraCS 2.0 was able to solve the standard 135-bit multiplexer problem for the second time ever reported in the literature while being the first to do so without reusing building blocks from progressively simpler multiplexer problems. Furthermore, algorithm usability was made simpler through the elimination of previously critical run parameters.

THECELLMAP.ORG: STORING AND VISUALIZING GENETIC INTERACTIONS IN S. CEREVISIAE

Matej Usaj[1], Yizhao Tan[1], Michael Costanzo[1], Chad L. Myers[2], Brenda Andrews[1], Charles Boone[1], Anastasia Baryshnikova[3]

[1]Donnelly Centre, University of Toronto; [2]Department of Computer Science and Engineering, University of Minnesota; [3]Lewis-Sigler Institute for Integrative Genomics, Princeton University

Providing access to quantitative genomic data is key to ensure large-scale data validation and promote new discoveries. Here we report the launch of a web application, theCellMap.org, that serves as the central repository for quantitative genetic interaction data for Saccharomyces cerevisiae produced by systematic Synthetic Genetic Array (SGA) experiments in the Boone/Andrews lab. TheCellMap.org provides a set of fundamental tools for analyzing genetic interaction data. In particular, theCellMap.org allows a user to easily visualize, explore and functionally annotate genetic interactions as well as extract and re-organize sub-networks using data-driven or annotation-driven network layouts in an intuitive and interactive manner.

SCALED SPARSE HIGH DIMENSIONAL TESTS FOR DETECTING SEQUENCE VARIANTS

Shaolong Cao, Huaizhen Qin, Hong-Wen Deng, Yu-Ping Wang

Tulane University

The detection of common and rare genetic variants has been a very challenging but significant problem, especially with the emergence of new generation sequencing technique. Existing statistical approaches have the low detection power for high dimensional set, where the number of samples is smaller than that of the markers to be detected. In this work, we propose a scaled sparse regression based approach for identifying variants. Our approach applies sparse regression with scaled Lp norm regularization ($0<p<1$) to generate a de-biased solution. Base on the solution, we construct the marker wise statistics for sparse high dimensional set test, where the dependence between each marker is incorporated to appropriately control set-based Type I error rates while maintain statistical power gain. Under extensive simulations, the new detection approach yields higher statistical power than SKAT and the other two regional adjustment methods. In addition, when applying the method to the analysis of Mexican Americans sequence data from the GAW18, we can identify seven significantly susceptible genomic regions.

HIGH-PERFORMANCE COMPUTING FOR ASSEMBLY AND ANALYSIS OF BIG GENOMICS DATA

Rene L. Warren, Benjamin P. Vandervalk, Anthony Raymond, Shaun D. Jackman, Hamid Mohamadi, Daniel Paulino, Justin Chu, Ewan Gibb, Inanç Birol

Genome Sciences Centre British Columbia Cancer Agency

DNA Sequencing technology is developing at an unprecedented pace, surpassing the rate of advances in computer hardware development. Limited compute resources for storing, processing and analyzing omics data have spurred the improvements of file compression formats, low memory footprint data structures, algorithms that use communication protocols for parallel programming, and astute approaches for handling large data on commodity hardware. Our research team oversees the development of such bioinformatics technologies. Past accomplishments include: enabling the first assembly with millions of very short sequence reads (Warren et al. 2006), assembly of the human genome from short reads with the first parallel assembler (Simpson et al. 2009) and last year, assembly of the then largest genome, that of the 20 Gbp white spruce (Birol et al. 2013). Here we discuss key enabling algorithms, specifically introducing data structures, processes, compression schemes within ABySS (Simpson et al. 2009), BBT (Chu et al. 2014), DIDA (Mohamadi et al. in preparation), Konnector (Vandervalk et al. 2014) and TASR (Warren et al. 2011) that are tailored to the needs of today's big sequence data reality. We describe limitations imposed by the data size, computing hardware or both and the circumventing technology that makes large data processing possible on modest and high-performance hardware. For each, we present how the resulting algorithm implementation led to the decryption, for the first time, of large GB-sized genomes such as that of conifers and the manageable comparative studies of these sequences. We also showcase comprehensive, low compute cost and time-efficient expression profiling of a newly discovered long non-coding RNA in over 7500 samples across 25 diseases from The Cancer Genome Atlas datasets and detail how the process can be applied to measure presence and expression of any target of interest in prohibitively large data. We will discuss applications of these new scalable tools, particularly BBT and Konnector, for genome assembly, finishing, sequence variant profiling and gene fusion discovery.

Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 23:500 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117 Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MM, Keeling CI, Brand D, Vandervalk BP et al. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 29:1492 Chu J, Sadeghi S, Raymond A, Jackman SD, Nip KM, Mar R, Mohamadi H, Butterfield YS, Robertson AG, Birol I. 2014. BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics. PMID: 25143290 Vandervalk B, Jackman SD, Raymond A, Mohamadi H, Yang C, Attali D A, Chu J, Warren RL, Birol I. 2014. Konnector: connecting paired-end reads using a Bloom filter de Bruijn graph. in Bioinformatics and Biomedicine (BIBM 2014), Belfast UK, 2014. Warren RL, Holt RA: Targeted assembly of short sequence reads. PLoS ONE 2011, 6:e19816.

FUNCTIONAL ANNOTATION OF MAIZE GENE MODELS: A MACHINE LEARNING APPROACH

Kokulapalan Wimalanathan[1,2], Carson Andorf[3], Carolyn J. Lawrence[1,2]

[1]Bioinformatics and Computational Biology, Iowa State University; [2]Department of Genetics Development and Cell Biology; Iowa State University; [3]USDA-ARS Corn Insects and Crop Genetics Research Unit; Iowa State University

Maize is a important agricultural crop which feeds the world. The Gene Ontology (GO) is a structured set of hierarchically related terms that describe molecular functions, biological processes, and cellular localization of gene products. Majority of the GO annotations for maize are derived primarily using large-scale pipelines such as Ensemble and UniProt. Historically, GO term assignment to gene models have been based on a simple methods whereby terms are simply inherited based on sequence similarity to a previously annotated genome. However, when only sequence similarity is used, an incorrect assignment in the original species is inherited by other species and errant functional annotations are propagated. Machine Learning (ML) approaches can be used to overcome this limitation by using a multiple types and sources of data and by assessing previously assigned annotations from a group of genes across various species prior to term assignment. We have focused on using ML to build and evaluate a set of classifiers which assign GO terms to maize genes by optimizing various components of these classifiers, and are currently generating a stringent dataset that forms the basis of this annotation project. Here we describe our pipeline to create high-confidence GO associations for maize gene models based on a supervised Machine Learning approach.

TRANSCRIPT EXPRESSION ESTIMATION WITH CONSIDERING DIFFERENCES OF RNA-SEQ READS

Yao-zhong Zhang, Rui Yamaguchi, Seiya Imoto, Satoru Miyano

The Institute of Medical Science The University of Tokyo

RNA-seq is a powerful technology which enables whole-genome scale transcriptome study. To explore raw RNA-seq data, one elementary task is to estimate transcript(isoform)/gene expression levels based on RNA-seq reads. One general approach to estimate isoform abundance is to build generative models which describe reads generating process and make use of reads quantative information to inference isoform abundance. As for the reads, they are different from a qualitative point of view: reads may contain ambiguities and convey different information of alternative splicing. The ambiguities of reads may not only come from multiple read alignment, but also come from unique genome location with multiple isoforms or genes. In the most of previous work, the differences of reads are ignored. Some work proposed to estimate isoform abundance only using unambiguous reads, but this results in splicing-dependent biases. In this work, we attempt to incorporate differences of reads into the isoform expression estimation. We quantified the ambiguity of a paired-end read based on entropy of the exon path the read covers and proposed using weighted EM algorithm to incorporate reads ambiguities into the inference process. We conducted experiments on both synthetic data and real data to show how will considering read ambiguities affect inference results. We also compared our method with several popular used isoform abundance estimation tools.

COMPUTATIONAL IDENTIFICATION OF DISEASE-SPECIFIC NETWORKS FROM PUBMED LITERATURE

Yuji Zhang

Division of Biostatistics and Bioinformatics, University of Maryland Greenebaum Cancer Center, and Department of Epidemiology and Public Health, University of Maryland School of Medicine

A huge amount of association relationships among biological entities (e.g., diseases, drugs, and genes) are scattered in biomedical literature. How to extract and analyze such heterogeneous data still remains a challenging task for most researchers in the biomedical field. Natural language processing (NLP) has the potential in extracting associations among biological entities from literature. However, association information extracted through NLP can be large, noisy, and redundant which poses significant challenges to biomedical researchers to use such information. To address this challenge, we proposed a two-step computational framework to facilitate the use of NLP results. In the first step, we applied the Latent Dirichlet Allocation (LDA) approach to discover disease-specific topics based on associations derived from literature abstracts. In the second step, we investigated top derived topics by constructing and analyzing topic-related association networks. We illustrated the framework through the construction of disease-specific networks from Semantic MEDLINE, an NLP-generated association database. The properties of each network were investigated at two levels: (1) network property such as hub nodes and degree distribution; (2) local network structure called network motifs. Genes associated with each disease topic were also investigated using gene set enrichment analysis. The results demonstrate that (1) LDA-based approach can group related diseases into the same disease topic; (2) the disease-specific association network follows the scale-free network property, in which hub nodes are enriched in related diseases, genes and drugs; (3) significant network motif patterns were detected in these disease-specific association networks and can infer specific biological meanings; (4) in each disease topic, their genes were enriched in disease-associated biological processes and canonical pathways.

FUNCTIONAL BASIS OF MICROORGANISM CLASSIFICATION

Chengsheng Zhu[1], Tom O. Delmont[2], Timothy M. Vogel[3], Yana Bromberg[1,4]

[1]Department of Biochemistry and Microbiology, Rutgers University; [2]Josephine Bay Paul Center, Marine biological Laboratory, Woods Hole;[3]Environmental Microbial Genomics, Laboratoire Ampere, École Centrale de Lyon, Université de Lyon; [4]Institute of Advanced Study, Technische Universitat Munchen

Correctly identifying nearest “neighbors” of a given microorganism is important in industrial and clinical applications, where close relationships imply similar uses and/or treatments. Microbial classification based on similarity of physiological and genetic organism traits (polyphasic similarity) is experimentally difficult and, arguably, subjective. Evolutionary relatedness, inferred from phylogenetic markers, facilitates classification but does not guarantee functional identity of the members of the same taxa or the lack of similarity between the members of different taxa. Using over thirteen hundred sequenced bacterial genomes we built a novel function-based microorganism classification scheme, functional-repertoire similarity-based organism network (FuSiON). Our scheme is phenetic, based on a network of quantitatively defined organism relationships across the known prokaryotic space. It correlates significantly with the current taxonomy, but the observed discrepancies reveal and quantify both (1) the inconsistency of levels of functional diversity among the different taxa and (2) an (unsurprising) bias towards prioritizing, for classification purposes, relatively minor organism traits of particular interest to humans. Here we show that our network-based organism classification is more robust in handling organism diversity than the traditional pairwise comparison-based metrics. FuSiON highlights the environmental impact as a major driving force of microorganism diversification. Our approach provides a complementary view to cladistic assignments and holds important clues for further exploration of the microbial lifestyles. FuSiON is a more practical fit for biomedical, industrial, and ecological applications, as many of these rely on understanding the functional capabilities of the microbes in their environment, and are less concerned with phylogenetic descent.

AUTHOR INDEX