PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017

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 #2-39 are not assigned poster space.

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

TABLE OF CONTENTS

PROCEEDINGS PAPERS WITH ORAL PRESENTATION

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

POSTER PRESENTATIONS

COMPUTATIONAL APPROACHES TO UNDERSTANDING THE EVOLUTION OF MOLECULAR FUNCTION

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

IDENTIFICATION AND ANALYSIS OF BACTERIAL GENOMIC METABOLIC SIGNATURES

Nathan Bowerman[1], Nathan Tintle[2], Matthew DeJongh[3], Aaron A. Best[1]

[1]Department of Biology, Hope College; [2]Department of Mathematics and Statistics, Dordt College, [3]Department of Computer Science, Hope College

With continued rapid growth in the number and quality of fully sequenced and accurately annotated bacterial genomes, we have unprecedented opportunities to understand metabolic diversity. We selected 101 diverse and representative completely sequenced bacteria and implemented a manual curation effort to identify 846 unique metabolic variants present in these bacteria. The presence or absence of these variants act as a metabolic signature for each of the bacteria, which can then be used to understand similarities and differences between and across bacterial groups. We propose a novel and robust method of summarizing metabolic diversity using metabolic signatures and use this method to generate a metabolic tree, clustering metabolically similar organisms. Resulting analysis of the metabolic tree confirms strong associations with well-established biological results along with direct insight into particular metabolic variants which are most predictive of metabolic diversity. The positive results of this manual curation effort and novel method development suggest that future work is needed to further expand the set of bacteria to which this approach is applied and use the resulting tree to test broad questions about metabolic diversity and complexity across the bacterial tree of life.

WHEN SHOULD WE NOT TRANSFER FUNCTIONAL ANNOTATION BETWEEN SEQUENCE PARALOGS?

Mengfei Cao, Lenore J. Cowen

Tufts University

Current automated computational methods to assign functional labels to unstudied genes often involve transferring annotation from orthologous or paralogous genes, however such genes can evolve divergent functions, making such transfer inappropriate. We consider the problem of determining when it is correct to make such an assignment between paralogs. We construct a benchmark dataset of two types of similar paralogous pairs of genes in the well-studied model organism S. cerevisiae: one set of pairs where single deletion mutants have very similar phenotypes (implying similar functions), and another set of pairs where single deletion mutants have very divergent phenotypes (implying different functions). State of the art methods for this problem will determine the evolutionary history of the paralogs with references to multiple related species. Here, we ask a first and simpler question: we explore to what extent any computational method with access only to data from a single species can solve this problem. We consider divergence data (at both the amino acid and nucleotide levels), and network data (based on the yeast protein-protein interaction network, as captured in BioGRID), and ask if we can extract features from these data that can distinguish between these sets of paralogous gene pairs. We find that the best features come from measures of sequence divergence, however, simple network measures based on degree or centrality or shortest path or diffusion state distance (DSD), or shared neighborhood in the yeast protein-protein interaction (PPI) network also contain some signal. One should, in general, not transfer function if sequence divergence is too high. Further improvements in classification will need to come from more computationally expensive but much more powerful evolutionary methods that incorporate ancestral states and measure evolutionary divergence over multiple species based on evolutionary trees.

PROSNET: INTEGRATING HOMOLOGY WITH MOLECULAR NETWORKS FOR PROTEIN FUNCTION PREDICTION

Sheng Wang, Meng Qu, Jian Peng

University of Illinois Urbana-Champaign

Automated annotation of protein function has become a critical task in the post-genomic era. Network-based approaches and homology-based approaches have been widely used and recently tested in large-scale community-wide assessment experiments. It is natural to integrate network data with homology information to further improve the predictive performance. However, integrating these two heterogeneous, high-dimensional and noisy datasets is non-trivial. In this work, we introduce a novel protein function prediction algorithm ProSNet. An integrated heterogeneous network is first built to include molecular networks of multiple species and link together homologous proteins across multiple species. Based on this integrated network, a dimensionality reduction algorithm is introduced to obtain compact low-dimensional vectors to encode proteins in the network. Finally, we develop machine learning classification algorithms that take the vectors as input and make predictions by transferring annotations both within each species and across different species. Extensive experiments on five major species demonstrate that our integration of homology with molecular networks substantially improves the predictive performance over existing approaches.

ON THE POWER AND LIMITS OF SEQUENCE SIMILARITY BASED CLUSTERING OF PROTEINS INTO FAMILIES

Christian Wiwie, Richard Röttger

University of Southern Denmark

Over the last decades, we have observed an ongoing tremendous growth of available sequencing data fueled by the advancements in wet-lab technology. The sequencing information is only the beginning of the actual understanding of how organisms survive and prosper. It is, for instance, equally important to also unravel the proteomic repertoire of an organism. A classical computational approach for detecting protein families is a sequence-based similarity calculation coupled with a subsequent cluster analysis. In this work we have intensively analyzed various clustering tools on a large scale. We used the data to investigate the behavior of the tools' parameters underlining the diversity of the protein families. Furthermore, we trained regression models for predicting the expected performance of a clustering tool for an unknown data set and aimed to also suggest optimal parameters in an automated fashion. Our analysis demonstrates the benefits and limitations of the clustering of proteins with low sequence similarity indicating that each protein family requires its own distinct set of tools and parameters. All results, a tool prediction service, and additional supporting material is also available online under http://proteinclustering.compbio.sdu.dk/

IMAGING GENOMICS

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

INTEGRATIVE ANALYSIS FOR LUNG ADENOCARCINOMA PREDICTS MORPHOLOGICAL FEATURES ASSOCIATED WITH GENETIC VARIATIONS

Chao Wang[1], Hai Su[2], Lin Yang[2], Kun Huang[1]

[1]The Ohio State University, [2]University of Florida

Lung cancer is one of the most deadly cancers and lung adenocarcinoma (LUAD) is the most common histological type of lung cancer. However, LUAD is highly heterogeneous due to genetic difference as well as phenotypic differences such as cellular and tissue morphology. In this paper, we systematically examine the relationships between histological features and gene transcription. Specifically, we calculated 283 morphological features from histology images for 201 LUAD patients from TCGA project and identified the morphological feature with strong correlation with patient outcome. We then modeled the morphology feature using multiple co-expressed gene clusters using Lasso-regression. Many of the gene clusters are highly associated with genetic variations, specifically DNA copy number variations, implying that genetic variations play important roles in the development cancer morphology. As far as we know, our finding is the first to directly link the genetic variations and functional genomics to LUAD histology. These observations will lead to new insight on lung cancer development and potential new integrative biomarkers for prediction patient prognosis and response to treatments.

IDENTIFICATION OF DISCRIMINATIVE IMAGING PROTEOMICS ASSOCIATIONS IN ALZHEIMER'S DISEASE VIA A NOVEL SPARSE CORRELATION MODEL

Jingwen Yan, Shannon L. Risacher, Kwangsik Nho, Andrew J. Saykin, Li Shen

Indiana University

Brain imaging and protein expression, from both cerebrospinal fluid and blood plasma, have been found to provide complementary information in predicting the clinical outcomes of Alzheimer's disease (AD). But the underlying associations that contribute to such a complementary relationship have not been previously studied yet. In this work, we will perform an imaging proteomics association analysis to explore how they are related with each other. While traditional association models, such as Sparse Canonical Correlation Analysis (SCCA), can not guarantee the selection of only disease-relevant biomarkers and associations, we propose a novel discriminative SCCA (denoted as DSCCA) model with new penalty terms to account for the disease status information. Given brain imaging, proteomic and diagnostic data, the proposed model can perform a joint association and multi-class discrimination analysis, such that we can not only identify disease-relevant multimodal biomarkers, but also reveal strong associations between them. Based on a real imaging proteomic data set, the empirical results show that DSCCA and traditional SCCA have comparable association performances. But in a further classification analysis, canonical variables of imaging and proteomic data obtained in DSCCA demonstrate much more discrimination power toward multiple pairs of diagnosis groups than those obtained in SCCA.

ENFORCING CO-EXPRESSION IN MULTIMODAL REGRESSION FRAMEWORK

Pascal Zille[1], Vince D. Calhoun[2], Yu-Ping Wang[1]

[1]Tulane University, [2]University of New Mexico

We consider the problem of multimodal data integration for the study of complex neurological diseases (e.g. schizophrenia). Among the challenges arising in such situation, estimating the link between genetic and neurological variability within a population sample has been a promising direction. A wide variety of statistical models arose from such applications. For example, Lasso regression and its multitask extension are often used to fit a multivariate linear relationship between given phenotype(s) and associated observations. Other approaches, such as canonical correlation analysis (CCA), are widely used to extract relationships between sets of variables from different modalities. In this paper, we propose an exploratory multivariate method combining these two methods. More Specifically, we rely on a 'CCA-type' formulation in order to regularize the classical multimodal Lasso regression problem. The underlying motivation is to extract discriminative variables that display are also co-expressed across modalities. We first evaluate the method on a simulated dataset, and further validate it using Single Nucleotide Polymorphisms (SNP) and functional Magnetic Resonance Imaging (fMRI) data for the study of schizophrenia.

METHODS TO ENSURE THE REPRODUCIBILITY OF BIOMEDICAL RESEARCH

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

EXPLORING THE REPRODUCIBILITY OF PROBABILISTIC CAUSAL MOLECULAR NETWORK MODELS

Ariella Cohain, Aparna A. Divaraniya, Kuixi Zhu, Joseph R. Scarpa, Andrew Kasarskis, Jun Zhu, Rui Chang, Joel T. Dudley, Eric E. Schadt

Icahn Institute and Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai

Network reconstruction algorithms are increasingly being employed in biomedical and life sciences research to integrate large-scale, high-dimensional data informing on living systems. One particular class of probabilistic causal networks being applied to model the complexity and causal structure of biological data is Bayesian networks (BNs). BNs provide an elegant mathematical framework for not only inferring causal relationships among many different molecular and higher order phenotypes, but also for incorporating highly diverse priors that provide an efficient path for incorporating existing knowledge. While significant methodological developments have broadly enabled the application of BNs to generate and validate meaningful biological hypotheses, the reproducibility of BNs in this context has not been systematically explored. In this study, we aim to determine the criteria for generating reproducible BNs in the context of transcription-based regulatory networks. We utilize two unique tissues from independent datasets, whole blood from the GTEx Consortium and liver from the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Team (STARNET) study. We evaluated the reproducibility of the BNs by creating networks on data subsampled at different levels from each cohort and comparing these networks to the BNs constructed using the complete data. To help validate our results, we used simulated networks at varying sample sizes. Our study indicates that reproducibility of BNs in biological research is an issue worthy of further consideration, especially in light of the many publications that now employ findings from such constructs without appropriate attention paid to reproducibility. We find that while edge-to-edge reproducibility is strongly dependent on sample size, identification of more highly connected key driver nodes in BNs can be carried out with high confidence across a range of sample sizes.

REPRODUCIBLE DRUG REPURPOSING: WHEN SIMILARITY DOES NOT SUFFICE

Emre Guney

Joint IRB-BSC-CRG Program in Computational Biology - Institute for Research in Biomedicine (IRB) Barcelona

Repurposing existing drugs for new uses has attracted considerable attention over the past years. To identify potential candidates that could be repositioned for a new indication, many studies make use of chemical, target, and side effect similarity between drugs to train classifiers. Despite promising prediction accuracies of these supervised computational models, their use in practice, such as for rare diseases, is hindered by the assumption that there are already known and similar drugs for a given condition of interest. In this study, using publicly available data sets, we question the prediction accuracies of supervised approaches based on drug similarity when the drugs in the training and the test set are completely disjoint. We first build a Python platform to generate reproducible similarity- based drug repurposing models. Next, we show that, while a simple chemical, target, and side effect similarity based machine learning method can achieve good performance on the benchmark data set, the prediction performance drops sharply when the drugs in the folds of the cross validation are not overlapping and the similarity information within the training and test sets are used independently. These intriguing results suggest revisiting the assumptions underlying the validation scenarios of similarity-based methods and underline the need for unsupervised approaches to identify novel drug uses inside the unexplored pharmacological space. We make the digital notebook containing the Python code to replicate our analysis that involves the drug repurposing platform based on machine learning models and the proposed disjoint cross fold generation method freely available at github.com/emreg00/repurpose.

EMPOWERING MULTI-COHORT GENE EXPRESSION ANALYSIS TO INCREASE REPRODUCIBILITY

Winston A. Haynes, Francesco Vallania, Charles Liu, Erika Bongen, Aurelie Tomczak, Marta Andres-Terrè, Shane Lofgren, Andrew Tam, Cole A. Deisseroth, Matthew D. Li, Timothy E. Sweeney, Purvesh Khatri

Stanford University

A major contributor to the scientific reproducibility crisis has been that the results from homogeneous, single-center studies do not generalize to heterogeneous, real world populations. Multi-cohort gene expression analysis has helped to increase reproducibility by aggregating data from diverse populations into a single analysis. To make the multi-cohort analysis process more feasible, we have assembled an analysis pipeline which implements rigorously studied meta-analysis best practices. We have compiled and made publicly available the results of our own multi-cohort gene expression analysis of 103 diseases, spanning 615 studies and 36,915 samples, through a novel and interactive web application. As a result, we have made both the process of and the results from multi-cohort gene expression analysis more approachable for non-technical users.

RABIX: AN OPEN-SOURCE WORKFLOW EXECUTOR SUPPORTING RECOMPUTABILITY AND INTEROPERABILITY OF WORKFLOW DESCRIPTIONS

Gaurav Kaushik, Sinisa Ivkovic, Janko Simonovic, Nebojsa Tijanic, Brandi Davis-Dusenbery, Deniz Kural

Seven Bridges Genomics

As biomedical data has become increasingly easy to generate in large quantities, the methods used to analyze it have proliferated rapidly. Reproducible and reusable methods are required to learn from large volumes of data reliably. To address this issue, numerous groups have developed workflow specifications or execution engines, which provide a framework with which to perform a sequence of analyses. One such specification is the Common Workflow Language, an emerging standard which provides a robust and flexible framework for describing data analysis tools and workflows. In addition, reproducibility can be furthered by executors or workflow engines which interpret the specification and enable additional features, such as error logging, file organization, optimizations to computation and job scheduling, and allow for easy computing on large volumes of data. To this end, we have developed the Rabix Executora, an open-source workflow engine for the purposes of improving reproducibility through reusability and interoperability of workflow descriptions.

DATA SHARING AND CLINICAL GENETIC TESTING: SUCCESSES AND CHALLENGES

Shan Yang[1], Melissa Cline[2], Can Zhang[2], Benedict Paten[2], Stephen E. Lincoln[1]

[1]Invitae, [2]University of California Santa Cruz

Open sharing of clinical genetic data promises to both monitor and eventually improve the reproducibility of variant interpretation among clinical testing laboratories. A significant public data resource has been developed by the NIH ClinVar initiative, which includes submissions from hundreds of laboratories and clinics worldwide. We analyzed a subset of ClinVar data focused on specific clinical areas and we find high reproducibility (>90% concordance) among labs, although challenges for the community are clearly identified in this dataset. We further review results for the commonly tested BRCA1 and BRCA2 genes, which show even higher concordance, although the significant fragmentation of data into different silos presents an ongoing challenge now being addressed by the BRCA Exchange. We encourage all laboratories and clinics to contribute to these important resources.

PATTERNS IN BIOMEDICAL DATA – HOW DO WE FIND THEM?

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

LEARNING ATTRIBUTES OF DISEASE PROGRESSION FROM TRAJECTORIES OF SPARSE LAB VALUES

Vibhu Agarwal[1], Nigam H. Shah[2]

[1]Biomedical Informatics Training Program, Stanford University, [2]The Center for Biomedical Informatics Research, Stanford University

There is heterogeneity in the manifestation of diseases, therefore it is essential to understand the patterns of progression of a disease in a given population for disease management as well as for clinical research. Disease status is often summarized by repeated recordings of one or more physiological measures. As a result, historical values of these physiological measures for a population sample can be used to characterize disease progression patterns. We use a method for clustering sparse functional data for identifying sub-groups within a cohort of patients with chronic kidney disease (CKD), based on the trajectories of their Creatinine measurements. We demonstrate through a proof-of-principle study how the two sub-groups that display distinct patterns of disease progression may be compared on clinical attributes that correspond to the maximum difference in progression patterns. The key attributes that distinguish the two sub-groups appear to have support in published literature clinical practice related to CKD.

COMPUTER AIDED IMAGE SEGMENTATION AND CLASSIFICATION FOR VIABLE AND NON-VIABLE TUMOR IDENTIFICATION IN OSTEOSARCOMA

Harish Babu Arunachalam[1], Rashika Mishra[1], Bogdan Armaselu[1], Ovidiu Daescu[1], Maria Martinez[1], Patrick Leavey[1], Dinesh Rakheja[2], Kevin Cederberg[2], Anita Sengupta[2], Molly Ni'Suilleabhain[2]

[1]University of Texas at Dallas, [2]University of Texas Southwestern Medical Center

Osteosarcoma is one of the most common types of bone cancer in children. To gauge the extent of cancer treatment response in the patient after surgical resection, the H&E stained image slides are manually evaluated by pathologists to estimate the percentage of necrosis, a time consuming process prone to observer bias and inaccuracy. Digital image analysis is a potential method to automate this process, thus saving time and providing a more accurate evaluation. The slides are scanned in Aperio Scanscope, converted to digital Whole Slide Images (WSIs) and stored in SVS format. These are high resolution images, of the order of 10^9 pixels, allowing up to 40X magnification factor. This paper proposes an image segmentation and analysis technique for segmenting tumor and non-tumor regions in histopathological WSIs of osteosarcoma datasets. Our approach is a combination of pixel-based and object-based methods which utilize tumor properties such as nuclei cluster, density, and circularity to classify tumor regions as viable and non-viable. A K-Means clustering technique is used for tumor isolation using color normalization, followed by multi-threshold Otsu segmentation technique to further classify tumor region as viable and non-viable. Then a Flood-fill algorithm is applied to cluster similar pixels into cellular objects and compute cluster data for further analysis of regions under study. To the best of our knowledge this is the first comprehensive solution that is able to produce such a classification for Osteosarcoma cancer. The results are very conclusive in identifying viable and non-viable tumor regions. In our experiments, the accuracy of the discussed approach is 100% in viable tumor and coagulative necrosis identification while it is around 90% for fibrosis and acellular/hypocellular tumor osteoid, for all the sampled datasets used. We expect the developed software to lead to a significant increase in accuracy and decrease in inter-observer variability in assessment of necrosis by the pathologists and a reduction in the time spent by the pathologists in such assessments.

MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS

Brett K. Beaulieu-Jones[1], Jason H. Moore[2], The Pooled Resource Open-Access ALS Clinical Trials Consortium

[1]Genomics and Computational Biology Graduate Group, Computational Genetics Lab, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania; [2]Computational Genetics Lab, Institute for Biomedical Informatics, University of Pennsylvania

Electronic health records (EHRs) have become a vital source of patient outcome data but the widespread prevalence of missing data presents a major challenge. Different causes of missing data in the EHR data may introduce unintentional bias. Here, we compare the effectiveness of popular multiple imputation strategies with a deeply learned autoencoder using the Pooled Resource Open-Access ALS Clinical Trials Database (PRO-ACT). To evaluate performance, we examined imputation accuracy for known values simulated to be either missing completely at random or missing not at random. We also compared ALS disease progression prediction across different imputation models. Autoencoders showed strong performance for imputation accuracy and contributed to the strongest disease progression predictor. Finally, we show that despite clinical heterogeneity, ALS disease progression appears homogenous with time from onset being the most important predictor.

DEVELOPMENT AND PERFORMANCE OF TEXT-MINING ALGORITHMS TO EXTRACT SOCIOECONOMIC STATUS FROM DE-IDENTIFIED ELECTRONIC HEALTH RECORDS

Brittany M. Hollister[1], Nicole A. Restrepo[2], Eric Farber-Eger[3], Dana C. Crawford[2], Melinda C. Aldrich[4], Amy Non[5]

[1]Vanderbilt Genetic Institute, Vanderbilt University; [2]Institute for Computational Biology and Department of Epidemiology and Biostatistics, Case Western Reserve University; [3]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University; [4]Department of Thoracic Surgery and Division of Epidemiology, Vanderbilt University Medical Center; [5]Department of Anthropology, University of California San Diego

Socioeconomic status (SES) is a fundamental contributor to health, and a key factor underlying racial disparities in disease. However, SES data are rarely included in genetic studies due in part to the difficultly of collecting these data when studies were not originally designed for that purpose. The emergence of large clinic-based biobanks linked to electronic health records (EHRs) provides research access to large patient populations with longitudinal phenotype data captured in structured fields as billing codes, procedure codes, and prescriptions. SES data however, are often not explicitly recorded in structured fields, but rather recorded in the free text of clinical notes and communications. The content and completeness of these data vary widely by practitioner. To enable gene-environment studies that consider SES as an exposure, we sought to extract SES variables from racial/ethnic minority adult patients (n=9,977) in BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified EHRs. We developed several measures of SES using information available within the de-identified EHR, including broad categories of occupation, education, insurance status, and homelessness. Two hundred patients were randomly selected for manual review to develop a set of seven algorithms for extracting SES information from de-identified EHRs. The algorithms consist of 15 categories of information, with 830 unique search terms. SES data extracted from manual review of 50 randomly selected records were compared to data produced by the algorithm, resulting in positive predictive values of 80.0% (education), 85.4% (occupation), 87.5% (unemployment), 63.6% (retirement), 23.1% (uninsured), 81.8% (Medicaid), and 33.3% (homelessness), suggesting some categories of SES data are easier to extract in this EHR than others. The SES data extraction approach developed here will enable future EHR-based genetic studies to integrate SES information into statistical analyses. Ultimately, incorporation of measures of SES into genetic studies will help elucidate the impact of the social environment on disease risk and outcomes.

DeMo DASHBOARD: VISUALIZING AND UNDERSTANDING GENOMIC SEQUENCES USING DEEP NEURAL NETWORKS

Jack Lanchantin, Ritambhara Singh, Beilun Wang, Yanjun Qi

University of Virginia

Deep neural network (DNN) models have recently obtained state-of-the-art prediction accuracy for the transcription factor binding (TFBS) site classification task. However, it remains unclear how these approaches identify meaningful DNA sequence signals and give insights as to why TFs bind to certain locations. In this paper, we propose a toolkit called the Deep Motif Dashboard (DeMo Dashboard) which provides a suite of visualization strategies to extract motifs, or sequence patterns from deep neural network models for TFBS classification. We demonstrate how to visualize and understand three important DNN models: convolutional, recurrent, and convolutional-recurrent networks. Our first visualization method is finding a test sequence's saliency map which uses first-order derivatives to describe the importance of each nucleotide in making the final prediction. Second, considering recurrent models make predictions in a temporal manner (from one end of a TFBS sequence to the other), we introduce temporal output scores, indicating the prediction score of a model over time for a sequential input. Lastly, a class-specific visualization strategy finds the optimal input sequence for a given TFBS positive class via stochastic gradient optimization. Our experimental results indicate that a convolutional-recurrent architecture performs the best among the three architectures. The visualization techniques indicate that CNN-RNN makes predictions by modeling both motifs as well as dependencies among them.

PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING: A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT

Khader Shameer[1,2], Kipp W. Johnson[1,2], Alexandre Yahi[7], Riccardo Miotto[1,2], Li Li[1,2], Doran Ricks[3], Jebakumar Jebakaran[4], Patricia Kovatch[1,4], Partho P. Sengupta[5], Annetine Gelijns[8], Alan Moskovitz[8], Bruce Darrow[5], David L. Reich[6], Andrew Kasarskis[1], Nicholas P. Tatonetti[7], Sean Pinney[5], Joel T. Dudley[1,2,8]*

[1]Department of Genetics and Genomics, Icahn Institute of Genomics and Multiscale Biology; [2]Institute of Next Generation Healthcare, Mount Sinai Health System, NY; [3]Decision Support, Mount Sinai Health System, NY; [4]Mount Sinai Data Warehouse, Icahn Institute of Genomics and Multiscale Biology, NY; [5]Zena and Michael A. Wiener Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, NY; [6]Department of Anesthesiology, Icahn School of Medicine at Mount Sinai, NY; [7]Departments of Biomedical Informatics, Systems Biology and Medicine, Columbia University Medical Center, NY; [8]Population Health Science and Policy, Mount Sinai Health System, NY

* Corresponding Author, Email: joel.dudley@mssm.edu

Reduction of preventable hospital readmissions that result from chronic or acute conditions like stroke, heart failure, myocardial infarction and pneumonia remains a significant challenge for improving the outcomes and decreasing the cost of healthcare delivery in the United States. Patient readmission rates are relatively high for conditions like heart failure (HF) despite the implementation of high-quality healthcare delivery operation guidelines created by regulatory authorities. Multiple predictive models are currently available to evaluate potential 30-day readmission rates of patients. Most of these models are hypothesis driven and repetitively assess the predictive abilities of the same set of biomarkers as predictive features. In this manuscript, we discuss our attempt to develop a data-driven, electronic-medical record-wide (EMR-wide) feature selection approach and subsequent machine learning to predict readmission probabilities. We have assessed a large repertoire of variables from electronic medical records of heart failure patients in a single center. The cohort included 1,068 patients with 178 patients were readmitted within a 30-day interval (16.66% readmission rate). A total of 4,205 variables were extracted from EMR including diagnosis codes (n=1,763), medications (n=1,028), laboratory measurements (n=846), surgical procedures (n=564) and vital signs (n=4). We designed a multistep modeling strategy using the Naïve Bayes algorithm. In the first step, we created individual models to classify the cases (readmitted) and controls (non-readmitted). In the second step, features contributing to predictive risk from independent models were combined into a composite model using a correlation-based feature selection (CFS) method. All models were trained and tested using a 5-fold cross-validation method, with 70% of the cohort used for training and the remaining 30% for testing. Compared to existing predictive models for HF readmission rates (AUCs in the range of 0.6-0.7), results from our EMR-wide predictive model (AUC=0.78; Accuracy=83.19%) and phenome-wide feature selection strategies are encouraging and reveal the utility of such data-driven machine learning. Fine tuning of the model, replication using multi-center cohorts and prospective clinical trial to evaluate the clinical utility would help the adoption of the model as a clinical decision system for evaluating readmission status.

METHODS FOR CLUSTERING TIME SERIES DATA ACQUIRED FROM MOBILE HEALTH APPS

Nicole Tignor[1], Pei Wang[1], Nicholas Genes[1], Linda Rogers[1], Steven G. Hershman[2], Erick R. Scott[1], Micol Zweig[1], Yu-Feng Yvonne Chan[1], Eric E. Schadt[1]

[1]Icahn School of Medicine at Mount Sinai, [2]Life Map Solutions

In our recent Asthma Mobile Health Study (AMHS), thousands of asthma patients across the country contributed medical data through the iPhone Asthma Health App on a daily basis for an extended period of time. The collected data included daily self-reported asthma symptoms, symptom triggers, and real time geographic location information. The AMHS is just one of many studies occurring in the context of now many thousands of mobile health apps aimed at improving wellness and better managing chronic disease conditions, leveraging the passive and active collection of data from mobile, handheld smart devices. The ability to identify patient groups or patterns of symptoms that might predict adverse outcomes such as asthma exacerbations or hospitalizations from these types of large, prospectively collected data sets, would be of significant general interest. However, conventional clustering methods cannot be applied to these types of longitudinally collected data, especially survey data actively collected from app users, given heterogeneous patterns of missing values due to: 1) varying survey response rates among different users, 2) varying survey response rates over time of each user, and 3) non-overlapping periods of enrollment among different users. To handle such complicated missing data structure, we proposed a probability imputation model to infer missing data. We also employed a consensus clustering strategy in tandem with the multiple imputation procedure. Through simulation studies under a range of scenarios reflecting real data conditions, we identified favorable performance of the proposed method over other strategies that impute the missing value through low-rank matrix completion. When applying the proposed new method to study asthma triggers and symptoms collected as part of the AMHS, we identified several patient groups with distinct phenotype patterns. Further validation of the methods described in this paper might be used to identify clinically important patterns in large data sets with complicated missing data structure, improving the ability to use such data sets to identify at-risk populations for potential intervention.

A NEW RELEVANCE ESTIMATOR FOR THE COMPILATION AND VISUALIZATION OF DISEASE PATTERNS AND POTENTIAL DRUG TARGETS

Modest von Korff, Tobias Fink, Thomas Sander

Research Information Management, Actelion Pharmaceuticals Ltd.

A new computational method is presented to extract disease patterns from heterogeneous and text-based data. For this study, 22 million PubMed records were mined for co-occurrences of gene name synonyms and disease MeSH terms. The resulting publication counts were transferred into a matrix Mdata. In this matrix, a disease was represented by a row and a gene by a column. Each field in the matrix represented the publication count for a co-occurring disease–gene pair. A second matrix with identical dimensions Mrelevance was derived from Mdata. To create Mrelevance the values from Mdata were normalized. The normalized values were multiplied by the column-wise calculated Gini coefficient. This multiplication resulted in a relevance estimator for every gene in relation to a disease. From Mrelevance the similarities between all row vectors were calculated. The resulting similarity matrix Srelevance related 5,000 diseases by the relevance estimators calculated for 15,000 genes. Three diseases were analyzed in detail for the validation of the disease patterns and the relevant genes. Cytoscape was used to visualize and to analyze Mrelevance and Srelevance together with the genes and diseases. Summarizing the results, it can be stated that the relevance estimator introduced here was able to detect valid disease patterns and to identify genes that encoded key proteins and potential targets for drug discovery projects.

DISCOVERY OF FUNCTIONAL AND DISEASE PATHWAYS BY COMMUNITY DETECTION IN PROTEIN-PROTEIN INTERACTION NETWORKS

Stephen J. Wilson, Angela D. Wilkins, Chih-Hsu Lin, Rhonald C. Lua, Olivier Lichtarge

Baylor College of Medicine

Advances in cellular, molecular, and disease biology depend on the comprehensive characterization of gene interactions and pathways. Traditionally, these pathways are curated manually, limiting their efficient annotation and, potentially, reinforcing field-specific bias. Here, in order to test objective and automated identification of functionally cooperative genes, we compared a novel algorithm with three established methods to search for communities within gene interaction networks. Communities identified by the novel approach and by one of the established method overlapped significantly ($q < 0.1$) with control pathways. With respect to disease, these communities were biased to genes with pathogenic variants in ClinVar ($p << 0.01$), and often genes from the same community were co-expressed, including in breast cancers. The interesting subset of novel communities, defined by poor overlap to control pathways also contained co-expressed genes, consistent with a possible functional role. This work shows that community detection based on topological features of networks suggests new, biologically meaningful groupings of genes that, in turn, point to health and disease relevant hypotheses.

PRECISION MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS IMPROVED HEALTH AND THERAPIES

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

OPENING THE DOOR TO THE LARGE SCALE USE OF CLINICAL LAB MEASURES FOR ASSOCIATION TESTING: EXPLORING DIFFERENT METHODS FOR DEFINING PHENOTYPES

Christopher R. Bauer, Daniel Lavage, John Snyder, Joseph Leader, J. Matthew Mahoney, Sarah A. Pendergrass

Geisinger Health System, University of Vermont

The past decade has seen exponential growth in the numbers of sequenced and genotyped individuals and a corresponding increase in our ability of collect and catalogue phenotypic data for use in the clinic. We now face the challenge of integrating these diverse data in new ways new that can provide useful diagnostics and precise medical interventions for individual patients. One of the first steps in this process is to accurately map the phenotypic consequences of the genetic variation in human populations. The most common approach for this is the genome wide association study (GWAS). While this technique is relatively simple to implement for a given phenotype, the choice of how to define a phenotype is critical. It is becoming increasingly common for each individual in a GWAS cohort to have a large profile of quantitative measures. The standard approach is to test for associations with one measure at a time; however, there are many justifiable ways to define a set of phenotypes, and the genetic associations that are revealed will vary based on these definitions. Some phenotypes may only show a significant genetic association signal when considered together, such as through principle components analysis (PCA). Combining correlated measures may increase the power to detect association by reducing the noise present in individual variables and reduce the multiple hypothesis testing burden. Here we show that PCA and k-means clustering are two complimentary methods for identifying novel genotype-phenotype relationships within a set of quantitative human traits derived from the Geisinger Health System electronic health record (EHR). Using a diverse set of approaches for defining phenotype may yield more insights into the genetic architecture of complex traits and the findings presented here highlight a clear need for further investigation into other methods for defining the most relevant phenotypes in a set of variables. As the data of EHR continue to grow, addressing these issues will become increasingly important in our efforts to use genomic data effectively in medicine.

TEMPORAL ORDER OF DISEASE PAIRS AFFECTS SUBSEQUENT DISEASE TRAJECTORIES: THE CASE OF DIABETES AND SLEEP APNEA

Mette Beck[1], David Westergaard[1], Leif Groop[2], Soren Brunak[1]

[1]Novo Nordisk Foundation Center for Protein Research; [2]Lund University Diabetes Centre, Department of Clinical Sciences

Most studies of disease etiologies focus on one disease only and not the full spectrum of multimorbidities that many patients have. Some disease pairs have shared causal origins, others represent common follow-on diseases, while yet other co-occurring diseases may manifest themselves in random order of appearance. We discuss these different types of disease co-occurrences, and use the two diseases “sleep apnea” and “diabetes” to showcase the approach which otherwise can be applied to any disease pair. We benefit from seven million electronic medical records covering the entire population of Denmark for more than 20 years. Sleep apnea is the most common sleep-related breathing disorder and it has previously been shown to be bidirectionally linked to diabetes, meaning that each disease increases the risk of acquiring the other. We confirm that there is no significant temporal relationship, as approximately half of patients with both diseases are diagnosed with diabetes first. However, we also show that patients diagnosed with diabetes before sleep apnea have a higher disease burden compared to patients diagnosed with sleep apnea before diabetes. The study clearly demonstrates that it is not only the diagnoses in the patient’s disease history that are important, but also the specific order in which these diagnosis are given that matters in terms of outcome. We suggest that this should be considered for patient stratification.

HUMAN KINASES DISPLAY MUTATIONAL HOTSPOTS AT COGNATE POSITIONS WITHIN CANCER

Jonathan Gallion, Angela D. Wilkins, Olivier Lichtarge

Baylor College of Medicine

The discovery of driver genes is a major pursuit of cancer genomics, usually based on observing the same mutation in different patients. But the heterogeneity of cancer pathways plus the high background mutational frequency of tumor cells often cloud the distinction between less frequent drivers and innocent passenger mutations. Here, to overcome these disadvantages, we grouped together mutations from close kinase paralogs under the hypothesis that cognate mutations may functionally favor cancer cells in similar ways. Indeed, we find that kinase paralogs often bear mutations to the same substituted amino acid at the same aligned positions and with a large predicted Evolutionary Action. Functionally, these high Evolutionary Action, non-random mutations affect known kinase motifs, but strikingly, they do so differently among different kinase types and cancers, consistent with differences in selective pressures. Taken together, these results suggest that cancer pathways may flexibly distribute a dependence on a given functional mutation among multiple close kinase paralogs. The recognition of this “mutational delocalization” of cancer drivers among groups of paralogs is a new phenomena that may help better identify relevant mechanisms and therefore eventually guide personalized therapy.

MUSE: A MULTI-LOCUS SAMPLING-BASED EPISTASIS ALGORITHM FOR QUANTITATIVE GENETIC TRAIT PREDICTION

Dan He, Laxmi Parida

IBM Thomas J. Watson Research Center

Quantitative genetic trait prediction based on high-density genotyping arrays plays an important role for plant and animal breeding, as well as genetic epidemiology such as complex diseases. The prediction can be very helpful to develop breeding strategies and is crucial to translate the findings in genetics to precision medicine. Epistasis, the phenomena where the SNPs interact with each other, has been studied extensively in Genome Wide Association Studies (GWAS) but received relatively less attention for quantitative genetic trait prediction. As the number of possible interactions is generally extremely large, even pairwise interactions is very challenging. To our knowledge, there is no solid solution yet to utilize epistasis to improve genetic trait prediction. In this work, we studied the multi-locus epistasis problem where the interactions with more than two SNPs are considered. We developed an efficient algorithm MUSE to improve the genetic trait prediction with the help of multi-locus epistasis. MUSE is sampling-based and we proposed a few different sampling strategies. Our experiments on real data showed that MUSE is not only efficient but also effective to improve the genetic trait prediction. MUSE also achieved very significant improvements on a real plant data set as well as a real human data set.

DIFFERENTIAL PATHWAY DEPENDENCY DISCOVERY ASSOCIATED WITH DRUG RESPONSE ACROSS CANCER CELL LINES

Gil Speyer[1], Divya Mahendra[1], Hai J. Tran[1], Jeff Kiefer[1], Stuart L. Schreiber[2], Paul A. Clemons[2], Harshil Dhruv[1], Michael Berens[1], Seungchan Kim[1]

[1]The Translational Genomics Research Institute, [2]Broad Institute of Harvard and MIT

The effort to personalize treatment plans for cancer patients involves the identification of drug treatments that can effectively target the disease while minimizing the likelihood of adverse reactions. In this study, the gene-expression profile of 810 cancer cell lines and their response data to 368 small molecules from the Cancer Therapeutics Research Portal (CTRP) are analyzed to identify pathways with significant rewiring between genes, or differential gene dependency, between sensitive and non-sensitive cell lines. Identified pathways and their corresponding differential dependency networks are further analyzed to discover essentiality and specificity mediators of cell line response to drugs/compounds. For analysis we use the previously published method EDDY (Evaluation of Differential DependencY). EDDY first constructs likelihood distributions of gene-dependency networks, aided by known gene- gene interaction, for two given conditions, for example, sensitive cell lines vs. non-sensitive cell lines. These sets of networks yield a divergence value between two distributions of network likelihoods that can be assessed for significance using permutation tests. Resulting differential dependency networks were then further analyzed to identify genes, termed mediators, which may play important roles in biological signaling in certain cell lines that are sensitive or non-sensitive to the drugs. Establishing statistical correspondence between compounds and mediators can improve understanding of known gene dependencies associated with drug response while also discovering new dependencies. Millions of compute hours resulted in thousands of these statistical discoveries. EDDY identified 8,811 statistically significant pathways leading to 26,822 compound-pathway-mediator triplets. By incorporating STITCH and STRING databases, we could construct evidence networks for 14,415 compound-pathway-mediator triplets for support. The results of this analysis are presented in a searchable website to aid researchers in studying potential molecular mechanisms underlying cells' drug response as well as in designing experiments for the purpose of personalized treatment regimens.

A METHYLATION-TO-EXPRESSION FEATURE MODEL FOR GENERATING ACCURATE PROGNOSTIC RISK SCORES AND IDENTIFYING DISEASE TARGETS IN CLEAR CELL KIDNEY CANCER

Jeffrey A. Thompson[1], Carmen J. Marsit[2]

[1]Dartmouth College, [2]Emory University

Many researchers now have available multiple high-dimensional molecular and clinical datasets when studying a disease. As we enter this multi-omic era of data analysis, new approaches that combine different levels of data (e.g. at the genomic and epigenomic levels) are required to fully capitalize on this opportunity. In this work, we outline a new approach to multi-omic data integration, which combines molecular and clinical predictors as part of a single analysis to create a prognostic risk score for clear cell renal cell carcinoma. The approach integrates data in multiple ways and yet creates models that are relatively straightforward to interpret and with a high level of performance. Furthermore, the proposed process of data integration captures relationships in the data that represent highly disease-relevant functions.

DE NOVO MUTATIONS IN AUTISM IMPLICATE THE SYNAPTIC ELIMINATION NETWORK

Guhan Ram Venkataraman[1], Chloe O'Connell[1], Fumiko Egawa[2], Dorna Kashef-Haghighi[1], Dennis Paul Wall[1]

[1]Stanford University, [2]St. George's University

Autism has been shown to have a major genetic risk component; the architecture of documented autism in families has been over and again shown to be passed down for generations. While inherited risk plays an important role in the autistic nature of children, de novo (germline) mutations have also been implicated in autism risk. Here we find that autism de novo variants verified and published in the literature are Bonferroni-significantly enriched in a gene set implicated in synaptic elimination. Additionally, several of the genes in this synaptic elimination set that were enriched in protein-protein interactions (CACNA1C, SHANK2, SYNGAP1, NLGN3, NRXN1, and PTEN) have been previously confirmed as genes that confer risk for the disorder. The results demonstrate that autism-associated de novos are linked to proper synaptic pruning and density, hinting at the etiology of autism and suggesting pathophysiology for downstream correction and treatment.

IDENTIFYING GENETIC ASSOCIATIONS WITH VARIABILITY IN METABOLIC HEALTH AND BLOOD COUNT LABORATORY VALUES: DIVING INTO THE QUANTITATIVE TRAITS BY LEVERAGING LONGITUDINAL DATA FROM AN EHR

Shefali S. Verma[1], Anastasia M. Lucas[1], Daniel R. Lavage[1], Joseph B. Leader[1], Raghu Metpally[2], Sarathbabu Krishnamurthy[1], Frederick Dewey[1], Ingrid Borecki[1], Alexander Lopez[3], John Overton[3], John Penn[3], Jeffrey Reid[3], Sarah A. Pendergrass[1], Gerda Breitwieser[2], Marylyn D. Ritchie[1]

[1]Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA; [2]Department of Functional and Molecular Genomics, Geisinger Health System, Danville, PA; [3]Regeneron Genetics Center, Tarrytown, NY

A wide range of patient health data is recorded in Electronic Health Records (EHR). This data includes diagnosis, surgical procedures, clinical laboratory measurements, and medication information. Together this information reflects the patient's medical history. Many studies have efficiently used this data from the EHR to find associations that are clinically relevant, either by utilizing International Classification of Diseases, version 9 (ICD-9) codes or laboratory measurements, or by designing phenotype algorithms to extract case and control status with accuracy from the EHR. Here we developed a strategy to utilize longitudinal quantitative trait data from the EHR at Geisinger Health System focusing on outpatient metabolic and complete blood panel data as a starting point. Comprehensive Metabolic Panel (CMP) as well as Complete Blood Counts (CBC) are parts of routine care and provide a comprehensive picture from high level screening of patients' overall health and disease. We randomly split our data into two datasets to allow for discovery and replication. We first conducted a genome-wide association study (GWAS) with median values of 25 different clinical laboratory measurements to identify variants from Human Omni Express Exome beadchip data that are associated with these measurements. We identified 687 variants that associated and replicated with the tested clinical measurements at p<5x10-08. Since longitudinal data from the EHR provides a record of a patient's medical history, we utilized this information to further investigate the ICD-9 codes that might be associated with differences in variability of the measurements in the longitudinal dataset. We identified low and high variance patients by looking at changes within their individual longitudinal EHR laboratory results for each of the 25 clinical lab values (thus creating 50 groups – a high variance and a low variance for each lab variable). We then performed a PheWAS analysis with ICD-9 diagnosis codes, separately in the high variance group and the low variance group for each lab variable. We found 717 PheWAS associations that replicated at a p-value less than 0.001. Next, we evaluated the results of this study by comparing the association results between the high and low variance groups. For example, we found 39 SNPs (in multiple genes) associated with ICD-9 250.01 (Type-I diabetes) in patients with high variance of plasma glucose levels, but not in patients with low variance in plasma glucose levels. Another example is the association of 4 SNPs in UMOD with chronic kidney disease in patients with high variance for aspartate aminotransferase (discovery p-value: 8.71x10-09 and replication p-value: 2.03x10-06). In general, we see a pattern of many more statistically significant associations from patients with high variance in the quantitative lab variables, in comparison with the low variance group across all of the 25 laboratory measurements. This study is one of the first of its kind to utilize quantitative trait variance from longitudinal laboratory data to find associations among genetic variants and clinical phenotypes obtained from an EHR, integrating laboratory values and diagnosis codes to understand the genetic complexities of common diseases.

STRATEGIES FOR EQUITABLE PHARMACOGENOMIC-GUIDED WARFARIN DOSING AMONG EUROPEAN AND AFRICAN AMERICAN INDIVIDUALS IN A CLINICAL POPULATION

Laura Wiley[1], Jacob VanHouten[2], David Samuels[2], Melinda Aldrich[3], Dan Roden[2], Josh Peterson[2], Joshua Denny[2]

[1]University of Colorado, [2]Vanderbilt University, [3]Vanderbilt University Medical Center

The blood thinner warfarin has a narrow therapeutic range and high inter- and intra-patient variability in therapeutic doses. Several studies have shown that pharmacogenomic variants help predict stable warfarin dosing. However, retrospective and randomized controlled trials that employ dosing algorithms incorporating pharmacogenomic variants under perform in African Americans. This study sought to determine if: 1) including additional variants associated with warfarin dose in African Americans, 2) predicting within single ancestry groups rather than a combined population, or 3) using percentage African ancestry rather than observed race, would improve warfarin dosing algorithms in African Americans. Using BioVU, the Vanderbilt University Medical Center biobank linked to electronic medical records, we compared 25 modeling strategies to existing algorithms using a cohort of 2,181 warfarin users (1,928 whites, 253 blacks). We found that approaches incorporating additional variants increased model accuracy, but not in clinically significant ways. Race stratification increased model fidelity for African Americans, but the improvement was small and not likely to be clinically significant. Use of percent African ancestry improved model fit in the context of race misclassification.

SINGLE-CELL ANALYSIS AND MODELLING OF CELL POPULATION HETEROGENEITY

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

PRODUCTION OF A PRELIMINARY QUALITY CONTROL PIPELINE FOR SINGLE NUCLEI RNA-SEQ AND ITS APPLICATION IN THE ANALYSIS OF CELL TYPE DIVERSITY OF POST-MORTEM HUMAN BRAIN NEOCORTEX

Brian Aevermann[1], Jamison McCorrison[1], Pratap Venepally[1], Rebecca Hodge[2], Trygve Bakken[2], Jeremy Miller[2], Mark Novotny[1], Danny N. Tran[1], Francisco Diez-Fuertes[3], Lena Christiansen[4], Fan Zhang[4], Frank Steemers[4], Roger S. Lasken[1], Ed Lein[2], Nicholas Schork[1], Richard H. Scheuermann[1]

[1]J. Craig Venter Institute, [2]Allen Institute for Brain Science, [3]Instituto de Salud Carlos III, [4]Illumina, Inc.

Next generation sequencing of the RNA content of single cells or single nuclei (sc/nRNA-seq) has become a powerful approach to understand the cellular complexity and diversity of multicellular organisms and environmental ecosystems. However, the fact that the procedure begins with a relatively small amount of starting material, thereby pushing the limits of the laboratory procedures required, dictates that careful approaches for sample quality control (QC) are essential to reduce the impact of technical noise and sample bias in downstream analysis applications. Here we present a preliminary framework for sample level quality control that is based on the collection of a series of quantitative laboratory and data metrics that are used as features for the construction of QC classification models using random forest machine learning approaches. We've applied this initial framework to a dataset comprised of 2272 single nuclei RNA-seq results and determined that ~79% of samples were of high quality. Removal of the poor quality samples from downstream analysis was found to improve the cell type clustering results. In addition, this approach identified quantitative features related to the proportion of unique or duplicate reads and the proportion of reads remaining after quality trimming as useful features for pass/fail classification. The construction and use of classification models for the identification of poor quality samples provides for an objective and scalable approach to sc/nRNA-seq quality control.

TRACING CO-REGULATORY NETWORK DYNAMICS IN NOISY, SINGLE-CELL TRANSCRIPTOME TRAJECTORIES

Pablo Cordero, Joshua M. Stuart

UC Santa Cruz Genomics Institute, University of California, Santa Cruz

The availability of gene expression data at the single cell level makes it possible to probe the molecular underpinnings of complex biological processes such as differentiation and oncogenesis. Promising new methods have emerged for reconstructing a progression 'trajectory' from static single-cell transcriptome measurements. However, it remains unclear how to adequately model the appreciable level of noise in these data to elucidate gene regulatory network rewiring. Here, we present a framework called Single Cell Inference of MorphIng Trajectories and their Associated Regulation (SCIMITAR) that infers progressions from static single-cell transcriptomes by employing a continuous parametrization of Gaussian mixtures in high-dimensional curves. SCIMITAR yields rich models from the data that highlight genes with expression and co-expression patterns that are associated with the inferred progression. Further, SCIMITAR extracts regulatory states from the implicated trajectory-evolving co-expression networks. We benchmark the method on simulated data to show that it yields accurate cell ordering and gene network inferences. Applied to the interpretation of a single-cell human fetal neuron dataset, SCIMITAR finds progression-associated genes in cornerstone neural differentiation pathways missed by standard differential expression tests. Finally, by leveraging the rewiring of gene-gene co-expression relations across the progression, the method reveals the rise and fall of co-regulatory states and trajectory-dependent gene modules. These analyses implicate new transcription factors in neural differentiation including putative co-factors for the multi-functional NFAT pathway.

AN UPDATED DEBARCODING TOOL FOR MASS CYTOMETRY WITH CELL TYPE-SPECIFIC AND CELL SAMPLE-SPECIFIC STRINGENCY ADJUSTMENT

Kristin I. Fread[1], William D. Strickland[2], Garry P. Nolan[3], Eli R. Zunder[1]

[1]Department of Biomedical Engineering, University of Virginia; [2]Department of Biomedical Sciences, University of Virginia; [3]Department of Microbiology and Immunology, Stanford University

Pooled sample analysis by mass cytometry barcoding carries many advantages: reduced antibody consumption, increased sample throughput, removal of cell doublets, reduction of cross-contamination by sample carryover, and the elimination of tube-to-tube-variability in antibody staining. A single-cell debarcoding algorithm was previously developed to improve the accuracy and yield of sample deconvolution, but this method was limited to using fixed parameters for debarcoding stringency filtering, which could introduce cell-specific or sample-specific bias to cell yield in scenarios where barcode staining intensity and variance are not uniform across the pooled samples. To address this issue, we have updated the algorithm to output debarcoding parameters for every cell in the sample-assigned FCS files, which allows for visualization and analysis of these parameters via flow cytometry analysis software. This strategy can be used to detect cell type-specific and sample-specific effects on the underlying cell data that arise during the debarcoding process. An additional benefit to this strategy is the decoupling of barcode stringency filtering from the debarcoding and sample assignment process. This is accomplished by removing the stringency filters during sample assignment, and then filtering after the fact with 1- and 2-dimensional gating on the debarcoding parameters which are output with the FCS files. These data exploration strategies serve as an important quality check for barcoded mass cytometry datasets, and allow cell type and sample-specific stringency adjustment that can remove bias in cell yield introduced during the debarcoding process.

IMAGING GENOMICS

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

ADAPTIVE TESTING OF SNP-BRAIN FUNCTIONAL CONNECTIVITY ASSOCIATION VIA A MODULAR NETWORK ANALYSIS

Chen Gao, Junghi Kim, Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota

Due to its high dimensionality and high noise levels, analysis of a large brain functional network may not be powerful and easy to interpret; instead, decomposition of a large network into smaller subcomponents called modules may be more promising as suggested by some empirical evidence. For example, alteration of brain modularity is observed in patients suffering from various types of brain malfunctions. Although several methods exist for estimating brain functional networks, such as the sample correlation matrix or graphical lasso for a sparse precision matrix, it is still difficult to extract modules from such network estimates. Motivated by these considerations, we adapt a weighted gene co-expression network analysis (WGCNA) framework to resting-state fMRI (rs-fMRI) data to identify modular structures in brain functional networks. Modular structures are identified by using topological overlap matrix (TOM) elements in hierarchical clustering. We propose applying a new adaptive test built on the proportional odds model (POM) that can be applied to a high-dimensional setting, where the number of variables (p) can exceed the sample size (n) in addition to the usual $p < n$ setting. We applied our proposed methods to the ADNI data to test for associations between a genetic variant and either the whole brain functional network or its various subcomponents using various connectivity measures. We uncovered several modules based on the control cohort, and some of them were marginally associated with the APOE4 variant and several other SNPs; however, due to the small sample size of the ADNI data, larger studies are needed.

EXPLORING BRAIN TRANSCRIPTOMIC PATTERNS: A TOPOLOGICAL ANALYSIS USING SPATIAL EXPRESSION NETWORKS

Zhana Kuncheva[1], Michelle L. Krishnan[2], Giovanni Montana[2]

[1]Imperial College London, [2]King's College London

Characterizing the transcriptome architecture of the human brain is fundamental in gaining an understanding of brain function and disease. A number of recent studies have investigated patterns of brain gene expression obtained from an extensive anatomical coverage across the entire human brain using experimental data generated by the Allen Human Brain Atlas (AHBA) project. In this paper, we propose a new representation of a gene's transcription activity that explicitly captures the pattern of spatial co-expression across different anatomical brain regions. For each gene, we define a Spatial Expression Network (SEN), a network quantifying co-expression patterns amongst several anatomical locations. Network similarity measures are then employed to quantify the topological resemblance between pairs of SENs and identify naturally occurring clusters. Using network-theoretical measures, three large clusters have been detected featuring distinct topological properties. We then evaluate whether topological diversity of the SENs reflects significant differences in biological function through a gene ontology analysis. We report on evidence suggesting that one of the three SEN clusters consists of genes specifically involved in the nervous system, including genes related to brain disorders, while the remaining two clusters are representative of immunity, transcription and translation. These findings are consistent with previous studies showing that brain gene clusters are generally associated with one of these three major biological processes.

PATTERNS IN BIOMEDICAL DATA – HOW DO WE FIND THEM?

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

A DEEP LEARNING APPROACH FOR CANCER DETECTION AND RELEVANT GENE IDENTIFICATION

Padideh Danaee, Reza Ghaeini, David Hendrix

Oregon State University

Cancer detection from gene expression data continues to pose a challenge due to the high dimensionality and complexity of these data. After decades of research there is still uncertainty in the clinical diagnosis of cancer and the identification of tumor-specific markers. Here we present a deep learning approach to cancer detection, and to the identification of genes critical for the diagnosis of breast cancer. First, we used Stacked Denoising Autoencoder (SDAE) to deeply extract functional features from high dimensional gene expression profiles. Next, we evaluated the performance of the extracted representation through supervised classification models to verify the usefulness of the new features in cancer detection. Lastly, we identified a set of highly interactive genes by analyzing the SDAE connectivity matrices. Our results and analysis illustrate that these highly interactive genes could be useful cancer biomarkers for the detection of breast cancer that deserve further studies.

GENOME-WIDE INTERACTION WITH SELECTED TYPE 2 DIABETES LOCI REVEALS NOVEL LOCI FOR TYPE 2 DIABETES IN AFRICAN AMERICANS

Jacob M. Keaton[1], Jacklyn N. Hellwege[1], Maggie C. Y. Ng[1], Nicholette D. Palmer[1], James S. Pankow[2], Myriam Fornage[3], James G. Wilson[4], Adolofo Correa[4], Laura J. Rasmussen-Torvik[5], Jerome I. Rotter[6], Yii-Der I. Chen[6], Kent D. Taylor[6], Stephen S. Rich[7], Lynne E. Wagenknecht[1], Barry I. Freedman[1], Donald W. Bowden[1]

[1]Wake Forest School of Medicine, [2]University of Minnesota, [3]University of Texas Health Science Center at Houston, [4]University of Mississippi Medical Center, [5]Northwestern University Feinberg School of Medicine, [6]Harbor-UCLA Medical Center, [7]University of Virginia

Type 2 diabetes (T2D) is the result of metabolic defects in insulin secretion and insulin sensitivity, yet most T2D loci identified to date influence insulin secretion. We hypothesized that T2D loci, particularly those affecting insulin sensitivity, can be identified through interaction with known T2D loci implicated in insulin secretion. To test this hypothesis, single nucleotide polymorphisms (SNPs) nominally associated with acute insulin response to glucose (AIRg), a dynamic measure of first-phase insulin secretion, and previously associated with T2D in genome-wide association studies (GWAS) were identified in African Americans from the Insulin Resistance Atherosclerosis Family Study (IRASFS; n=492 subjects). These SNPs were tested for interaction, individually and jointly as a genetic risk score (GRS), using GWAS data from five cohorts (ARIC, CARDIA, JHS, MESA, WFSM; n=2,725 cases, 4,167 controls) with T2D as the outcome. In single variant analyses, suggestively significant (Pinteraction < 5x10-6) interactions were observed at several loci including DGKB (rs978989), CDK18 (rs12126276), CXCL12 (rs7921850), HCN1 (rs6895191), FAM98A (rs1900780), and MGMT (rs568530). Notable beta-cell GRS interactions included two SNPs at the DGKB locus (rs6976381; rs6962498). These data support the hypothesis that additional genetic factors contributing to T2D risk can be identified by interactions with insulin secretion loci.

META-ANALYSIS OF CONTINUOUS PHENOTYPES IDENTIFIES A GENE SIGNATURE THAT CORRELATES WITH COPD DISEASE STATUS

Madeleine Scott[1], Francesco Vallania[2], Purvesh Khatri[3]

[1]Stanford Medical School, Stanford University, Stanford, California; [2]Stanford Institute for Immunity, Transplantation, and Infection, Stanford University, Stanford, California; [3]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California

The utility of multi-cohort two-class meta-analysis to identify robust differentially expressed gene signatures has been well established. However, many biomedical applications, such as gene signatures of disease progression, require one-class analysis. Here we describe an R package, MetaCorrelator, that can identify a reproducible transcriptional signature that is correlated with a continuous disease phenotype across multiple datasets. We successfully applied this framework to extract a pattern of gene expression that can predict lung function in patients with chronic obstructive pulmonary disease (COPD) in both peripheral blood mononuclear cells (PBMCs) and tissue. Our results point to a disregulation in the oxidation state of the lungs of patients with COPD, as well as underscore the classically recognized inflammatory state that underlies this disease.

LEARNING PARSIMONIOUS ENSEMBLES FOR UNBALANCED COMPUTATIONAL GENOMICS PROBLEMS

Ana Stanescu, Gaurav Pandey

Icahn School of Medicine at Mount Sinai

Prediction problems in biomedical sciences are generally quite difficult, partially due to incomplete knowledge of how the phenomenon of interest is influenced by the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor(s) for specific problems. In these situations, a powerful approach to improving prediction performance is to construct ensembles that combine the outputs of many individual base predictors, which have been successful for many biomedical prediction tasks. Moreover, selecting a {\it parsimonious} ensemble can be of even greater value for biomedical sciences, where it is not only important to learn an accurate predictor, but also to interpret what novel knowledge it can provide about the target problem. Ensemble selection is a promising approach for this task because of its ability to select a collectively predictive subset, often a relatively small one, of all input base predictors. One of the most well-known algorithms for ensemble selection, CES (Caruana {\it et al.}'s Ensemble Selection), generally performs well in practice, but faces several challenges due to the difficulty of choosing the right values of its various parameters. Since the choices made for these parameters are usually ad-hoc, good performance of CES is difficult to guarantee for a variety of problems or datasets. To address these challenges with CES and other such algorithms, we propose a novel heterogeneous ensemble selection approach based on the paradigm of reinforcement learning (RL), which offers a more systematic and mathematically sound methodology for exploring the many possible combinations of base predictors that can be selected into an ensemble. We develop three RL-based strategies for constructing ensembles and analyze their results on two unbalanced computational genomics problems, namely the prediction of protein function and splice sites in eukaryotic genomes. We show that the resultant ensembles are indeed substantially more parsimonious as compared to the full set of base predictors, yet still offer almost the same classification power, especially for larger datasets. The RL ensembles also yield a better combination of parsimony and predictive performance as compared to CES.

NETWORK MAP OF ADVERSE HEALTH EFFECTS AMONG VICTIMS OF INTIMATE PARTNER VIOLENCE

Kathleen Whiting[1], Larry Y. Liu[2], Mehmet Koyutürk [2], Gunnur Karakurt[2]

[1]Uniformed Services University, [2]Case Western Reserve University

Intimate partner violence (IPV) is a serious problem with devastating health consequences. Screening procedures may overlook relationships between IPV and negative health effects. To identify IPV-associated women's health issues, we mined national, aggregated de-identified electronic health record data and compared female health issues of domestic abuse (DA) versus non-DA records, identifying terms significantly more frequent for the DA group. After coding these terms into 28 broad categories, we developed a network map to determine strength of relationships between categories in the context of DA, finding that acute conditions are strongly connected to cardiovascular, gastrointestinal, gynecological, and neurological conditions among victims.

PRECISION MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS IMPROVED HEALTH AND THERAPIES

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

A POWERFUL METHOD FOR INCLUDING GENOTYPE UNCERTAINTY IN TESTS OF HARDY-WEINBERG EQUILIBRIUM

Andrew Beck[1], Alexander Luedtke[2], Keli Liu[3], Nathan Tintle[4]

[1]University of Michigan, [2]University of California - Berkeley, [3]Harvard University, [4]Dordt College

The use of posterior probabilities to summarize genotype uncertainty is pervasive across genotype, sequencing and imputation platforms. Prior work in many contexts has shown the utility of incorporating genotype uncertainty (posterior probabilities) in downstream statistical tests. Typical approaches to incorporating genotype uncertainty when testing Hardy-Weinberg equilibrium tend to lack calibration in the type I error rate, especially as genotype uncertainty increases. We propose a new approach in the spirit of genomic control that properly calibrates the type I error rate, while yielding improved power to detect deviations from Hardy-Weinberg Equilibrium. We demonstrate the improved performance of our method on both simulated and real genotypes.

MICRORNA-AUGMENTED PATHWAYS (MIRAP) AND THEIR APPLICATIONS TO PATHWAY ANALYSIS AND DISEASE SUBTYPING

Diana Diaz[1], Michele Donato[2], Tin Nguyen[1], Sorin Draghici[1]

[1]Wayne State University, [2]Stanford University Medical Center

MicroRNAs play important roles in the development of many complex diseases. Because of their importance, the analysis of signaling pathways including miRNA interactions holds the potential for unveiling the mechanisms underlying such diseases. However, current signaling pathway databases are limited to interactions between genes and ignore miRNAs. Here, we use the information on miRNA targets to build a database of miRNA-augmented pathways (mirAP), and we show its application in the contexts of integrative pathway analysis and disease subtyping. Our miRNA-mRNA integrative pathway analysis pipeline incorporates a topology-aware approach that we previously implemented. Our integrative disease subtyping pipeline takes into account survival data, gene and miRNA expression, and knowledge of the interactions among genes. We demonstrate the advantages of our approach by analyzing nine sample-matched datasets that provide both miRNA and mRNA expression. We show that integrating miRNAs into pathway analysis results in greater statistical power, and provides a more comprehensive view of the underlying phenomena. We also compare our disease subtyping method with the state-of-the-art integrative analysis by analyzing a colorectal cancer database from TCGA. The colorectal cancer subtypes identified by our approach are significantly different in terms of their survival expectation. These miRNA-augmented pathways offer a more comprehensive view and a deeper understanding of biological pathways. A better understanding of the molecular processes associated with patients' survival can help to a better prognosis and an appropriate treatment for each subtype.

FREQUENT SUBGRAPH MINING OF PERSONALIZED SIGNALING PATHWAY NETWORKS GROUPS PATIENTS WITH FREQUENTLY DYSREGULATED DISEASE PATHWAYS AND PREDICTS PROGNOSIS

Arda Durmaz, Tim A.D. Henderson, Douglas Brubaker, Gurkan Bebek

Case Western Reserve University

Motivation: Large scale genomics studies have generated comprehensive molecular characterization of numerous cancer types. Subtypes for many tumor types have been established; however, these classifications are based on molecular characteristics of a small gene sets with limited power to detect dysregulation at the patient level. We hypothesize that frequent graph mining of pathways to gather pathways functionally relevant to tumors can characterize tumor types and provide opportunities for personalized therapies. Results: In this study we present an integrative omics approach to group patients based on their altered pathway characteristics and show prognostic differences within breast cancer ($p < 9.57E-10$) and glioblastoma multiforme ($p < 0.05$) patients. We were able validate this approach in secondary RNA-Seq datasets with $p < 0.05$ and $p < 0.01$ respectively. We also performed pathway enrichment analysis to further investigate the biological relevance of dysregulated pathways. We compared our approach with network-based classifier algorithms and showed that our unsupervised approach generates more robust and biologically relevant clustering whereas previous approaches failed to report specific functions for similar patient groups or classify patients into prognostic groups. Conclusions: These results could serve as a means to improve prognosis for future cancer patients, and to provide opportunities for improved treatment options and personalized interventions. The proposed novel graph mining approach is able to integrate PPI networks with gene expression in a biologically sound approach and cluster patients in to clinically distinct groups. We have utilized breast cancer and glioblastoma multiforme datasets from microarray and RNA-Seq platforms and identified disease mechanisms differentiating samples.

CERNA SEARCH METHOD IDENTIFIED A MET-ACTIVATED SUBGROUP AMONG EGFR DNA AMPLIFIED LUNG ADENOCARCINOMA PATIENTS

Halla Kabat, Leo Tunkle, Inhan Lee

miRcore

Given the diverse molecular pathways involved in tumorigenesis, identifying subgroups among cancer patients is crucial in precision medicine. While most targeted therapies rely on DNA mutation status in tumors, responses to such therapies vary due to the many molecular processes involved in propagating DNA changes to proteins (which constitute the usual drug targets). Though RNA expressions have been extensively used to categorize tumors, identifying clinically important subgroups remains challenging given the difficulty of discerning subgroups within all possible RNA-RNA networks. It is thus essential to incorporate multiple types of data. Recently, RNA was found to regulate other RNA through a common microRNA (miR). These regulating and regulated RNAs are referred to as competing endogenous RNAs (ceRNAs). However, global correlations between mRNA and miR expressions across all samples have not reliably yielded ceRNAs. In this study, we developed a ceRNA-based method to identify subgroups of cancer patients combining DNA copy number variation, mRNA expression, and microRNA (miR) expression data with biological knowledge. Clinical data is used to validate identified subgroups and ceRNAs. Since ceRNAs are causal, ceRNA-based subgroups may present clinical relevance. Using lung adenocarcinoma data from The Cancer Genome Atlas (TCGA) as an example, we focused on EGFR amplification status, since a targeted therapy for EGFR exists. We hypothesized that global correlations between mRNA and miR expressions across all patients would not reveal important subgroups and that clustering of potential ceRNAs might define molecular pathway-relevant subgroups. Using experimentally validated miR-target pairs, we identified EGFR and MET as potential ceRNAs for miR-133b in lung adenocarcinoma. The EGFR-MET up and miR-133b down subgroup showed a higher death rate than the EGFR-MET down and miR-133b up subgroup. Although transactivation between MET and EGFR has been identified previously, our result is the first to propose ceRNA as one of its underlying mechanisms. Furthermore, since MET amplification was seen in the case of resistance to EGFR-targeted therapy, the EGFR-MET up and miR-133b down subgroup may fall into the drug non-response group and thus preclude EGFR target therapy.

IMPROVED PERFORMANCE OF GENE SET ANALYSIS ON GENOME-WIDE TRANSCRIPTOMICS DATA WHEN USING GENE ACTIVITY STATE ESTIMATES

Thomas Kamp, Micah Adams, Craig Disselkoen, Nathan Tintle

Dordt College

Gene set analysis methods continue to be a popular and powerful method of evaluating genome-wide transcriptomics data. These approach require a priori grouping of genes into biologically meaningful sets, and then conducting downstream analyses at the set (instead of gene) level of analysis. Gene set analysis methods have been shown to yield more powerful statistical conclusions than single-gene analyses due to both reduced multiple testing penalties and potentially larger observed effects due to the aggregation of effects across multiple genes in the set. Traditionally, gene set analysis methods have been applied directly to normalized, log-transformed, transcriptomics data. Recently, efforts have been made to transform transcriptomics data to scales yielding more biologically interpretable results. For example, recently proposed models transform log-transformed transcriptomics data to a confidence metric (ranging between 0 and 100%) that a gene is active (roughly speaking, that the gene product is part of an active cellular mechanism). In this manuscript, we demonstrate, on both real and simulated transcriptomics data, that tests for differential expression between sets of genes using are typically more powerful when using gene activity state estimates as opposed to log-transformed gene expression data. Our analysis suggests further exploration of techniques to transform transcriptomics data to meaningful quantities for improved downstream inference.

METHYLDMV: SIMULTANEOUS DETECTION OF DIFFERENTIAL DNA METHYLATION AND VARIABILITY WITH CONFOUNDER ADJUSTMENT

Pei Fen Kuan, Junyan Song, Shuyao He

Stony Brook University

DNA methylation has emerged as promising epigenetic markers for disease diagnosis. Both the differential mean (DM) and differential variability (DV) in methylation have been shown to contribute to transcriptional aberration and disease pathogenesis. The presence of confounding factors in large scale EWAS may affect the methylation values and hamper accurate marker discovery. In this paper, we propose a flexible framework called methylDMV which allows for confounding factors adjustment and enables simultaneous characterization and identification of CpGs exhibiting DM only, DV only and both DM and DV. The proposed framework also allows for prioritization and selection of candidate features to be included in the prediction algorithm. We illustrate the utility of methylDMV in several TCGA datasets. An R package methylDMV implementing our proposed method is available at http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV.

IDENTIFY CANCER DRIVER GENES THROUGH SHARED MENDELIAN DISEASE PATHOGENIC VARIANTS AND CANCER SOMATIC MUTATIONS

Meng Ma[1], Changchang Wang[2], Benjamin Glicksberg[1], Eric E. Schadt[1], Shuyu Li[1], Rong Chen[1]

[1]Icahn School of Medicine at Mount Sinai, [2]Anhui University

Genomic sequencing studies in the past several years have yielded a large number of cancer somatic mutations. There remains a major challenge in delineating a small fraction of somatic mutations that are oncogenic drivers from a background of predominantly passenger mutations. Although computational tools have been developed to predict the functional impact of mutations, their utility is limited. In this study, we applied an alternative approach to identify potentially novel cancer drivers as those somatic mutations that overlap with known pathogenic mutations in Mendelian diseases. We hypothesize that those shared mutations are more likely to be cancer drivers because they have the established molecular mechanisms to impact protein functions. We first show that the overlap between somatic mutations in COSMIC and pathogenic genetic variants in HGMD is associated with high mutation frequency in cancers and is enriched for known cancer genes. We then attempted to identify putative tumor suppressors based on the number of distinct HGMD/COSMIC overlapping mutations in a given gene, and our results suggest that ion channels, collagens and Marfan syndrome associated genes may represent new classes of tumor suppressors. To elucidate potentially novel oncogenes, we identified those HGMD/COSMIC overlapping mutations that are not only highly recurrent but also mutually exclusive from previously characterized oncogenic mutations in each specific cancer type. Taken together, our study represents a novel approach to discover new cancer genes from the vast amount of cancer genome sequencing data.

IDENTIFYING CANCER SPECIFIC METABOLIC SIGNATURES USING CONSTRAINT-BASED MODELS

André Schultz[1], Sanket Mehta[1], Chenyue W. Hu[1], Fieke W. Hoff[2], Terzah M. Horton[3], Steven M. Kornblau[2], Amina A. Qutub[1]

[1]Rice University, [2]University of Texas MD Anderson Cancer Center, [3]Baylor College of Medicine and Texas Children's Hospital

Cancer metabolism differs remarkably from the metabolism of healthy surrounding tissues, and it is extremely heterogeneous across cancer types. While these metabolic differences provide promising avenues for cancer treatments, much work remains to be done in understanding how metabolism is rewired in malignant tissues. To that end, constraint-based models provide a powerful computational tool for the study of metabolism at the genome scale. To generate meaningful predictions, however, these generalized human models must first be tailored for specific cell or tissue sub-types. Here we first present two improved algorithms for (1) the generation of these context-specific metabolic models based on omics data, and (2) Monte-Carlo sampling of the metabolic model flux space. By applying these methods to generate and analyze context-specific metabolic models of diverse solid cancer cell line data, and primary leukemia pediatric patient biopsies, we demonstrate how the methodology presented in this study can generate insights into the rewiring differences across solid tumors and blood cancers.

SINGLE-CELL ANALYSIS AND MODELLING OF CELL POPULATION HETEROGENEITY

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

MAPPING NEURONAL CELL TYPES USING INTEGRATIVE MULTI-SPECIES MODELING OF HUMAN AND MOUSE SINGLE CELL RNA SEQUENCING

Travis Johnson, Zachary Abrams, Yan Zhang, Kun Huang

Ohio State University

Mouse brain transcriptomic studies are important in the understanding of the structural heterogeneity in the brain. However, it is not well understood how cell types in the mouse brain relate to human brain cell types on a cellular level. We propose that it is possible with single cell granularity to find concordant genes between mouse and human and that these genes can be used to separate cell types across species. We show that a set of concordant genes can be algorithmically derived from a combination of human and mouse single cell sequencing data. Using this gene set, we show that similar cell types shared between mouse and human cluster together. Furthermore we find that previously unclassified human cells can be mapped to the glial/vascular cell type by integrating mouse cell type expression profiles.

A SPATIOTEMPORAL MODEL TO SIMULATE CHEMOTHERAPY REGIMENS FOR HETEROGENEOUS BLADDER CANCER METASTASES TO THE LUNG

Kimberly R. Kanigel Winner[1], James C. Costello[2]

[1]Computational Bioscience Program, Department of Pharmacology, Univeristy of Colorado Cancer Center; [2]University of Colorado Anschutz Medical Campus

Tumors are composed of heterogeneous populations of cells. Somatic genetic aberrations are one form of heterogeneity that allows clonal cells to adapt to chemotherapeutic stress, thus providing a path for resistance to arise. In silico modeling of tumors provides a platform for rapid, quantitative experiments to inexpensively study how compositional heterogeneity contributes to drug resistance. Accordingly, we have built a spatiotemporal model of a lung metastasis originating from a primary bladder tumor, incorporating in vivo drug concentrations of first-line chemotherapy, resistance data from bladder cancer cell lines, vascular density of lung metastases, and gains in resistance in cells that survive chemotherapy. In metastatic bladder cancer, a first-line drug regimen includes six cycles of gemcitabine plus cisplatin (GC) delivered simultaneously on day 1, and gemcitabine on day 8 in each 21-day cycle. The interaction between gemcitabine and cisplatin has been shown to be synergistic in vitro, and results in better outcomes in patients. Our model shows that during simulated treatment with this regimen, GC synergy does begin to kill cells that are more resistant to cisplatin, but repopulation by resistant cells occurs. Post-regimen populations are mixtures of the original, seeded resistant clones, and/or new clones that have gained resistance to cisplatin, gemcitabine, or both drugs. The emergence of a tumor with increased resistance is qualitatively consistent with the five-year survival of 6.8% for patients with metastatic transitional cell carcinoma of the urinary bladder treated with a GC regimen. The model can be further used to explore the parameter space for clinically relevant variables, including the timing of drug delivery to optimize cell death, and patient-specific data such as vascular density, rates of resistance gain, disease progression, and molecular profiles, and can be expanded for data on toxicity. The model is specific to bladder cancer, which has not previously been modeled in this context, but can be adapted to represent other cancers.

SCALABLE VISUALIZATION FOR HIGH-DIMENSIONAL SINGLE-CELL DATA

Juho Kim, Nate Russell, Jian Peng

University of Illinois at Urbana-Champaign

Single-cell analysis can uncover the mysteries in the state of individual cells and enable us to construct new models about the analysis of heterogeneous tissues. State-of-the-art technologies for single-cell analysis have been developed to measure the properties of single-cells and detect hidden information. They are able to provide the measurements of dozens of features simultaneously in each cell. However, due to the high-dimensionality, heterogeneous complexity and sheer enormity of single-cell data, its interpretation is challenging. Thus, new methods to overcome high-dimensionality are necessary. Here, we present a computational tool that allows efficient visualization of high-dimensional single-cell data onto a low-dimensional (2D or 3D) space while preserving the similarity structure between single-cells. We first construct a network that can represent the similarity structure between the high-dimensional representations of single-cells, and then, embed this network into a low-dimensional space through an efficient online optimization method based on the idea of negative sampling. Using this approach, we can preserve the high-dimensional structure of single-cell data in an embedded low-dimensional space that facilitates visual analyses of the data.

COMPUTATIONAL APPROACHES TO UNDERSTANDING THE EVOLUTION OF MOLECULAR FUNCTION

POSTER PRESENTATIONS

CLUSTER-BASED GENOTYPE-ENVIRONMENT-PHENOTYPE CORRELATION ALGORITHM

Ernesto Borrayo, Ryoko Machida-Hirano

Gene Research Center, University of Tsukuba

The interactions between genotype and environment give rise to phenotypic plasticity. However, these interactions are dynamic and complex. What is considered as a phenotype at one evaluation, can be considered as an environmental condition at some other, as that previous phenotype will affect particular conditions for the new one. Also, under a specific perspective a determined genetic material can be considered as an environmental condition for other loci. These concepts elucidate that the "one gene, one trait" rationale is rather the exception than the rule, and in order to adequately predict the possible phenotype expected at any biological level, the specific interaction between environment and genotype should be analyzed carefully. In order to infer the degree of influence of both a genotype and an environment over certain phenotypic traits, we developed a cluster-based algorithm that renders the way phenotypical traits can be explained by either that genotype or such environmental conditions. Although this approach is still far from being able to consider all possible aspects that may explain a phenotypic condition, it is a first approach to successfully analyzing the mentioned genotype-environment-phenotype interactions in a comprehensive manner. To test the algorithm along with synthetic data, real genetic, environmental and agromorphological traits of Theobroma cacao and Sechium edule were also analyzed. We expect that further exploration of different classifiers will help to adequately predict phenotypic expression at different biological levels—with significant applications in diverse fields such as crop improvement, genomics, clinical diagnosis/prognosis/treatment and metabolomics—and that it will enhance our understanding of genomics, metabolomics and adaptation/evolutionary processes.

QUANTITATING TRANSLATIONAL CONTROL: MRNA ABUNDANCE - DEPENDENT AND INDEPENDENT CONTRIBUTIONS

Jingyi Jessica Li[1], Guo-Liang Chew[2], Mark D. Biggin[3]

[1]Department of Statistics and Department of Human Genetics, UCLA; [2]Computational Biology Program, Fred Hutchinson Cancer Research Center; [3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory

Translation rate per mRNA molecule correlates positively with mRNA abundance. As a result, protein levels do not scale linearly with mRNA levels, but instead scale with the abundance of mRNA raised to the power of an "amplification exponent". Here we show that to quantitate translational control it is necessary to decompose the translation rate into two components. One component, TRmD, depends on the mRNA level and defines the amplification exponent. The other component, TRmIND, is independent of mRNA amount and impacts the correlation coefficient between protein and mRNA levels. We show that in S. cerevisiae TRmD represents ~30% of the variance in translation and results in an amplification exponent of ~1.20–1.27. TRmIND constitutes the remaining 70% of the variance in translation and explains <5% of the variance in protein expression. When protein degradation is also considered, the correlation between the abundances of protein and mRNA is R2prot–RNA > 0.92. We also investigate which mRNA sequence elements explain the variance in TRmD and TRmIND. We find that TRmIND is most strongly determined by the length of the open reading frame, while TRmD is more strongly determined by an A rich, highly unfolded element that spans nucleotides -35 to +28 relative to the initiating AUG codon, implying that TRmIND is under different evolutionary selective pressures than TRmD. Our work introduces methods for correctly scaling mRNA and protein abundance data using internally controlled standards. It provides quite different, more accurate estimates of translational control than any previous. By decomposing translation rates, we also provide insights into the mRNA sequence dependencies of translation that would not be apparent otherwise.

PROSNET: INTEGRATING HOMOLOGY WITH MOLECULAR NETWORKS FOR PROTEIN FUNCTION PREDICTION

Sheng Wang, Meng Qu, Jian Pen

University of Illinois Urbana Champaign

Automated annotation of protein function has become a critical task in the post-genomic era. Network-based approaches and homology-based approaches have been widely used and recently tested in large-scale community-wide assessment experiments. It is natural to integrate network data with homology information to further improve the predictive performance. However, integrating these two heterogeneous, high-dimensional and noisy datasets is non-trivial. In this work, we introduce a novel protein function prediction algorithm ProSNet. An integrated heterogeneous network is first built to include molecular networks of multiple species and link together homologous proteins across multiple species. Based on this integrated network, a dimensionality reduction algorithm is introduced to obtain compact low-dimensional vectors to encode proteins in the network. Finally, we develop machine learning classification algorithms that take the vectors as input and make predictions by transferring annotations both within each species and across different species. Extensive experiments on five major species demonstrate that our integration of homology with molecular networks substantially improves the predictive performance over existing approaches.

GENERAL

POSTER PRESENTATIONS

IDENTIFICATION OF DIFFERENTIALLY PHOSPHORYLATED MODULES IN PROTEIN INTERACTION NETWORKS

Marzieh Ayati, Danica Wiredja, Daniela Schlatzer, Goutham Narla, Mark Chance, Mehmet Koyutürk

Case Western Reserve University

Advances in high-throughput omics technologies revolutionized our understanding of the genomic underpinnings of cancer. However, many challenges remain in understanding how patients with common driver mutations may display diverging phosphoproteomic responses to the same treatment. Thus, an examination of the signaling landscape will provide essential molecular information for modeling personalized patient treatment design. However, integrative bioinformatics approaches to identify phosphoproteomics-based molecular states are in their infancy. To address this challenge, we adapt our algorithm MoBaS, which has been originally developed to identify phenotype-associated subnetworks in the context of genome-wide association studies. MoBaS takes as input a PPI network and a score for each protein indicating the protein's differential phosphorylation level. It then identifies protein subnetworks that are (i) composed of densely interacting proteins, and (ii) enriched in proteins with high scores. MoBaS also assesses the statistical significance of the identified subnetworks using permutation tests that effectively handle multiple hypothesis testing. We apply MoBaS to compare and contrast the drug-induced global signaling alterations of two KRAS mutated non-small cell lung cancer (NSCLC) cell lines, A549 and H358, treated with a novel activator of the tumor suppressor Protein Phosphatase 2A (PP2A) versus DMSO control. Applying kinase enrichment analysis on identified subnetworks, we identify Aurora KB as a key kinase differentially regulated between the two cell lines in response to our compound. Further corroborating this finding, we show that Aurora KB is downregulated at the protein and mRNA levels with our treatment in A549 but not in H358.

CLUSTERING METHOD FOR PRIORITIZING BREAST CANCER RISK GENES AND MIRNAS

Yongsheng Bai, Naureen Aslam, Ali Salman

Indiana State University

Background MicroRNAs (miRNA) are short nucleotides that interact with their target mRNAs through 3' untranslated regions (UTRs). The Cancer Genome Atlas (TCGA) project initiated in 2006 has achieved to sequence tissue collection with matched tumor and normal samples from 11,000 patients in 33 cancer types and subtypes, including 10 rare cancers. There is an urgent need to develop innovative methodologies and tools that can cluster mRNA-miRNA interaction pairs into groups and characterize functional consequences of cancer risk genes while analyzing the tumor and normal samples simultaneously. Rationale An undirected graph can be used to represent gene and miRNA relationships in an interaction network. Specifically, interactions between genes and miRNAs are rendered as a bipartite graph with genes or miRNAs as vertices and their calculated correlation as edges. Our hypothesis is: If a highly scored gene/miRNA cluster in a given tumor sample shows a significantly altered regulation relative to a similar gene/miRNA cluster in the corresponding non-tumor sample, the cluster is biologically significant. Results We developed a powerful mathematical model to identify clusters of significant mRNA and miRNA interaction pairs and decipher mRNA and miRNA regulation network using TCGA miRNA sequencing and mRNA sequencing data. We ran the cluster detection algorithm implemented in Python3 on TCGA Breast Invasive Carcinoma (BRCA) transcriptome (both RNA-Seq and miRNA-Seq) data sets. Using different cluster size (or bin) and different selection of miRNA and mRNA pairs for creating clusters will generate different topology of clusters, therefore, resulting in different numbers of common clusters between tumor and normal samples as well. We ran 1,000 different random selections of target pairs to generate different cluster topology and combined all results together to obtain 105,850 distinctive candidate clusters for prioritization. Conclusions We think our methodology for identifying cancer driver genes in personal genomes in which clinicians seek to develop better treatment strategies is valuable to the field. Our proposed method should be applicable across a range of diseases and cancers.

FUSIONDB: ASSESSING MICROBIAL DIVERSITY AND ENVIRONMENTAL PREFERENCES VIA FUNCTIONAL SIMILARITY

Chengsheng Zhu[1], Yannick Mahlich[1, 2, 3, 4], Yana Bromberg[1, 4]

[1]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ, USA; [2]Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), TUM, Garching, Germany; [3]Department of Informatics, Bioinformatics & Computational Biology - I12, TUM, Garching, Germany; [4]Institute of Advanced Study (TUM-IAS), Garching, Germany

Summary: Microbial functional diversification is driven by environmental factors. In some cases, microbes differ more across environments than across taxa. Here we introduce fusionDB, a novel database of microbial functional similarities, indexed by available environmental preferences. fusionDB entries represent nearly fourteen hundred taxonomically-distinct bacteria annotated with available metadata: habitat, temperature, and oxygen use. Each microbe is encoded as a set of functions represented by its proteome, and individual microbes are connected via common functions. Database searches produce easily visualizable XML-formatted network files of selected organisms, along with their shared functions. fusionDB thus provides a fast means of associating specific environmental factors with organism functions.
Availability: http://bromberglab.org/databases/fusiondb and as a sql-dump by request.
Contact: czhu@bromberglab.org, ymahlich@bromberglab.org

THE GEORGE M. O'BRIEN KIDNEY TRANSLATIONAL CORE CENTER AT THE UNIVERSITY OF MICHIGAN

Frank C. Brosius[1], Wenjun Ju[1], Keith Bellovich[2], Zeenat Bhat[3], Crystal Gadegbeku[4], Debbie Gipson[1], Jennifer Hawkins[1], Julia Herzog[1], Susan Massengill[5], Richard C. McEachin[1], Subramaniam Pennathur[1], Kalyani Perumal[6], Roger Wiggins[1], Matthias Kretzler[1]

[1]University of Michigan, [2]Renaissance Renal Research Institute, [3]Wayne State University, [4]Temple University, [5]Levine Children's Hospital, [6]Univiversity of Illinois at Chicago

Recent advances have allowed the development of molecular maps to define chronic kidney disease (CKD) in new, accurate and personalized ways. These developments make possible the prediction of outcomes and response to therapy and the identification of key molecular targets for treatment of CKD in individual patients. Identification of such targets entails close collaboration between teams of investigators to collect and annotate samples from well characterized CKD subjects. In addition, technologies are needed that support information exchange, robust databanks, and data integration to define key pathways driving CKD pathogenesis. The O'Brien Kidney Translational Core Center at the University of Michigan provides such biobanking, databank structure and bioinformatic support to basic and clinical investigators to allow them to pursue critical precision medicine investigations of humans with CKD. The Clinical Phenotyping and Biobank Core has enrolled over 1200 patients with CKD from 5 sites and banked their samples and clinical information providing a valuable resource for efficient discovery. Multiple specific research studies have now successfully utilized these resources. The Applied Systems Biology Core and its online analytical tool, Nephroseq, have assisted hundreds of investigators around the world in approaches to the analysis of large transcriptomic datasets and other systems-level, biological studies of patients with CKD. The Center's Bioinformatics Core provides access to computational applications and skilled professional support in bioinformatics and biostatistics and will now be providing back-end maintenance of Nephroseq. The Administrative Core directs pilot and small grants, student training and discount programs with the goal of helping new and established researchers utilize systems biological and translational research tools. Together these cores provide a comprehensive translational research support for novel research into classification and treatment of chronic kidney diseases. All interested academic investigators around the world are invited to make use of these services and to contact us for information and consultation.

MINING DIRECTIONAL DRUG INTERACTION EFFECTS ON MYOPATHY USING THE FAERS DATABASE

Danai Chasioti[1], Xiaohui Yao[1], Pengyue Zhang[2], Xia Ning[3], Lang Li[2], Li Shen[4]

[1]IUPUI School of Informatics and Computing; [2]Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine; [3]IUPUI Department of Computer Science; [4]Center for Neruoimaging, Department of Radiology and Imaging Sciences, Indiana University School of Medicine

Background: Mining high-order drug-drug interaction (DDI) induced adverse drug effects (ADEs) from electronic health record (EHR) databases is an emerging area, and very few studies have explored the relationships between DDIs. To bridge this gap, we study a novel pharmacovigilance problem for mining directional drug interaction effect on myopathy using the FDA Adverse Event Reporting System (FAERS) database. Method: The analysis was performed on a case–control dataset extracted from the FAERS database. The dataset contains 1,763 drugs, and includes 136,860 myopathy events and 3,940,587 control events. Given two sets of drug combinations D1 and D2 (a superset of D1), we define the directional ADE effect from D1 to D2, as the altered ADE risk associated with the change from taking D1 to taking D2. The ADE risks were estimated using odd ratios (ORs). To address both computational and statistical challenges, this study was focused on computing ORs for frequent D2's (i.e., the number of occurrences a user-specified minimum support). The Apriori algorithm was employed to identify frequent D2's. Results: Using the minimum support of 1000, we identified 764 frequent drugs, 7036 frequent 2 drug combinations, and 4280 frequent 3 drug combinations. The top ten ADE ORs for single drugs range from 4.1 to 5.6, for two drug combinations from 12.6 to 21.5, and for three drug combinations from 14.8 to 19.5. The top ten directional ADE ORs between one drug and two drugs range from 13.5 to 28.2; those between one drug and three drugs range from 13.1 to 20.3; and those between two drugs and three drugs range from 11.3 to 34.4. Multiple promising directional ADE findings were identified. For example, the risk of myopathy is 28.2 times higher when adding Gadopentetate dimeglumine on top of Gadobenate dimeglumine. Both drugs are Gadolinium-based contrast agents (GBCAs) used in magnetic resonance imaging. GBCAs have been shown to be associated with Nephrogenic systemic fibrosis (NSF) which may present as progressive myopathy. Conclusion: The directional drug interactions capture the ADE risks introduced by additional drugs taken on top of a set of baseline drugs, and provide novel and valuable pharmacovigilance knowledge with potential to impact clinical decision support. Mining frequent patterns using Apriori is a promising approach for effective discovery of high-order directional drug interaction effects.

DECIPHERING NEURONAL BROAD HISTONE H3K4ME3 DOMAINS ASSOCIATED WITH GENE-REGULATORY NETWORKS AND CONSERVED EPIGENOMIC LANDSCAPES IN THE HUMAN BRAIN

Aslihan Dincer[1], Eric E. Schadt[2], Bin Zhang[2], Joel T. Dudley[2], Davin Gavin[3], Schahram Akbarian[4]

[1]Department of Neuroscience, Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York; [2]Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York; [3]Department of Psychiatry, Jesse Brown Veterans Affairs Medical Center, Chicago; [4]Department of Psychiatry, Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York

Only few histone modifications have been mapped in human brain. Trimethylation of histone H3 at lysine 4 (H3K4me3) is a chromatin modification known to mark the transcription start sites (TSS) of active gene promoters. Regulators of H3K4me3 mark are significantly associated with the genetic risk architecture of common neurodevelopmental disease, including schizophrenia and autism. Here, through integrative computational analysis of epigenomic and transcriptomic data based on next generation sequencing, we investigated H3K4me3 landscapes of FACS sorted neuronal and non-neuronal nuclei in human postmortem, non-human primate (chimpanzee and macaque) and mouse prefrontal cortex (PFC), and blood. We characterized the broad H3K4me3 histone domains from human PFC in the context of cell-type specific regulation, association with neuronal and non-neuronal gene expression and potential implications for normal and diseased development. We first addressed the occurrence and the biological significance of the broad H3K4me3 histone domains in three different cell types, including NeuN+ PFC neurons, NeuN- PFC cells, and nucleated blood cells and then identified novel regulators of these three different cell types by focusing on top 5% broadest H3K4me3 peaks (length in base pairs). In PFC neurons, broadest peaks ranged in size from 3.9 to 12kb, with extremely broad peaks (~10kb or broader) related to synaptic function and GABAergic signaling (DLX1, ELFN1, GAD1, LINC00966). Broadest neuronal peaks showed distinct motif signatures, and were centrally positioned in prefrontal gene bayesian regulatory networks. Approximately 120 of the broadest H3K4me3 peaks in human PFC neurons, including many genes related to glutamatergic and dopaminergic signaling, were fully conserved in chimpanzee, macaque and mouse cortical neurons. Exploration of spread and breadth of lysine methylation markings in specific cell types could provide novel insights into epigenetic mechanism of normal and diseased brain development, aging and evolution of neuronal genomes.

NORMALIZATION TECHNIQUES AND MACHINE LEARNING CLASSIFICATION FOR ASSIGNING MOLECULAR SUBSETS IN AUTOIMMUNE DISEASE AND CANCER

Jennifer M. Franks[1,2], Guoshuai Cai[1], Jaclyn N. Taroni[3,4], Michael L. Whitfield[1,2]

[1]Department of Molecular and Systems Biology; [2]Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth; [3]Department of Systems Pharmacology and Translational Therapeutics; [4]Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine

Systemic sclerosis (SSc) is a complex connective tissue disease involving skin and internal organ fibrosis, vascular damage, and immunologic abnormalities. To characterize disease heterogeneity and molecular pathogenesis, transcriptomics have elucidated common biological processes in subsets of SSc patients using intrinsic gene expression analyses. Four intrinsic subsets characterized by distinct molecular signatures have been validated by multiple independent cohorts. Technical biases inherent to different gene expression profiling platforms present a unique problem when analyzing data generated from multiple studies. While microarray and RNA-seq data have been shown to have a high correlation, differences in overall processing and quantification result in distinct data distributions. Here, we introduce an accurate and reproducible classifier for SSc molecular subtypes and have developed a method to normalize data when platform-specific artifacts arise. We used three independent, well-characterized and validated experimental microarray data sets (Hinchcliff et al., 2013; Milano et al., 2008; Pendergrass et al., 2012) to train a supervised classifier using three-fold cross-validation repeated ten times, performing at an average of >88% accuracy. Data from other platforms, including RNA-seq, are analyzed for platform-based bias using guided PCA analysis (Reese et al., 2013). We developed a method to eliminate platform bias by normalizing on a gene-by-gene basis using the microarray training data as the target distribution. We find that this method successfully removes platform-specific effects from the data. Following normalization, each sample is assigned to a molecular subset based on support vector machine (SVM) classification. Our preliminary analyses find that these methods work extremely well on a validation RNA-seq dataset in SSc (100% accuracy, n=12, Li et al., in preparation). We also applied our methods to breast cancer DNA microarray and RNA-seq data from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas, 2012) where five intrinsic gene expression subsets have been previously identified and described with PAM50 (Parker et al., 2009). Tumor and tumor-adjacent normal biopsies of breast cancer, for which intrinsic subtype information was available, were used to train and test a SVM and evaluate our normalization technique. We achieve 93% accuracy in assigning subtypes for normalized RNA-seq data using our classifier trained exclusively on microarray data. Until recently, clinical trials and diagnosing physicians have not considered molecular heterogeneity in the context of immunosuppressive therapy, which may explain improvement in select SSc patients (Martyanov & Whitfield, 2016). Advancing personalized medicine by using intrinsic molecular subsets may prove particularly beneficial to this field. With our newly developed techniques, we can successfully leverage information from validated expression data in new analyses despite different platforms used for gene expression profiling.

MULTI-OMICS DATA INTEGRATION TO STRATIFY POPULATION IN HEPATOCELLULAR CARCINOMA

Kumardeep Chaudhary, Olivier Poirion, Liangqun Lu, Lana Garmire

University of Hawaii Cancer Center, Honolulu

High mortality rate of Hepatocellular Carcinoma (HCC) is in part due to the vast heterogeneity of the cancer. Identifying robust molecular subgroups of HCC helps to guide precise targeted therapeutics. This could be realized by integrating different layers of omics datasets from the same cohort. To achieve this, we present a deep learning (DL) based method to inspect the different subpopulations of patients within HCC from TCGA. We obtained the information of 360 HCC patients available in TCGA with 3 omics data types (RNA-seq, miRNA-seq and methylation). To identify the different subpopulations, our pipeline implements a DL-based autoencoder, identifies hidden layers linked to survival, and performs k-means clustering using these new features. To assign new samples to the identified subpopulations, a supervised classification procedure was conducted using Support Vector Machine (SVM). To assess the performance of the model, we used 5-folds cross-validation scheme to estimate c-index and brier scores. We also used 60:40 ratio to split the data in 10 folds in order to assess the significance of the coxph regression in the test dataset. Finally, we inferred the cluster labels of two external cohorts based on the gene expression data. Autoencoder framework was used to combine the 3 omics as input features (~40,000) and to produce 100 transformed new features. Among these new features, we identified 36 features significantly linked with survival, which were further used to infer 2 optimal clusters of patients with significant survival differences. Using cross-validation procedure, we obtained average c-index and brier score values of 0.70 and 0.20 respectively, for the test sets. Also, the coxph regression shows significant survival estimation when using the test samples. Finally, our framework is validated on two external dataset: 221 HCC samples from GEO study and 230 HCC samples from LIRI-JP (RIKEN) cohort. Moreover, we proved that each of the individual omic feature sets can be used successfully to infer the 2 survival profiles. However, the combination of the 3 omics is more powerful. We also compared the DL methodology with new features produced by PCA instead. The clinical and molecular differences (in terms of survival, pathways, and driver mutation profiles) were significantly different for the two subpopulations. This is the first study to employ deep learning as a robust framework to identify non-linear combination of multi-omics features linked to identification of subclasses of HCC patients. Using multi-omics datasets, our pipeline successfully combines these different features and identifies two HCC subpopulations exhibiting different survival profiles. We then used this model in combination with supervised machine-learning approaches to predict HCC subpopulation assignment for test and validation datasets.

TOWARDS STANDARDS-BASED CLINICAL DATA WEB APPLICATION LEVERAGING SHINY R AND HL7 FHIR

Na Hong, Naresh Prodduturi, Chen Wang, Guoqian Jiang

Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Introduction: The Fast Healthcare Interoperability Resources (FHIR) is an emerging clinical data standard developed at HL7, which enables the representation and exchange of the electronic health records (EHR) data in a standard structure. FHIR has strong executable ability based on the RESTful service architecture and multiple flexible data exchange formats. Shiny is a web application framework with a simplified web deployment mechanism that enables powerful R functions to support the graphical and interactive analysis. Therefore, with the goal of building reusable and extensible clinical statistics and analysis applications, we aim to design, develop and evaluate a flexible framework using the HL7 FHIR standard and the R-powered web application - Shiny. Methods: We first established a local FHIR server to manage our clinical data. This part of work is focused on the analysis and implementation of the FHIR data models (i.e., core resources), data exchange formats (e.g., XML and JSON) and invoking an open source HAPI FHIR API. Second, we designed two analysis workflows that are focused on patient-centered data analysis and cohort-based data analysis respectively. According to the workflow design, we developed an open application platform known as Shiny FHIR using the Shiny web framework and the established FHIR server. Results: We built a local FHIR server using the HAPI DSTU2 API. In total, 140 patient records, 476 observation records, 496 condition records and 107 procedure records were populated into the FHIR server for testing. With the support of R packages, including 'jsonlite', 'dygraph' and 'timeline', our platform can be used for a variety of use cases of clinical data analysis, including patient blood pressure observation timeline analysis, patient cohort gender/age distribution statistics, etc. The results of the experiment show that the Shiny FHIR integration approach offers the feasibility of web-based interactive statistics analysis on standardized FHIR-based clinical data. Discussions: The implementations of FHIR have already attracted a lot of interests from healthcare practitioners. Our Shiny FHIR implementation provides a useful framework that would be complementary to other FHIR-based applications (e.g., SMART on FHIR). Shiny FHIR is designed to visualize the FHIR-conformant data through capturing the user experiences and habits, and offers rapid support for clinical research while combining the limitless statistical power of R. However, there are several issues need to be solved in the future, such as the support of the FHIR extensions and custom models and the system performance enhancement. In this study, we described our efforts in building a standardized clinical statistics and analysis application leveraging Shiny. We consider that the designed workflows can be applied to other EHRs data that follows the FHIR standard, and other public available FHIR servers can be used to validate the utility of our framework.

A DATA LAKE PLATFORM OF CONTEXTUAL BIOLOGICAL INFORMATION FOR AGILE TRANSLATIONAL RESEARCH

Austin Huang[1], Dmitri Bichko[1], Mathieu Boespflug[2], Edsko deVries[3], Facundo Dominguez[2], Daniel Ziemek[1]

[1]Pfizer, [2]Tweag I/O, [3]Well-Typed

Researchers need to aggregate contextual biological information in order to interpret experimental and clinical study results. These needs vary greatly depending on the scientific question. Creating large-scale, structured data repositories requires substantial investment that is not amenable to the rapidly-evolving needs of translational research. On the other hand, performing data analyses using adhoc collections of local data files (excel sheets, csv tables, etc.) allows rapid and flexible execution, it also creates technical debt. In the long term, these workflows result in missed opportunities to accumulate institutional knowledge and are associated with poor reproducibility. We have implemented a data platform that can achieve the benefits of a more principled handling of data persistence with minimal analyst overhead. This is achieved by automating schema inference, metadata curation, versioning, and RESTful service production through a simple, Git-like ingestion tool. Data scientists can retrieve data via familiar client language APIs such as dplyr in R. The platform is built on open source database (Postgres, with an architecture that allows alternative backends) and functional programming (Haskell, PostgREST) technologies. Our objective is to accelerate data sharing/discoverability on analyst teams and drastically reduce the effort of persisting data in a systematic mechanism. We therefore provide a technology foundation for rapid data service production and improving reproducibility and reusability in data analyses.

GENOME READ IN-MEMORY (GRIM) FILTER: FAST LOCATION FILTERING IN DNA READ MAPPING USING EMERGING MEMORY TECHNOLOGIES

Jeremie Kim[1], Damla Senol[1], Hongyi Xin[2], Donghyuk Lee[1,3], Mohammed Alser[4], Hasan Hassan[5], Oguz Ergin[5], Can Alkan[4], Onur Mutlu[1,6]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA; [2]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA; [3]NVIDIA Research, Austin, TX; [4]Department of Computer Engineering Bilkent University, Ankara, Turkey; [5]Department of Computer Engineering, TOBB University of Economics and Technology, Söğütözü, Ankara, Turkey; [6]Department of Computer Science, Systems Group ETH, Zürich, Switzerland

High-throughput sequencing (HTS) technology has resulted in a massive influx of available genetic data. Using HTS technology, genomes are sequenced relatively quickly and result in many short DNA sequences (reads) that are used to analyze the donor's genome across multiple days when using state-of-the-art methods. The first step of genome analysis, *read mapping*, determines origins for billions of reads within a reference genome to identify the donor's genomic variants. Hash-table based read mappers are a common type of comprehensive read mappers. They operate by fetching from a pre-generated hash-table, potential mapping locations of a read in the reference genome, which are verified by local alignment, a computationally-expensive dynamic programming algorithm that determines similarity between the read and the potential mapping segment of the reference genome. Alignment has traditionally been the computational bottleneck of read mapping, but recently, many works have been proposing a new step called Location-Filtering in order to alleviate this bottleneck.

Location-Filtering is a critical step where many incorrect potential locations from the hash-table are discarded before local alignment verifies such locations. FastHASH, SHD, and GateKeeper propose variations of Location-Filtering that discard only incorrect locations to reduce end-to-end runtime of hash-table based read mapping. Location-Filtering is now the computational bottleneck of read mapping.

Our goal is to create an efficient Location-Filter that quickly discards as many false negative locations as possible before alignment, while retaining a zero false positive rate. Efficiently filtering incorrect mappings before alignment significantly improves throughput and latency of hash-table based read mapping. We propose a novel filtering algorithm that quickly eliminates from consideration reference genome segments where alignment would yield no matches. Our algorithm's novelty mainly stems from its design to exploit 3D-stacked memory systems. 3D-stacked memory is an emerging technology that tightly integrates computation and high-capacity memory in a single die stack, thereby enabling concurrent processing of large data chunks at low latency and high bandwidth. The key ideas of our design consist of 1) a new representation of coarse-grained reference genome segments such that the genome can be operated on in parallel using bitwise operations and 2) exploiting the parallel computation capability of 3D-stacked memory to run massively-parallel in-memory operations on the new genome representation. We call our resulting filter the GRIM-Filter.

This work shows how GRIM-Filter can be used with any hash-table based read mapping algorithm and how it effectively exploits processing-in-memory capabilities of 3D-stacked memory. We show that when running with 5% error tolerance, GRIM-Filter reduces false positive locations by *5.59x-6.41x* and provides a *1.81x-3.65x* end-to-end speedup over the state-of-the-art read mapper mrFAST with FastHASH

BCL-2 FAMILY MEMBERS AS REGULATORS OF RESPONSIVENESS TO BORTEZOMIB IN A MULTIPLE MYELOMA MODEL

Melissa E. Ko[1,2], Charis Teh[3,4], Christopher S. Playter[5], Eli R. Zunder[6], Daniel H. Gray[4,7], Wendy J. Fantl[8], Sylvia K. Plevritis[9], Garry P. Nolan[2]

[1]Cancer Biology Program, Stanford School of Medicine, Stanford, CA; [2]Baxter Laboratory for Stem Cell Biology, Stanford School of Medicine, Stanford, CA; [3]Molecular Genetics of Cancer Division, Immunology Division, The Walter and Eliza Hall Institute, Parkville, VIC, Australia; [4]Department of Medical Biology, The University of Melbourne, Parkville, VIC, Australia; [5]Department of Biological Sciences, Purdue University, Lafayette, IN; [6]Department of Biomedical Engineering, University of Virginia, Charlottesville, VA; [7]The Walter and Eliza Hall Institute, Parkville, VIC, Australia; [8]Department of Obstetrics and Gynecology, Stanford School of Medicine, Stanford, CA; [9]Department of Radiology, Stanford School of Medicine, Stanford, CA

Survival rates for B cell malignancies have steadily improved over the last five decades reaching levels of over 50% as a result of therapeutic agents such as dexamethasone, bortezomib, and lenalidomide. However, despite their success in producing clinical responses, the cellular mechanisms by which these agents kill tumor cells are poorly understood. We hypothesized that the Bcl-2 family of proteins, which are known to control initiation of apoptosis and are frequently dysregulated in cancerous B cells such as multiple myeloma, can influence responsiveness to these therapeutic agents. Thus, with a focus on multiple myeloma, we aimed to comprehensively profile individual cells for their expression levels of Bcl-2 family members simultaneously with activated intracellular signaling proteins upon exposure of cells to drugs used to treat B-cell malignancies. We applied single-cell mass cytometry to investigate the interplay of pro-survival and pro-apoptotic Bcl-2 family members in MM1S B lymphoblastic cells exposed to different drugs. This dataset was analyzed with FLOW-MAP, a computational tool developed in the Nolan Lab that organizes high-dimensional single-cell data into an interpretable 2D graph structure. FLOW-MAP enabled the apoptotic progression of individual cells to be visualized and showed changes in expression levels of Bcl-2 family members and signaling factors across cells with different drug sensitivities. Our extensive study revealed heterogeneous responses of cell subsets to therapeutic agents used to treat multiple myeloma patients. For example, our results showed that bortezomib, a proteasome inhibitor approved for treatment of multiple myeloma, potently induces apoptosis within 24 hours to a greater extent compared to other treatments. Induction of apoptosis in single cells treated with bortezomib coincided with a selective reduction of a subset of pro-survival Bcl-2 members. Furthermore, our analysis suggests that a metric that reflects the balance of pro-survival and pro-apoptotic Bcl-2 proteins may best separate and predict cells with differential sensitivity to bortezomib. This paradigm is supported by statistical modeling wherein we developed a classifier of bortezomib-resistant vs. sensitive cells using Bcl-2 family information or a single Bcl-2 score with significant accuracy. Our study provides a general framework for understanding differential sensitivity of tumor populations to anti-cancer drugs. Our results are likely to identify previously unknown death-inducing mechanisms as well as pinpoint potential synergies between standard-of-care therapies and newly developed therapies, such as Bcl-2 family inhibitors.

BIOMEDICAL TEXT-MINING APPLICATIONS FOR THE SYSTEM DEEPDIVE

Emily K. Mallory, Chris Re, Russ B. Altman

Stanford University

A complete repository of biomedical relationships is key for understanding the processes underlying both human disease and drug response. After decades of experimental research, the majority of known biomedical relationships exist solely in textual form in the literature and are thus computationally inaccessible. While curated databases have experts manually annotate relevant relationships or interactions from text, these databases struggle to keep up with the rapid growth of the biomedical literature. To address the need for biomedical relationship extraction, there have been numerous biological entity and relationship extraction challenges; however, extraction systems in the biomedical space tend to be task specific and do not provide a general framework for quickly developing future systems to address new extraction tasks. In this work, we developed multiple entity and relationship applications (called "extractors") for the system DeepDive to extract biomedical relationships from full text articles. DeepDive is a trained system for extracting information from a variety of sources, including text. Application developers create features and training examples, and DeepDive assigns a probability that a given entity or relationship is correct or true in the original sentence. We developed entity extractors for genes, drugs, and diseases; and relationship extractors for gene-gene, gene-disease, and gene-drug relationships. We evaluated the gene-gene work previously with a corpus of articles from three PLOS journals, and we are currently evaluating the other two relationship extractors on a corpus from PubMed Central. The precision of our entity extractors ranged from 80 to 90%. For the task of extracting gene-gene relationships, our system achieved 76% precision and 49% recall in extracting direct and indirect interactions previously curated by the Database of Interacting Proteins (DIP). For randomly curated extractions, the system achieved between 62% and 83% precision based on direct or indirect interactions, as well as sentence-level and document-level precision. Our current gene-disease and gene-drug extractors achieved over 70% precision on a random subset of documents from over 340,000 full text articles in the PubMed Central Open Access Subset. We are currently tuning these extractors to increase performance. This work will enable not only full text literature extraction for biomedical relationships, but also computational methods development based on these relationships.

PROFILING ADAPTIVE IMMUNE REPERTOIRES ACROSS MULTIPLE HUMAN TISSUES BY RNA SEQUENCING

Serghei Mangul[1], Igor Mandric[2], Harry Taegyun Yang[1], Dennis Montoya[1], Nicolas Strauli[3], Jeremy Rotman[1], Benjamin Statz[1], Will Van Der Wey[1], Alex Zelikovsky[2], Roberto Spreafico[1], Maura Rossetti[1], Sagiv Shifman[1], Mark Ansel[3], Noah Zaitlen[3], Eleazar Eskin[1]

[1]University of California Los Angeles, [2]Georgia State University, [3]University of California San Francisco

Assay-based approaches provide a detailed view of the adaptive immune system by profiling T-and B-cell receptors. However, these methods come at a high cost and lack the scale of regular RNA sequencing (RNA-seq). We developed ImReP, a novel computational method that utilizes RNA-seq data to profile the adaptive immune repertoire. ImReP is able to quantify individual immune responses from RNA-Seq data based on a recombination landscape of genes encoding B- and T-cell receptors. We applied ImReP to 8,555 samples from 544 individuals and 53 diverse human tissues, and constructed the complementarity determining regions 3 (CDR3), which is the most variable part of the antigen-binding site. We assembled 3.8 million distinct CDR3 sequences. Analyzing this dataset, we identified the normal, healthy, adaptive immune profile for different tissues. We describe the variation in immune profiles, and the distribution of clonal lineages across individuals and tissues. Base on the immune profiles generated by ImReP, we were able to identify inflammation and various diseases, as confirmed from the histological images. The atlas of T and B cell repertoires, freely available at https://sergheimangul.wordpress.com/atlas-of-t-and-b-cell-repertoires/ , is the largest recourse in terms of the number of CDR3 sequences and tissue types involved. We anticipate this recourse to enhance future studies in areas such as immunology and advance development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/.

THE CMH VARIANT WAREHOUSE - A CATALOG OF GENETIC VARIATION IN PATIENTS OF A CHILDREN'S HOSPITAL

Neil MIller[1], Greyson Twist[1], Byunggil Yoo[1], Andrea Gaedigk[2]

[1]Center for Pediatric Genomic Medicine, Children's Mercy, Kansas City; [2]Division of Clinical Pharmacology & Therapeutic Innovation, Children's Mercy, Kansas City, School of Medicine, University of Missouri-Kansas City

Advances in high-throughput DNA sequencing have enabled the comprehensive identification of individual genetic variation on an unprecedented scale, powering the diagnosis of disease and personalized treatment. As the ability to detect genetic variation has grown, clinicians and researchers struggle to interpret the functional significance of the millions of variants found in each individual genome. The Variant Warehouse at the Center for Pediatric Genomic Medicine at Children's Mercy, Kansas City, is a resource containing a record of over 160 million genomic variants detected in more than 5000 patients sequenced by the Center since 2011. Each variant has been characterized by the CPGM's Rapid Understanding of Nucleotide Effect Software (RUNES) pipeline, which records database cross references, predicted functional consequences and a variant classification score (1-5) based on preliminary guidelines from the American College of Medical Genetics and Genomics (ACMG). Additionally, a local allele frequency is calculated for each variant every 6 hours enabling clinicians and researchers to rapidly identify rare variants. Despite extensive cross-referencing with databases such as dbSNP, ClinVar, ExAC and COSMIC the CMH variant warehouse contains a significant number of novel variants not present in external databases. 59% of the total variants in the warehouse are novel with a local allele frequency of less than 0.25%. Of these, 1% are category 1-3 variants expected to have some functional impact. We have observed 82,578 variants among a panel of 58 pharmacogenes (including CPIC genes), of which 59% are novel and 2% are category 1-3 variants. The amount of novelty observed in this patient population suggests that efforts to comprehensively catalog human variation remain a work in progress and that interpretation of variant data will require some level of interpretation of novel variants for the foreseeable future. These observations are increasingly relevant in pharmacogenomics applications where drug compatibility is determined through association to known haplotypes; in this context, the presence of novel and rare variants must be anticipated and accounted for in automated haplotype determination. The CMH variant warehouse is publicly available at http://warehouse.cmh.edu. Tools to search and view variants by gene, category and allele frequency are provided as well as bulk downloads of data. Programmatic access to data is provided through implementations of the Global Alliance for Genomics and Health variant annotation API.

MUTPRED2 AND ITS APPLICATION TO THE INFERENCE OF MOLECULAR SIGNATURES OF DISEASE

Vikas Pejaver[1], Lilia M. Iakoucheva[2], Sean D. Mooney[3], Predrag Radivojac[1]

[1]Department of Computer Science and Informatics, School of Informatics and Computing, Indiana University Bloomington; [2]Department of Psychiatry, University of California San Diego; [3]Department of Biomedical Informatics and Medical Education, University of Washington Seattle

Over the past decade, several methods have been developed for the computational prioritization of missense mutations. However, the identification of the effects of such mutations on protein structure and function still remain a major challenge. Previously, we developed MutPred, a random forest-based model for the classification of pathogenic missense variants and the automated inference of molecular mechanisms of disease. Here, we build on our previous work and present MutPred2 as an improved approach for these tasks. For pathogenicity prediction, MutPred2 particularly benefits from a larger and heterogeneous training set, the inclusion of new features, the encoding of local sequence context and the use of a neural network ensemble. Through cross-validation experiments and a test on an independent data set, we show that MutPred2 outperforms MutPred and other state-of-the-art methods. In particular, we observe that MutPred2 predicts fewer pathogenic mutations than PolyPhen-2, when applied to homozygous mutations from healthy individuals. Additionally, MutPred2 has over 50 built-in structural and functional property predictors, which greatly increase the number of possible downstream consequences that can be associated with a given amino acid substitution. We introduce a novel ranking approach that utilizes a positive-unlabeled learning framework to derive posterior probabilities for the disruption of these properties and, thus, infer the most likely molecular mechanism of pathogenicity. We then demonstrate the utility of MutPred2 in two situations. First, we identify prominent structural and functional signatures in a data set of mostly Mendelian diseases (from MutPred2's training set) and recapitulate known associations between these diseases and ordered and structured regions of proteins. We also make novel predictions about the role of allosteric residues in such diseases. Second, we apply MutPred2 to a data set of de novo mutations from patients diagnosed with neuropsychiatric disorders, along with healthy siblings as controls. On this data set, MutPred2 pathogenicity scores alone are sufficient to distinguish between neuropsychiatric cases and controls, without any additional gene-based or variant-based filtering. We also observe that disruptions in protein-protein interactions (PPIs), phosphorylation and acetylation are frequent mechanisms, suggesting that neuropsychiatric disorders are largely characterized by a breakdown in molecular signaling. Finally, we identify candidate mutations predicted to disrupt PPIs and validate them experimentally.

HIV-TRACE: MONITORING THE HIV EPIDEMIC IN NEAR REAL TIME USING LARGE NATIONAL AND GLOBAL SCALE MOLECULAR EPIDEMIOLOGY

Sergei Pond[1], Steven Weaver[1], Joel Wertheim[2], Andrew J. Leigh Brown[3]

[1]Temple University, [2]University of California San Diego, [3]University of Edinburgh

Many pathogens, including HIV, propagate along sexual and social contact networks. It is now clear that HIV transmission networks belong to the scale free family and the spread of infections in scale free networks is critically enhanced by highly connected individuals or "hubs". The structure of the transmission network has major implications for interrupting an epidemic. Since pathogen transmission networks are not observed directly, they are inferred and characterized based on indirect measurements, and methods to do this properly remains an open research challenge. Because of their rapid and host-specific evolution and chronic disease states, HIV sequence isolates are essentially unique to each infected person. This sequence uniqueness can be used to confirm or reject the hypothesis that two individuals are "linked" by a recent transmission or belong to the same transmission cluster There are ~ 1,000,000 HIV sequences isolated from different individuals over the last 4 decades. National and international surveillance and drug resistance programs are generating high resolution sequencing data on hundreds of thousands of isolates annually. We developed HIV Transmission Cluster Engine (HIV-TRACE) in order to make the process of cluster (and network) inference automated, fast, convenient, and more robust. It is an efficient open-source application designed to scale well and enable near real-time inference and analysis of large networks: it can process 100,000 sequences in ~15-30 minutes on a 64 core backend system. HIV-TRACE (hiv-trace.org) is an open-source web application built on robust and popular modern libraries. User interaction and result visualization is done entirely in the browser, processing is done asynchronously on a server backend. Components and versions of HIV-TRACE are used by the CDC (VARS, HICSB), Canadian public health officials, NYC Department of Public health, San Diego primary infection cohort, and the UK Drug Resistance Database. We illustrate the utility of HIV-TRACE on four real-world examples of essential questions in public health and epidemiology of HIV-1: 1). Are there rapidly growing transmission clusters, and what is driving their growth? 2). How does HIV spread at different geographic scales, and among different risk groups? 3). How can treatment and intervention be deployed in optimal ways to reduce incidence and prevalence? 4). Can vaccine and prevention efficacy be measured more accurately using network-level information.

THE EXTREME MEMORY® CHALLENGE: A SEARCH FOR THE HERITABLE FOUNDATIONS OF EXCEPTIONAL MEMORY

Mary A. Pyc, Emily Giron, Philip Cheung, Douglas Fenger, J. Steven de Belle, Tim Tully

Dart NeuroScience

We are interested in discovering new candidate targets for drug therapies to enhance cognitive vitality in humans throughout life, and to remediate memory deficits associated with brain injury and brain-related diseases such as Alzheimer's and Parkinson's. To achieve our goal we need a comprehensive and objective understanding of the human genome contribution to variation in memory performance in healthy individuals. We are implementing a Genome-Wide Association Study (GWAS) to identify genetic loci varying among individuals who possess exceptional and normal memory abilities. These genes and those in associated networks will inform drug discovery and development. Our first step is to identify exceptional members of the population. Thus, we have created an online memory test – the Extreme Memory Challenge (XMC, accessible at http://www.extrememorychallenge.com) – to conveniently screen through an unlimited number of subjects to find individuals with exceptional memory consolidation abilities. Identified subjects are (1) validated by a battery of secondary memory tasks, and (2) providing saliva samples from which we can isolate DNA for GWAS. . Ten pilot experiments were conducted to parameterize the XMC screen. Participants learned face-name pairs for a delayed recall test. After initial study, each name was presented and participants were asked to select the correct face among four (distracters were other faces paired with different names). One day later participants completed a final test trial. We are primarily interested in forgetting across sessions, as this provides an estimate of consolidation across a 24-hour time interval. Pilot studies indicated the optimal protocol should include 30 face-name pairs, presented at a 4 second rate. To date, 17,849 participants from 176 nations have been screened in the XMC. Of these, 11,311 have completed both sessions. Individuals in our sample are most frequently Caucasians (55%), post-secondary school-educated (63%), reported being most alert in the morning (51%), and right handed (89.5%). The average age was 34, and the gender distribution was split evenly. The forgetting rate (decrease in performance from day 1 to day 2) was 10%. We have identified 49 individuals with perfect performance on day 2 of the test and 24 with exceptional consolidation abilities (defined as 3 SDs from the mean). We have begun the genomics phase of the study with 33 individuals who have completed additional behavioral testing.

RESCUE THE MISSING VARIANTS-LESSONS LEARNED FROM LARGE SEQUENCING PROJECTS

Yingxue Ren[1], Joseph S. Reddy[1], Vivekananda Sarangi[2], Jason P. Sinnwell[2], Steve G. Younkin[3], Nilüfer Ertekin-Taner[3], Owen A. Ross[3], Rosa Rademakers[3], Shannon K. McDonnell[2], Joanna M. Biernacka[2], Yan W. Asmann[1]

[1]Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL; [2]Department of Health Sciences Research, Mayo Clinic, Rochester, MN; [3]Department of Neuroscience, Mayo Clinic, Jacksonville, FL

Identifying novel disease variants through next generation sequencing (NGS) has been a fruitful practice in medical research in recent years, leading to the discoveries of new disease mechanisms as well as therapeutic strategies. The GATK best practices have since been established to provide general recommendations on core processing steps required to go from raw reads to final variant call sets. However, with the sample size drastically increasing in today's sequencing experiments, many default variant calling strategies and the choice of tools call for a closer examination. Our study utilized the whole exome sequencing data provided by the Alzheimer's Disease Sequencing Project (ADSP) to test for different variant calling strategies and tools involved in the variant discovery workflow in the context of sample sizes. We first investigated the impact of using different sequence aligners on variant callsets while keeping the default GATK settings of the variant calling and QC steps identical. We selected 1952 samples to align by both BWA and NovoAlign, and compared the variant callsets in 50, 100, 200, 500, 1000 and 1952 samples. We discovered that the percentage of variants unique to aligner increased dramatically with increasing sample sizes. At sample size of 1952, the unique variants generated by BWA and NovoAlign account for more than 20% of total called variants. These unique variants have good variant quality metrics: ~80% have Genotype Quality (GQ) score of 60 or above, and their distribution of B allele concentration (BAC) centers around 0.5 and 1, consistent with what is expected of diploid genomes. What's more, over 96% of the unique variants have population B allele frequency (BAF) of less than 0.01, indicating that these variants are rare in the population. All these metrics suggest that these unique variants are important to be included in downstream variant analysis. In addition to aligner comparison, we also evaluated single-sample variant calling versus the default, single sample variant calling followed by joint multi-sample genotyping strategy in 50, 100, 500, 2000, and 5000 samples. Our data showed that, with increasing sample sizes, the single-sample calling strategy added increasing percentage of unique variants. At sample size of 5000, single-sample calling added 58,884 variants, accounting for 5.55% of total variants called by both strategies. 7331 of these unique variants passed Variant Quality Score Recalibration (VQSR) and have GQ of 60 or above in at least 5 samples. Our study identified a large number of good-quality variants from the ADSP exome sequencing project that were missed by using one aligner or using multi-sample genotyping strategy alone. Our findings revealed the relationships between bioinformatics pipelines and biomedical research results, and suggested that alternative variant calling strategies may be beneficial for optimal variant discovery in face of today's large sequencing scale.

TOWARD EFFECTIVE MICRORNA QUANTIFICATION FROM SMALL RNA-SEQ

Pamela Russell[1], Richard Radcliffe[2], Brian Vestal[1], Wen Shi[1], Pratyaydipta Rudra[1], Laura Saba[2], Katerina Kechris[1]

[1]Department of Biostatistics and Informatics, Colorado School of Public Health;
[2]Department of Pharmaceutical Sciences, University of Colorado Skaggs School of Pharmacy and Pharmaceutical Sciences

Extensive work has led to robust quantification methods for RNA-seq data primarily derived from large RNAs. Many studies have used these methods "out of the box" to estimate microRNA (miRNA) expression from small RNA-seq data. However, these methods do not effectively address issues particular to miRNAs. First of all, reference bias is amplified due to the small size of sequencing reads derived from miRNAs (~22nt). That is, with shorter reads, a true mismatch between a sample and the reference can lead to incorrect alignments or inability to align reads at all, creating a count bias toward those samples with the reference allele. With longer reads, single mismatches have less impact on alignment algorithms. Second, any bias for individual miRNAs is more impactful overall due to the relatively small repertoire of miRNAs compared to mRNAs. Inaccurate counts for a handful of miRNAs can significantly alter overall library counts and thus affect normalization. We refer to this issue as repertoire bias. Also, most miRNA studies seek to identify functional mature miRNA molecules regardless of the position in the genome that they are originally transcribed from or small non-functional differences between miRNAs of the same family. Tools designed for large RNAs do not address the repetitive nature and family structure of miRNAs, by default returning estimated counts for multiple targets that should be considered equivalent by typical miRNA study paradigms. Genome-based methods often map miRNA reads to multiple loci encoding the same mature miRNA. Methods based on mapping directly to a miRNA database do not suffer from multiple alignments due to identical regions of the genome but do typically distinguish among members of each miRNA family. Both sources of multiple mappings can lead to misleading counts when the goal is to elucidate function. Here we explore all these issues in the context of commonly used methods. We then propose a new high throughput approach that (1) incorporates individual genetic variation into the reference sequence used for alignment, reducing reference bias, and (2) assigns each read to a single functional group such as a miRNA family. We demonstrate the accuracy of this approach compared to other popular methods using a dataset derived from 206 mouse brain samples. Funded by NIH/NIAAA AA016597, R01AA021131 and R24AA013162

NANOPORE SEQUENCING TECHNOLOGY AND TOOLS: COMPUTATIONAL ANALYSIS OF THE CURRENT STATE, BOTTLENECKS AND FUTURE DIRECTIONS

Damla Senol[1], Jeremie Kim[1], Saugata Ghose[1], Can Alkan[2], Onur Mutlu[1,3]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA; [2]Department of Computer Engineering, Bilkent University, Bilkent, Ankara, Turkey; [3]Department of Computer Science, Systems Group, ETH Zürich, Switzerland

Nanopore sequencing, a promising single-molecule DNA sequencing technology, exhibits many attractive qualities and, in time, could potentially surpass current sequencing technologies. Nanopore sequencing promises higher throughput, lower cost, and increased read length, and it does not require a prior amplification step. Nanopore sequencers rely solely on the electrochemical structure of the different nucleotides for identification and measure the change in the ionic current as long strands of DNA (ssDNA) pass through the nano-scale protein pores.

Biological nanopores for DNA sequencing was first proposed in the 1990s, but it was only just recently made commercially available in May 2014 by Oxford Nanopore Technologies (ONT). The first commercial nanopore sequencing device, MinION, is an inexpensive, pocket-sized, portable, high-throughput sequencing apparatus that produces real-time data. These properties enable new potential applications of genome sequencing, such as rapid surveillance of Ebola, Zika or other epidemics, near-patient testing, and other applications that require real-time data analysis. In addition, this technology is capable of generating very long reads (~50,000bp) with minimal sample preparation. Despite all these advantageous characteristics, it has one major drawback: high error rates. In order to provide higher accuracy and higher speed, in May 2016, ONT released a new version of MinION with a new nanopore chemistry called R9, which replaced the previous version R7. Although R9 chemistry improves the data accuracy, the tools used for nanopore sequence analysis are of critical importance as they should overcome the high error rates of the technology.

Our goal in this work is to comprehensively analyze tools for nanopore sequence analysis, with a focus on understanding the advantages, disadvantages, and bottlenecks of the various tools. To this end, we rigorously examine multiple steps in the nanopore genome analysis pipeline. The first step, *basecalling*, translates the raw signal output of MinION into nucleotides to generate DNA sequences. Currently, *Nanocall* and *Nanonet* are publicly available nanopore basecallers. The second step performs genome assembly with assemblers for noisy long reads. Using only the basecalled DNA reads, assemblers generate longer contiguous fragments called *draft assemblies*. Currently, *Canu* and *Miniasm* are the commonly used long-read assemblers. After this step, an improved consensus sequence is generated from the draft assembly with *Nanopolish*, and a complete whole genome is obtained.

We analyze the five aforementioned nanopore sequencing tools in terms of their speed and accuracy, with the goals of determining their bottlenecks and finding improvements to these tools. We also discuss potential future works in nanopore basecallers and assemblers, to take better advantage of nanopore sequencing and to overcome its current disadvantage of high error rates.

DETECTING OUTLIERS FROM MULTIDIMENSIONAL DATA WITH APPLICATION IN CANCER

Kyle Smith[1], Subhajyoti De[2], Debashis Gosh[1]

[1]University of Colorado, [2]Rutgers University

Outliers, which are very different from the typical cases in a cohort, bring in unexpected challenges for decision making in many different disciplines. The issue is more acute in oncology, since most types of cancer are highly heterogeneous diseases. Even within any cancer subtype, patients show extensive variation in their molecular profiles and clinical outcomes. Even within a cohort of cancer patients who have apparently the same biomarkers and received identical treatment, there are exceptional responders and exceptional non-responders, who are outliers. It is suspected that their atypical molecular and clinical profiles contribute to their exceptional response. While identifying such outlier cases can benefit precision medicine initiatives, methods to detect them from multidimensional data has received limited attention. Here, we propose a novel framework to identify outlier cancer patients with atypical profiles from multidimensional genomic data. We argue that detection of outlier patients with atypical profiles can help identify exceptional responders and tailor precision medicine in oncology initiatives.

HUEMR: INTUITIVE MINING OF ELECTRONIC MEDICAL RECORDS

Abiodun Otolorin[1], Nana Osafo[2], William Southerland[2]

[1]Department of Community and Family Medicine, Howard University, Washington, DC;
[2]Department of Biochemistry & Molecular Biology and the Center for Computational Biology and Bioinformatics, Howard University, Washington, DC

Despite the widespread adoption of electronic medical record systems and advances in genomics, a major barrier to research endeavors is the lack of intuitive user-friendly interactive tools that enable researchers to access and analyze data readily. In light of this, innovative tools have been developed to address the problem. However, we hypothesized that an interactive data visualization tool that is capable of stand-alone or plugin functionality that also leverages common data query methodologies would contribute to research efforts requiring interrogation of clinical research databases. Howard University Hospital (HUH) is a tertiary academic medical center with over 50,000 emergency department visits and 8,000 inpatient admissions per year and primarily provides care to the minority population in the District of Columbia metropolitan area. Using de-identified HUH electronic medical records data, a HUH clinical research database was developed. Additionally, the Howard University electronic Medical Records (HUeMR) query tool was developed as a web-based client-server application using javascript and php. HUeMR may function in stand-alone or plugin mode. Its graphical interface was built using Google Charts, an interactive open source visualization library. HUeMR supports complex boolean search operations specified by an interactive query tool. Ontology is presented using linked drop down menus and query construction is displayed in natural language form. Data is displayed using editable interactive charts. Multiple rows of charts may be created that contain different types of data concepts. Queries may be refined by clicking on the charts followed by selection of one or more additional query parameters. Diagnosis based on ICD codes or keywords may also be searched. These features are illustrated in a diabetes use-case investigation. In summary, HUeMR is a secure data analytics that can be use in stand-alone or plugin mode to querying clinical research databases. It has a highly interactive user interface that allows rapid data analysis for cohort discovery.
This work was supported by grant #5G12MD007597 from the National Institute on Minority Health and Health Disparities from the NIH.

DECIPHERING LUNG ADENOCARCINOMA MORPHOLOGY AND PROGNOSIS BY INTEGRATING OMICS AND HISTOPATHOLOGY

Kun-Hsing Yu[1], Gerald J. Berry[2], Daniel L. Rubin[1], Christopher Ré[3], Russ B. Altman[1], Michael Snyder[4]

[1]Biomedical Informatics Program, Stanford University; [2]Department of Pathology, Stanford University; [3]Department of Computer Science, Stanford University; [4]Department of Genetics, Stanford University

Adenocarcinoma accounts for more than 40% of lung malignancy, and microscopic pathology evaluation is indispensable to its diagnosis. However, how histopathology findings relate to molecular abnormalities remains largely unknown. To address this problem, we obtained hematoxylin and eosin stained whole-slide histopathology images, pathology reports, RNA-sequencing, and proteomics data of 538 lung adenocarcinoma patients from The Cancer Genome Atlas. We profiled gene expression, protein expression and modifications, and extracted more than 9,000 objective features from the histopathology images of each patient. We successfully predicted histology grade with transcriptomics and proteomics signatures (area under curve > 0.75) and identified the associated molecular pathways, such as cell cycle regulation, which provide biological insights into tumor cell differentiation grades. We further built an integrative histopathology-transcriptomics model to generate superior prognostic predictions for stage I patients ($P<0.01$) compared with gene expression or histopathology analysis alone. These results suggest that the integration of histopathology and omics studies can reveal the molecular mechanisms of pathology findings and enhance clinical prognostic prediction, which will contribute to the development of precision cancer medicine. Our methods are generalizable to other types of malignancy or diseases.

EXPLORING DEEP LEARNING FOR COPY NUMBER VARIATION DETECTION WITH NGS DATA

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

Institute of Medical Science, University of Tokyo

Copy number variations (CNVs) are an important type of genetic variations widely used for profiling cancer and other complex diseases. Accurate detection and summarization of CNVs help identify oncotarget and cancer subtypes for precision medicine. In using NGS data for CNVs detection, various heterogeneous biases, such as GC-content bias and other noises are needed to be properly processed. This becomes especially important for CNVs detection on single-cell NGS data. In this study, we extend traditional HMM approaches for CNVs detection with deep learning. We extract feature representation, which integrate the information from read count and observable genomic sequences, as the new observable sequence of genomic bins and iteratively train a DNN-HMM model for CNVs detection. We compare our method with other HMM based CNVs detection methods.

IMAGING GENOMICS

POSTER PRESENTATIONS

PERIPHERAL EPIGENETIC ASSOCIATIONS WITH BRAIN GRAY MATTER IN SCHIZOPHRENIA

Dongdong Lin[1], Vince D. Calhoun[2], Juan R. Bustillo[3], Nora Perrone-Bizzozero[4], Jingyu Liu[1]

[1]The Mind Research Network and Lovelace Biomedical and Environmental Research Institute, Albuquerque; [2]Dept. of Electronic and Computer Engineering, University of New Mexico, Albuquerque; [3]Dept. of Psychiatry, University of New Mexico, Albuquerque; [4]Dept. of Neurosciences, University of New Mexico, Albuquerque

Epigenetic regulation by DNA methylation and histone modification has been increasingly recognized for its relevance to schizophrenia (SZ). Beyond the genetic variation, epigenetics through regulation of gene transcription and expression can potentially explain the 'missing' heritability and mediate the effect of genetic risks in disease. Specific to DNA methylation, recent studies have demonstrated that 6-7% of CpG sites across the genome show significant correspondence between brain and blood, supporting the investigation of easily accessible tissues for brain and mental disorders. In this study, we analyzed DNA methylation of 163 CpG sites from saliva and whole brain gray matter density of 108 SZ patients and 105 healthy controls. We are aware of cellularity differences between blood and saliva, and to our best knowledge no detailed saliva-brain correspondence study has been done except general comparison of overall patterns, which indicate saliva may be a more close indicator to brain than blood. The 163 CpG sites are located within the 108 schizophrenic risk regions reported by the Psychiatric Genomics Consortium schizophrenia working group, and also showed strong cross-tissue similarity based on the genome-wide methylation study of blood and brain tissues by Hannon, et al. Quality control and normalization for methylation data were implemented using minfi R package to remove batch effect, and cell type proportion effect. Gray matter density maps were segmented by SPM12 with a smooth kennel of 8 mm3. We applied independent component analysis to both brain imaging data and methylation data, and extracted 25 gray matter networks, and 15 methylation components. Among them, two methylation components were significantly correlated to three gray matter networks (false discovery rate <0.05). The first methylation component comprised two CpG sites within and near gene ZSCAN12, and was associated with a bilateral middle/superior temporal network (r=0.25), and a bilateral superior frontal network (r=-0.24). The higher the methylation component is, the lower the gray matter density in superior frontal gyrus and the higher in middle temporal gyrus are. Moreover, SZ patients showed significant gray matter reduction in superior frontal gyrus (p=7.9x10-5). The second methylation component consisted of CpG sites from two chromosome regions (Chr.10 AS3MT and NT5C2 genes, and Chr. 12 ARL6IP4 and OGFOD2 genes), and was associated with caudate and thalamus regions. All analyses were controlled for age and gender. Although we did not find SZ specific methylation differences within SZ risk regions, our results suggest that DNA methylation patterns in saliva are associated with brain gray matter variation, and some of this variation is related to schizophrenia. The main limitation of this study includes 1) the lack of replication data to verify the findings, and 2) the lack of direct saliva and brain tissue correspondence verification.

THE INTERPLAY BETWEEN OLIGO-TARGET SPECIFIC AND GENOME-WIDE OFF-TARGET INTERACTIONS

Olga V. Matveeva[1], Nafisa N. Nazipova[2], Aleksey Y. Ogurtsov[3], Svetlana A. Shabalina[3]

[1]Biopolymer Design LLC, Acton, MA; [2]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia; [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

Many techniques of molecular biology involve interaction of specific oligonucleotides with DNA or RNA as a basic step. DNA targeting of single-guided (sg)RNAs for genome editing procedures, oligonucleotide array gene expression monitoring or anti-sense-mediated gene down-regulation, and the Genomic Comparison Hybridization (GCH) array experiments are examples of techniques involving RNA-DNA and DNA-DNA interactions. RNAi approaches with siRNA and shRNA molecules are based on RNA-RNA interactions. The main problem of any oligo-probe experiment is that the specific oligo-target interaction, based on fully paired duplex, are usually combined with non-specific parallel reactions, where oligo-probe could interact with many partially paired DNA or RNA sequences. The interplay between specific and genome-wide off-target interactions is poorly studied despite its crucial role in the efficacy of these techniques. In this study, we investigated oligo-probe characteristics, which are responsible for the interplay, and which most improve the oligo-probe design. We defined specificity of interaction as a ratio between oligo-target specific and genome-wide off-target interactions. Microarray databases, derived from the GCH experiments using the Affymetrix platforms, and containing two different types of probes were used for the analysis based on the thermodynamic features and nucleotide sequences of oligo-probes. The first type of oligo-probe does not have a specific target on the genome and their hybridization signals are derived from genome-wide cross-hybridization alone. The second type includes oligonucleotides that have a specific target on the genomic DNA and their signals are derived from specific and cross-hybridization components combined together in a total signal. The analysis has revealed that hybridization specificity was negatively affected by low stability of the fully-paired oligo-target duplex, stable probe self-folding, G-rich content, including GGG motifs, low sequence Symmetrical Complexity (SC) score. The SC-score characterizes nucleotide composition symmetry and probe's vulnerability to off-target interactions. Filtering out the probes with these characteristics significantly increases hybridization specificity by decreasing genome-wide cross-hybridization or by increasing specific interactions. Selected oligo-probes have three times higher hybridization specificity on average, compared to the probes that were filtered out from the analysis by applying suggested cut-off thresholds to the described parameters. Multiple regression models with described parameters were successfully applied for predictions of interaction specificity and off-target effects and supported parameter choice ($P<0.001$). We also compared probe characteristics selected for the analysis in microarray databases with applicable features of siRNA/shRNA design from our earlier studies. We applied all selected oligonucleotide features and described parameters to new sets of sgRNAs. Our study examined the thermodynamics and sequence-intrinsic properties of sgRNA-DNA duplexes and analyzed additional selection criteria that are critical for guide efficacy. Finally, we identify universal features of oligo-probes, si/shRNAs and guides for optimal design including the SC-score.

PATTERNS IN BIOMEDICAL DATA – HOW DO WE FIND THEM?

POSTER PRESENTATIONS

WARS2 IMPLICATED AS A COMMON MODIFIER OF METFORMIN METABOLITE BIOMARKERS IN A BIOBANK COHORT

Alyssa I. Clay[1], Richard M. Weinshilboum[2], K. Sreekumaran Nair[3], Rima F. Kaddurah-Daouk[4], Liewei Wang[2], Matthew K. Breitenstein[1]

[1]Division of Epidemiology, Mayo Clinic; [2]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic; [3]Division of Endocrinology, Mayo Clinic; [4]Duke University

Background Metformin is one of the most widely prescribed drugs worldwide and a first line treatment for type 2 diabetes mellitus(T2D). Metformin has many mechanisms of action, with varying levels of understanding. Metformin is being evaluated as a potential chemoprevention agent for cancer treatment, with inhibition of angiogenesis as one affect of metformin being strongly pursued. However, contradictory evidence exists for a potential mechanism of angiogenesis inhibition (Carcinogenesis 2014;(35)5). Building on our prior work that identified stratum of statistically correlated metabolites, we aimed to identify overlapping metformin pharmacogenomic(PGx) SNP associations, using pharmacometabolomics informed PGx paired with an agnostic computational approach. Methods To elucidate overlapping PGx signals of metformin exposure, we included metabolites (n=5) with correlated plasma concentration, adjusted for metformin exposure, in a biobank cohort-based, case-control study. Cases (n=274) were exposed to metformin monotherapy with T2D; healthy controls (n=274) had no known drug exposures. Cases and controls were matched by age and gender, and adjusted for BMI and batch. A panel of amino acid metabolite (n=42) concentrations was quantitatively measured using tandem liquid chromatography-mass spectrometry from fasting platelet poor plasma samples collected in EDTA. Genotyping was performed using the 700k SNP Illumina Omni Express array platform from 250ng of DNA. Normalized metabolite concentrations were utilized as endpoints to inform genome wide associations. Results Increased plasma metabolite concentrations for leucine(t=4.47,p=<0.0001), isoleuceine (t=4.63,p=<0.0001), and valine(t=4.48,p=<0.0001) were observed with exposure to metformin. Variant rs17023164(MAF=0.31), in the Tryptophanyl TRNA Synthetase 2, Mitochondrial (WARS2) gene region of chromosome 1 and an eQTL for WARS2 in fibroblasts, was a common downward modifier of leucine(β=-11.69,p=1.79e-7), isoleuceine (β =-6.99,p=2.40e-6), and valine(β =-14.55,p=1.04e-5) with metformin exposure. No SNPs in neighboring genes regions were in high LD (R^2>0.5) with rs17023164. Conclusion Increased plasma metabolite concentrations for leucine, valine, and isoleucine were observed with metformin exposure. A common variant, rs17023164 in WARS2, was identified as a strong downward modifier of these metabolites with metformin exposure. Independently, WARS2 is proposed as a determinant of angiogenesis (Nat Com 2016;(7)12061). We posit a hypothesis: modification of metabolite biomarker concentration associated with metformin exposure by WARS2 variants is a potential link between metformin and angiogenesis. Functional characterization of a potential mechanism for metformin inhibition of angiogenesis, modified by WARS2, is ongoing.

ESTIMATION OF FALSE NEGATIVE RATES VIA EMBEDDING SIMULATED EVENTS

Stephen V. Gliske[1], Katy L. Lau[1], Benjamin H. Brinkman[2], Greg A. Worrell[2], Cris G. Fink[3], William C. Stacey[1]

[1]University of Michigan, [2]Mayo Clinic, [3]Ohio Wesleyan University

Automated event detection is the result of many types of data-driven pattern recognition methods. One of the general challenges to these analyzes is the quantification and correction for false negative detections, i.e., cases where the event (pattern) is present in the data but was not detected. Estimating the false positive rate is much easier, as human review of a subsample of detected events is sufficient. However, determining the false negative rate by human review would require manual searching through the raw data, which is impractical, if not completely infeasible. This challenge is not unique to biomedical data and is commonly addressed in high energy physics. The approach is called embedding. It is applicable to any analysis where at least one of the signal or background can be modeled well by simulations. By placing specific events at known locations, one can then run the automated detector and report the fraction of embedded events that were detected. We present the first application of embedding to neurological data, specifically the automated detection of a biomarker of epilepsy (high frequency oscillations) recorded in intracranial electroencephalogram (EEG) data. The false negative rate is found to be consistent across both recording channel and across patients.

INTEGRATIVE, INTERPRETABLE DEEP LEARNING FRAMEWORKS FOR REGULATORY GENOMICS AND EPIGENOMICS

Chuan Sheng Foo, Avanti Shrikumar, Johnny Israeli, Peyton Greenside, Chris Probert, Anna Scherbina, Rahul Mohan, Nathan Boley, Anshul Kundaje

Stanford University

We present generalizable and interpretable supervised deep learning frameworks to predict regulatory and epigenetic state of putative functional genomic elements by integrating raw DNA sequence with diverse chromatin assays such as ATAC-seq, DNase-seq or MNase-seq. First, we develop novel multi-channel, multi-modal CNNs that integrate DNA sequence and chromatin accessibity profiles (DNase-seq or ATAC-seq) to predict in-vivo binding sites of a diverse set of transcription factors (TF) across cell types with high accuracy. Our integrative models provide significant improvements over other state-of-the-art methods including recently published deep learning TF binding models. Next, we train multi-task, multi-modal deep CNNs to simultaneously predict multiple histone modifications and combinatorial chromatin state at regulatory elements by integrating DNA sequence, RNA-seq and ATAC-seq or a combination of DNase-seq and MNase-seq. Our models achieve high prediction accuracy even across cell-types revealing a fundamental predictive relationship between chromatin architecture and histone modifications. Finally, we develop DeepLIFT (Deep Linear Importance Feature Tracker), a novel interpretation engine for extracting predictive and biological meaningful patterns from deep neural networks (DNNs) for diverse genomic data types. DeepLIFT can integrate the combined effects of multiple cooperating filters and compute importance scores accounting for redundant patterns. We apply DeepLIFT on our models to obtain unified TF sequence affinity models, infer high resolution point binding events of TFs, dissect regulatory sequence grammars involving homodimer and heterodimeric binding with co-factors, learn predictive chromatin architectural features and unravel the sequence and architectural heterogeneity of regulatory elements.

VISUALIZATION OF COMPLEX DISEASES AND RELATED GENE SETS

Modest von Korff, Tobias Fink, Thomas Sander

Actelion Pharmaceuticals Ltd., Allschwil, Switzerland

The relations between genes and diseases form complex patterns. Visualization of these patterns enables the scientist to obtain an overview of the most important gene–disease relations. These gene–disease relations are of high importance in drug discovery. Proteins encoded by disease-related genes are potential targets for new drugs or may become biomarkers for disease diagnosis. Both a novel drug target and a biomarker should be highly specific for the aimed disease. In our publication for this conference, we introduce a relevance estimator. This relevance estimator is a measure of the specificity of a gene–disease relationship that also takes into consideration all other known gene–disease relationships. We analyzed gene–disease relationships from 22 million PubMed records and obtained a matrix with relevance estimators for about 5000 diseases and 15,000 genes. This relevance matrix enabled us to express the similarity between diseases with simple vector-based distance measures. A meaningful disease–gene–disease visualization, consisting of several layers, was derived from these disease–disease similarity measures and the relevance estimators. The multidimensional visualizations presented here give an overview of complex diseases like asthma, Alzheimer's disease and hypertension.

PRECISION MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS IMPROVED HEALTH AND THERAPIES

POSTER PRESENTATIONS

FINDINGS FROM THE FOURTH CRITICAL ASSESSMENT OF GENOME INTERPRETATION, A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Steven E. Brenner[1], Gaia Andreoletti[1], Roger A Hoskins[1], John Moult[2], CAGI Participants,

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

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. CAGI participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors.

The fourth CAGI experiment concluded this year. It included 11 challenges which reflected: non-synonymous variants and their biochemical impact measured by targeted assays; noncoding regulatory variants and their impact on gene expression; research exomes for prediction of complex traits; personal genomes and trait profiles; and clinical sequences and associated referring indications.

There were notable discoveries throughout the CAGI experiment, and general themes emerged. The independent assessment found that top missense prediction methods are highly statistically significant, but individual variant accuracy is limited. Moreover, missense methods tend to correlate better with each other than with experiments (for reasons that may reflect the predictive methods and the assays themselves). However, there might be potential for missense interpretation at the extreme of the distribution. Structure-based missense methods excel in a few cases, while evolutionary-based methods have more consistent performance. Bespoke approaches often enhance performance.

On the clinical studies, predictors were able to identify causal variants that were overlooked by the clinical laboratory, and it appears that physicians may not always order the most relevant genetic test for their patients. CAGI data show that running multiple uncalibrated methods and considering their consensus often provides undue confidence in their correlation; we therefore advise against running multiple uncalibrated variant interpretation tools in clinical analysis.

The results showed that predicting complex traits from exomes is fraught. Interpretation of non-coding variants shows promise but is not at the level of missense. Beyond this, creating a genetic study that provides a reliable gold standard is remarkably difficult. However, there were notable improvements in the ability to match genomes to trait profiles.

Complete information about CAGI may be found at https://genomeinterpretation.org.

ASTROLABE: EXPANSION TO CYP2C9 AND CYP2C1

Andrea Gaedigk[1], Greyson P. Twist[2], Sarah Soden[2], Emily G. Farrow[2], Neil A. Miller[2]

[1]Division of Clinical Pharmacology & Therapeutic Innovation, Children's Mercy, Kansas City, School of Medicine, University of Missouri-Kansas City; [2]Center for Pediatric Genomic Medicine, Children's Mercy, Kansas City

Background: CYP2C9 and 19 are highly polymorphic pharmacogenes metabolizing numerous drugs. Both are genes with CPIC guidelines underscoring their clinical relevance. To facilitate haplotype calling and translation into phenotype, we have developed a probabilistic scoring system, Astrolabe (initially called Constellation; Twist et al 2016, Gen Med 1:15007) that enables automated CYP2D6 diplotype calling from whole genome sequencing. We report here the extension of Astrolabe to CYP2C9 and 2C19. Methods: The study was approved by the Institutional Review Board of Children's Mercy and included 85 subjects (7 HapMap; 78 patients/parents). Allele definitions are according to the P450 Nomenclature Database (cypalleles.ki.se/) with some modifications. Exons and 100bp of flanking introns were used for Astrolabe calls as well as -2990 to -440 of CYP2C9 and -1063 to -180 of CYP2C19 harboring SNPs defining CYP2C9*8 and CYP2C19*27, respectively. All but 3 subjects were genotyped for CYP2C9*2, *3, *5 and *8 and CYP2C19*2-*4, *17, *27 and *35 using TaqMan assays to validate Astrolabe calls. WGS data were reanalyzed with the DRAGEN Bio-IT processor (Edico Genome) to improve variation call quality. To account for haplotype and diplotype combinations not observed in our sample set simulations of all possible diplotype combinations were performed using the ART read simulator and DRAGEN analysis pipeline. Astrolabe is available at https://www.childrensmercy.org/genomesoftwareportal/ Results: To maximize Astrolabe call accuracy, intron regions were adjusted to include informative SNPs while excluding those that occur on numerous haplotypes and/or are not part of a defined allele. The CYP2C9 exon1 region, e.g. was limited to 57bp of intron1 to exclude 251T>C, which is present in 1155/3540 subjects (CMH variant warehouse database). This SNP defines CYP2C*29, but interfered with Astrolabe calls by overcalling CYP2C*29 in the absence of its key SNP (33437C>A). Optimized calling target regions were then used to compare Astrolabe with genotype calls. Astrolabe correctly called 68/75 (90.67%) and 71/75 (94.67%) of subjects for CYP2C9 and 19, respectively. Among the alleles detected by Astrolabe and genotyping were CYP2C9*2, *3 and *8 and CYP2C19*2, *17, *27 and *35. Astrolabe also identified subjects carrying the rare CYP2C9*9 and *11 and CYP2C19*15 alleles which were not covered by genotyping. Astrolabe correctly called 1077/1128 simulated CYP2C19 diplotypes (95% recall; 45 missed and 6 multiple calls). All missed calls were *12 called as *1. For CYP2C9, Astrolabe correctly called 2186/2278 simulated diplotypes (95% recall; 61 missed and 31 multiple calls). All missed calls were *25 called as *1. Discussion: Astrolabe's functionality was successfully expanded to CYP2C9 and 19. Phenotype prediction based on Astrolabe was superior over that derived from a limited genotype panel. Continued improvement and expansion of the nomenclature definitions will allow us to resolve the miscalled haplotypes represented in the simulation set and improve Astrolabe calling across all diplotypes.

HUMAN KINASES DISPLAY MUTATIONAL HOTSPOTS AT COGNATE POSITIONS WITHIN CANCER

Jonathan Gallion, Angela D. Wilkins, Olivier Lichtarge

Baylor College of Medicine

The discovery of driver genes is a major pursuit of cancer genomics, usually based on observing the same mutation in different patients. But the heterogeneity of cancer pathways plus the high background mutational frequency of tumor cells often cloud the distinction between less frequent drivers and innocent passenger mutations. Here, to overcome these disadvantages, we grouped together mutations from close kinase paralogs under the hypothesis that cognate mutations may functionally favor cancer cells in similar ways. Indeed, we find that kinase paralogs often bear mutations to the same substituted amino acid at the same aligned positions and with a large predicted Evolutionary Action. Functionally, these high Evolutionary Action, non-random mutations affect known kinase motifs, but strikingly, they do so differently among different kinase types and cancers, consistent with differences in selective pressures. Taken together, these results suggest that cancer pathways may flexibly distribute a dependence on a given functional mutation among multiple close kinase paralogs. The recognition of this “mutational delocalization” of cancer drivers among groups of paralogs is a new phenomena that may help better identify relevant mechanisms and therefore eventually guide personalized therapy.

SCOTCH: A NOVEL METHOD TO DETECT INSERTIONS AND DELETIONS FROM NGS DATA

Rachel Goldfeder, Euan Ashley

Stanford University

Clinical-grade genome sequencing and interpretation requires accurate and complete genotype calls across the entire genome. While single nucleotide variant detection is highly accurate and consistent, these variants explain only a small fraction of disease risk. Other types of variation that disrupt the open reading frame, such as insertions and deletions (INDELs), are more likely to be harmful. However, current methods have low sensitivity for larger (>= five bases) INDELs, primarily due to challenges surrounding aligning sequence reads that span INDELs. We present Scotch, a novel INDEL detection method that leverages signatures of poor read alignment, read depth information, and machine learning approaches to accurately identify INDELs from next-generation DNA sequencing data. Using biologically realistic simulated genomes and sequence reads with technologically representative error profiles (generated by ART), we evaluate Scotch and several currently available INDEL callers. We show that Scotch has higher sensitivity than current methods, particularly for larger INDELs. Finally, we validate INDELs that Scotch discovered in one individual, NA12878, and show that Scotch has high positive predictive value. This method will enable researchers and clinicians to more accurately identify INDELs associated with previously unexplained genetic conditions.

MAYO OMICS REPOSITORY FOR TRANSLATIONAL MEDICINE

Iain Horton, Jeanette Eckel-Passow, Steven Hart, Shannon McDonnell, David Mead, Gay Reed, Greg Dougherty, Jason Ross, Julie Swank, Mark Myers, Mathieu Wiepert, Rama Volety, Tony Stai, Yaxiong Lin, Robert Freimuth

Mayo Clinic

The Mayo Clinic Genomic Data Warehouse has established the infrastructure foundation, processes, and applications to meet the translational needs of the Mayo Clinic Center for Individualized Medicine (CIM). Through the streamlined and automated data pipeline, the next-gen sequencing (NGS) results are loaded and integrated with clinical data, providing the foundation for the development of revolutionary solutions and discovery in the clinical practice and genomic research. Initiated in 2012, with production data ingestion beginning in early 2014, Mayo Clinic's Translational Research Center (TRC) has provided the cornerstone platform for data centric activities within CIM. Data generated from both the clinical pipeline and research pipeline are automatically loaded into TRC with each new bit adding value and power to the system. Two key solutions with significant potential of impacting patient care and scientific discovery have been built on this genomic data warehouse. First is the Molecular Decision Support system, a rule-based pharmacogenomics system that enables Mayo Clinic clinicians to integrate actionable information based on a patient's genotype information at the point of care using NGS data. Second is the Mayo Variant Summary application, a cloud-native system which empowers Mayo Clinic researchers to identify rare and actionable genomic variants through dynamic filtering and grouping of subject phenotype and specimen metadata.

PHARMACOGENOMICS CLINICAL ANNOTATION TOOL (PHARMCAT)

T.E. Klein[1], M. Whirl-Carrillo[1], R.M. Whaley[1], M. Woon[1], K. Sangkuhl[1], Lester G. Carter[1], H.M. Dunnenberger[2], P.E. Empey[3], A.T. Frase[4], R.R. Freimuth[5], A. Gaedigk[6], A. Gordon[7], C. Haidar[8], J.K. Hicks[9], J.M. Hoffman[8], M.T. Lee[10], N. Miller[11], S.D. Mooney[12], T.N. Person[13], J.F. Peterson[14], M.V. Relling[8], S.A. Scott[15], G. Twist[11], A. Verma[13], M.S. Williams[10], C. Wu[16], W. Yang[8], M.D. Ritchie[4,13]

[1]Dept Genetics, Stanford Univ, Stanford, CA; [2]Center for Molecular Medicine, NorthShore University Health System, Evanston IL; [3]Department of Pharmacy and Therapeutics, School of Pharmacy, University of Pittsburgh; [4]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA; [5]Department of Health Sciences Research, Mayo Clinic, Rochester MN; [6]Division of Clinical Pharmacology, Toxicology & Therapeutic Innovation, Children's Mercy-Kansas City, Kansas City, MO; [7]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; [8]St. Jude Children's Research Hospital, Memphis, TN; [9]DeBartolo Family Personalized Medicine Institute, H. Lee Moffitt Cancer Center, Tampa, FL; [10]Genomic Medicine Institute, Geisinger Health System, Danville, PA; [11]Center for Pediatric Genomic Medicine, Children's Mercy, Kansas City, MO; [12]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA; [13]Biomedical and Translational Informatics, Geisinger Health System, Danville, PA; [14]Vanderbilt University Medical Center, Nashville, TN; [15]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; [16]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA

Pharmacogenomics (PGx) decision support and return of results is an active area of genomic medicine implementation at many health care organizations and academic medical centers. The Clinical Pharmacogenetics Implementation Consortium (CPIC) has established guidelines surrounding gene-drug pairs that can and should lead to prescribing modifications based on genetic variant(s). One of the challenges in implementing PGx is extracting genomic variants and assigning haplotypes (including star- alleles) from genetic data derived from sequencing and genotyping technologies in order to apply the prescribing recommendations of CPIC guidelines. In a collaboration between the PGRN Statistical Analysis Resource (P-STAR), The Pharmacogenomics Knowledgebase (PharmGKB), the Clinical Genome Resource (ClinGen), and CPIC, we are developing a software tool to extract all variants from CPIC level-A genes with the exception of G6PD and HLA, from a genetic dataset resulting from sequencing or genotyping technologies (represented as a .vcf), interpret the variant alleles, infer diplotypes, and generate an interpretation report based on CPIC guidelines. The CPIC pipeline report can then be used to inform prescribing decisions. We assembled a focus group of thought leaders in PGx to brainstorm the issues and to design the software pipeline. We hosted a one-week Hackathon at the PharmGKB at Stanford University to bring together computer programmers with scientific curators to implement the first version of this tool. Through this process, we have uncovered many of the challenges surrounding PGx implementation. For example, the inference of diplotypes is challenging for several CPIC level-A genes. This software pipeline will be made available under the Mozilla Public License (MPL 2.0) and disseminated in Github for the scientific and clinical community to test, explore, and improve. PharmCAT will provide a solution that will enable sites implementing PGx a way to more consistently interpret genomic results and link those results to published clinical guidelines. Furthermore, we are assembling (and will be maintaining) the translation tables that underlie the tool, which will significantly reduce the effort required to implement PGx clinically and ensure more uniform interpretations of PGx knowledge. As precision medicine continues to move into clinical practice, implementation workflows for PGx, like PharmCAT, would enable standardized and consistent implementation of PGx genes.

PCSK9 MODULATING VARIANTS IN FAMILIAL HYPERCHOLESTEROLEMIA

Sarathbabu Krishnamurthy[1], Diane Smelser[1], Manickam Kandamurugu[1], Joseph Leader[1], Noura S. Abul-Husn[2], Alan R. Shuldiner[2], David H. Ledbetter[1], Frederick E. Dewey[2], David J. Carey[1], Michael F. Murray[1], Raghu P. R. Metpally[1]

[1]Geisinger Health System; [2]Regeneron Genetics Center

BACKGROUND: Highly penetrant autosomal dominant familial hypercholesterolemia (FH) is known to be caused by pathogenic loss of function (LOF) variants in LDLR and gain of function variants in PCSK9 and APOB genes. In addition to its causative role in FH, PCSK9 LOF variants are associated with lowering of serum low density lipoprotein cholesterol (LDL-C) and total cholesterol. The aims of this study were to 1. Identify rare novel PCSK9 gene variants that lead to complete or partial loss of protein function in the DiscovEHR cohort. 2. Explore prevalence of PCSK9 LOF variants in a subset of FH patients and 3. Examine whether FH patients carrying PCSK9 LOFs show association with lowering the plasma low density LDL-C and cardiovascular risk. METHODS: We analyzed whole exome sequences from 51,289 individuals in the DiscovEHR cohort, who consented to participate in the Geisinger Health System's MyCode Community Health Initiative. Rare missense and predictive loss of function (pLOF) coding variants in PCSK9 were identified by integrating bioinformatics and evaluating LDL-C and total cholesterol measures from the electronic health records (EHR). RESULTS: In the overall DiscovEHR cohort, we identified 20 missense and 13 pLOFs (2 splice donor, 6 stop gained and 5 frame shift) rare variants in PCSK9, including 15 novel variants that were associated with lower LDL-C and total cholesterol levels. LDL-C in pLOF carriers was significantly lower than in missense carriers with presumed partial loss of function ($p < 0.0012$). Patients with PCSK9 rare missense with presumed partial LOF or LOF variants had significant reduction in the incidence of coronary events compared to the control group ($p < 0.0001$). In FH patients, the LDL-lowering PCSK9 R46L variant previously reported as 3% prevalence was found to be enriched at 9.6% and was associated with lower LDL-C compared to FH patients not carrying an R46L allele. A novel PCSK9 missense variant (G316S) was also present in FH patients with a prevalence of 0.8% and also showed an LDL-lowering phenotypic effect in an imputed family pedigree. CONCLUSIONS: Overall 11.8% of the FH patients in the DiscovEHR cohort were identified to also carry a PCSK9 variant which modulates their LDL-C and serum cholesterol levels.

INTEGRATIVE NETWORK ANALYSIS OF PROSTATE TISSUE LINCRNA-MRNA EXPRESSION PROFILES REVEALS POTENTIAL REGULATORY MECHANISMS OF PROSTATE CANCER RISK LOCI

Nicholas B. Larson[1], Shannon McDonnell[1], Zach Fogarty[1], Melissa Larson[1], John Cheville[2], Shaun Riska[1], Saurabh Baheti[1], Asha A. Nair[1], Daniel O'Brien[1], Jaime Davila[1], Daniel Schaid[1], Stephen N. Thibodeau[2]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN; [2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN

Large-scale genome-wide association studies have identified 146 loci associated with risk of developing prostate cancer (PRCA). However, most of these loci do not lie in close proximity to protein coding genes and are presumed to be regulatory in nature. Downstream regulation of protein coding genes related to PRCA development may be mediated by cis-acting regulation of nearby transcripts, also known as cis-mediated trans-eQTLs. This cis-mediator causal relationship is comprised of a regulatory variant, a nearby cis-regulated gene, and the downstream regulated trans target gene. Cis-mediators may include transcription factors, signaling proteins, and long intergenic non-coding RNAs (lincRNAs). LincRNAs correspond to a host of regulatory functions such as chromatin remodeling and transcriptional co-activation, and have previously been identified as diagnostic and prognostic biomarkers for a number of cancers. However their role in cancer development and progression is poorly understood. To explore the hypothesis that cis-mediated trans eQTLs may play a role in PRCA risk, we leveraged an eQTL dataset of 471 samples of normal prostate tissue from prostate/bladder cancer patients with available RNA-Seq and imputed Illumina Infinium 2.5M genotype data. We first conducted an initial transcriptome-wide eQTL screening of all lincRNAs and mRNAs with 8,073 SNPs in high linkage disequilibrium ($r^2>0.5$) with previously identified PRCA risk-associated variants, identifying approximately 5000 transcripts (FDR<0.10) to be putatively associated (cis or trans). We then constructed an undirected Gaussian graphical regulatory network from the expression profiles of this transcript subset, identifying 87,468 connections. To identify candidate cis-mediator node-pairs in the expression network, we isolated a subset of cis-associated transcripts (lincRNA or mRNA) at a strict Bonferroni significance threshold. We then identified all connected mRNA nodes to these cis-nodes that distal to the cis-variant (>1 Mb) and had evidence of a trans-association with the cis variant (P<1E-04), resulting in 9 candidate cis-mediator trios. Finally, we applied causal mediation analysis to test the proportion of the trans-association that is mediated by the cis-regulated transcript, resulting in 7/9 significant cis-mediator relationships. Transcription factor HNF1B was identified to be a significant mediator in the trans-associations between rs11263762 and three mRNAs: SRC, MIA2, and SEMA6A. All three exhibited concomitant upregulation with HNF1B. Notably, HNF1A has been shown to stimulate SRC expression via an alternative promoter, while MIA2 is also a known HNF1A target. Dysregulation of SEMA6A has been observed in PRCA metastases and plays a potential role in angiogenesis interacting with VEGFR2. MSMB and NDRG1 both demonstrate androgen-stimulated expression in prostate tissue, and indicated a recessive pattern of expression dysregulation with rs10993994. Despite a small sample size, we replicated multiple trans-eQTLs from these cis-mediator trios in the GTEx prostate tissue eQTL dataset (P<0.05). Together, our findings suggest dysregulation of RNA expression may play a role in genetic predisposition to PRCA.

INTEGRATED ANALYSIS OF GENOMICS, PROTEOMICS, AND PHOSPHOPROTEOMICS IN CELLS AND TUMOR SAMPLES

Jason E. McDermott[1], Tao Liu[1], Samuel Payne[1], Vladislav Petyuk[1], Richard Smith[1], Philipp Mertins[2], Steven Carr[2], Karin Rodland[1]

[1]Pacific Northwest National Laborator, [2]Broad Institute

As part of the Clinical Proteomic Tumor Analysis Consortium (CPTAC), we have recently published the first large-scale proteomic and phosphoproteomic analysis of high-grade serous ovarian tumors. We observed that phosphorylation status was an excellent indicator of pathway activity and could discriminate between patient survival times. In the current work we have combined this data with comparable data from breast cancer tumors and cancer cell lines treated with kinase inhibitors, to answer several fundamental questions about the role of phosphorylation in cellular processes and cancer. The total dataset comprised over 150 samples with very deep proteomic coverage (>20,000 phosphopeptides confidently identified). We first found that the correlation between kinase protein abundance and abundance of phosphorylated target peptides was very low, indicating that kinase abundance is not a good proxy for phosphorylation status overall. However, highly correlated kinase-substrate pairs were significantly more likely to be true relationships (from existing knowledge), demonstrating that this method could be used to predict novel kinase targets in some cases. We used this analysis to identify several novel kinase-substrate relationships that were differential between tumor subtypes, and that correlated with pathways where phosphorylation was affected by drug treatment. These relationships are currently under investigation as potential novel targets for therapeutic intervention. To better analyze cancer-relevant pathway activity we developed a novel approach that characterizes correlation, differential abundance, and statistical interactions between components to analyze multiple omics types in the context of signaling and functional pathways. We used this approach, called the Layered Enrichment Analysis of Pathways (LEAP), to identify active pathways in molecular subtypes of ovarian and breast cancer, and several novel subpopulations of patients displaying uniquely dysregulated pathways. Our results show that integration of multiple omics types has great potential in the area of development of novel therapeutic approaches for personalized medicine.

NETDX: PATIENT CLASSIFICATION USING INTEGRATED PATIENT SIMILARITY NETWORKS

Shraddha Pai, Shirley Hui, Ruth Isserlin, Hussam Kaka, Gary D. Bader

The Donnelly Centre, University of Toronto

Patient classification has widespread biomedical and clinical applications, including diagnosis, prognosis, disease subtyping and treatment response prediction. A general purpose and clinically relevant prediction algorithm should be accurate, generalizable, be able to integrate diverse data types (e.g. clinical, genomic, metabolomic, imaging), handle sparse data and be intuitive to interpret. We describe netDx, a supervised patient classification framework based on patient similarity networks, that meets the above criteria (Ref 1). netDx models input data as patient networks, and uses network integration and machine learning for feature selection. We demonstrate the utility of netDx by integrating gene expression and copy number variants to classify breast cancer tumours as being of the Luminal A subtype (N=348 tumours; Ref 2). Using gene expression data, netDx performed as well as or better than established state of the art machine learning methods, achieving a mean accuracy of 89% (2% s.d.) in classifying Luminal A. In the second application, we predict case/control status in autism spectrum disorders based on the occurrence of rare copy number deletions in metabolic pathways (N=3,291 patients; Ref 3); this predictor achieved better performance than previously published methods. netDx uses pathway features to aid biological interpretability and results can be visualized as an integrated patient similarity network to aid clinical interpretation. Upon publication, netDx software will be made publicly available via github; the software provides worked examples and easy-to-use functions for design of custom predictor workflows. More at http://netdx.org References: 1. netDx preprint: http://dx.doi.org/10.1101/084418 2. The Cancer Genome Atlas (2012) Nature 490: 61. 3. Pinto et al. (2014). Am J Hum Gen. 94 (5):677.

PREVALENCE AND DETECTION OF LOW-ALLELE-FRACTION VARIANTS IN CLINICAL CANCER SAMPLES

Hyun-Tae Shin[1,2], Jae Won Yun[1,2], Nayoung K. D. Kim[1], Yoon-La Choi[2,3], Woong-Yang Park[1,2,4], Peter J. Park[5]

[1]Samsung Genome Institute, Samsung Medical Center, Seoul, Korea; [2]Samsung Advanced Institute of Health Science and Technology, Sungkyunkwan University, Seoul, Korea; [3]Department of Pathology & Translational Genomics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea; [4]Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Seoul, Korea; [5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Clinical application of sequencing-based assays requires high sensitivity and specificity for detecting genomic alterations. Our analysis of more than 5000 cancer samples reveals that a significant fraction of clinically-actionable somatic variants may have low variant allele fractions (VAF), indicating the importance of very high coverage sequencing for these patients. As a case study, we describe refractory cancer patients with clinical response to therapies that target low VAF alterations.

A METHYLATION-TO-EXPRESSION FEATURE MODEL FOR GENERATING ACCURATE PROGNOSTIC RISK SCORES AND IDENTIFYING DISEASE TARGETS

Jeffrey A. Thompson[1], Carmen J. Marsit[2]

[1]Dartmouth College, [2]Emory University

Many researchers now have available multiple high-dimensional molecular and clinical datasets when studying a disease. As we enter this multi-omic era of data analysis, new approaches that combine different levels of data (e.g. at the genomic and epigenomic levels) are required to fully capitalize on this opportunity. In this work, we outline a new approach to multi-omic data integration, which creates a model of methylation dysregulation and its effect on gene expression and then combines this molecular information with clinical predictors as part of a single analysis to create a prognostic risk score for clear cell renal cell carcinoma. The approach integrates data in multiple ways and yet creates models that are relatively straightforward to interpret and with a high level of performance. Over 100 random splits of the data into training and testing sets, our model had the highest median C-index of any method we tried, at .792. Furthermore, we demonstrated that our molecular risk predictor is independent of clinical covariates and that the combined model results in statistically significantly higher accuracy than either data type alone. Additionally, the proposed process of data integration itself captures relationships in the data that represent highly disease-relevant functions. The gene signature we identify for clear cell renal cell carcinoma prognosis is enriched for genes that are central nodes in a protein-protein interaction network associated with the JAK-STAT signaling cascade, which itself is a known factor in kidney cancer progression. Our signature is also enriched for genes in pathways involved in immune response, which are increasingly targeted by novel cancer therapies. We call this model the methylation-to-expression feature model (M2EFM). Although one of the other approaches we considered also resulted in a highly accurate model, M2EFM performed better with a far more parsimonious model that sheds light on the potential relationship between abnormal gene regulation and cancer prognosis. Given our results, we think that further development of this approach is warranted.

CYP2D6 DIPLOTYPE CALLING FROM WGS USING ASTROLABE: UPDATE

Andrea Gaedigk[1], Greyson P. Twist[2], Sarah Soden[2], Emily G. Farrow[2], Neil A. Miller[2]

[1]Division of Clinical Pharmacology & Therapeutic Innovation, Children's Mercy, Kansas City, School of Medicine, University of Missouri-Kansas City; [2]Center for Pediatric Genomic Medicine, Children's Mercy, Kansas City

Background: To facilitate haplotype calling and translation into phenotype, we have previously developed a probabilistic scoring system, Astrolabe (initially called Constellation; Twist et al 2016, Gen Med 1:15007) enabling automated CYP2D6 diplotype calling from whole genome sequencing. We have implemented a series of improvements to increase call accuracy as well as ease of use. Methods: The Study was approved by the Institutional Review Board of Children's Mercy Kansas City and included a total of 85 subjects (7 HapMap; 78 patients/parents). WGS data were reanalyzed with the DRAGEN Bio-IT processor (Edico Genome) to improve the quality of variation calls. The Astrolabe CYP2D6 allele definition table was expanded to include a) additional variants available through the P450 Nomenclature Database; b) variants characterized by our laboratory, but not available through the Nomenclature Database; c) resequencing of some alleles (e.g. *10, *17) for which only exons are annotated by the Nomenclature Database. Programming errors in the scoring algorithm were repaired and unit tested as well as a broad range of variant file input types were included (vcf, gvcf, tabix, .gz). Improvements also include versioning of the Astrolabe tool and the nomenclature data from which calls are generated. To account for haplotype and diplotype combinations not observed in our sample set simulations of all possible diplotype combinations were performed using the ART read simulator and DRAGEN analysis pipeline. Astrolabe is available at https://www.childrensmercy.org/genomesoftwareportal/. Results: To maximize Astrolabe call accuracy, we removed CYP2D6*1E, *3B, *4A-L, *4N, *6D, *10C-D, and *45B from the call set, because of incomplete allele definitions (based on exons only), or SNP(s) that are not unique to an allele. For example, 1749A>G is part of the CYP2D6*3B and *103 definitions, but also appears to be present on some *1 subvariants. Likewise, 3288A>G is not limited to CYP2D6*6D as implied by the nomenclature database, thus causing erroneous Astrolabe calls. Calls with our revised definitions were compared with those obtained by genotyping. Astrolabe also accurately identified subjects with copy number variations including the CYP2D6*5 deletion (n=5) and gene duplications (n=2). Also, increased variant calling accuracy of the DRAGEN pipeline improved the calling of several samples (n=). Astrolabe correctly called 7731/8128 simulated diplotypes (95% recall); 133 missed and 264 multiple calls). Of the missed calls 124 were due to *38 called as *1. Discussion: The series of improvements to Astrolabe increased call accuracy and minimized the number of no calls. Phenotype prediction based on Astrolabe was superior over that derived from a limited genotype panel. Continued refinement of existing allele definitions and the inclusion of novel haplotype definitions will further improve the Astrolabe tool. We are currently applying Astrolabe to other NGS datasets including exomes and targeted NGS panels.

INTEGRATION, INTERPRETATION AND DISPLAY OF MULTI-OMIC DATA FOR PRECISION MEDICINE

David S. Wishart[1], Ana Marcu[1], AnChi Guo[1], Ash Anwar[2], Solveig Johannessen[3], Craig Knox[4], Michael Wilson[4], Christoph H. Borchers[5], Pieter Cullis[6], Robert Fraser[2]

[1]University of Alberta, [2]Molecular You Inc., [3]Educe Design Inc., [4]OMx Inc., [5]University of Victoria, [6]University of British Columbia

The goal of precision medicine is to use advanced multi-omic technologies to improve the accuracy of medical diagnoses and enhance the individualization of medical treatment. The fundamental challenge in precision medicine is not in the measurement or collection of multi-omic data but in its delivery. In particular, the integration, interpretation and display of multi-omic data has proven to be particularly problematic. Here we describe some of our experiences in tackling this problem and outline a number of important findings that we believe are worth sharing. Our most important finding was the need to use high quality, quantitative 'omics data. Measuring absolutely quantitative 'omics data ensures greater reproducibility and permits direct comparisons to well-established clinical reference values. Several 'omics laboratories offering quantitative services have been identified and these are described here. Second, we discovered that custom databases containing biomarker-disease data are essential. Very few of these kinds of databases exist, but they are necessary for the comparison and full integration of multi-omic data. In particular, they provide the information needed to integrate multi-omic measures and to determine disease risk. A brief description of a few of these biomarker-disease databases is provided. Third, we discovered that color-coded graphs, which are hyperlinked to detailed textual explanations, are necessary for the facile interpretation of the multi-omic data – both by patients and physicians. An example of a well-designed, web-enabled "dashboard" is shown to highlight these findings. Finally we found that comprehensive databases of actionable responses must be prepared so that detailed, customizable medical, lifestyle, diet or pharmacological guidance can be provided to treat or prevent conditions detected by these multi-omic measurements. Examples of several omics-derived, actionable responses are provided to clarify this point. These findings, along with several associated software tools and databases, have recently been integrated into an automatic workflow that allows a wide range of multi-omic measurements to be integrated, interpreted and displayed for precision or personalized medicine applications.

BIOTHINGS APIS: LINKED HIGH-PERFORMANCE APIS FOR BIOLOGICAL ENTITIES

Jiwen Xin[1], Cyrus Afrasiabi[1], Sebastien Lelong[1], Ginger Tsueng[1], Sean D. Mooney[2], Andrew I. Su[1], Chunlei Wu[1]

[1]The Scripps Research Institute, [2]The University of Washington

The accumulation of biological knowledge and the advance of web and cloud technology are growing in parallel. Recently, many biological data providers start to provide web-based APIs (Application Programming Interfaces) for accessing data in a simple and reliable manner, in addition to the traditional raw flat-file downloads. Web APIs provide many benefits over traditional file downloads. For instance, users can request specific data such as a list of genes of interest without having to download the entire dataset, thereby providing the latest data on demand and reducing computation and data transfer times. This means that programmers can spend less time on wrangling data, and more time on analysis and discovery. Building and deploying scalable and high-performance web APIs requires sophisticated software engineering techniques. We previously developed high-performance and scalable web APIs for gene and genetic variant annotations, accessible at MyGene.info and MyVariant.info. These two services are a tangible implementation of our expertise and collectively serve over 4 million requests every month from thousands of unique users. Crucially, the underlying design and implementation of these systems are in fact not specific to genes or variants, but rather can be easily adapted to other biomedical data types such drugs, diseases, pathways, species, genomes, domains and interactions. We are currently expanding the scope of our platform to other biological entities. Collectively, we refer them as "BioThings APIs" (http://biothings.io). We also applied JSON-LD (JSON for Linking Data) technology in the development of BioThings APIs. JSON-LD provides a standard way to add semantic context to the existing JSON data structure, for the purpose of enhancing the interoperability between APIs. We have demonstrated the applications of JSON-LD with BioThings APIs, including data discrepancy checks as well as the cross-linking between APIs.

SINGLE-CELL ANALYSIS AND MODELLING OF CELL POPULATION HETEROGENEITY

POSTER PRESENTATIONS

SINGLE CELL SIGNALING STATES REVEAL INDUCTION OF NON-GENETIC VARIATION IN RESISTANCE TO TRAIL-INDUCED APOPTOSIS

Reema Baskar, Harris Fienberg, Garry Nolan, Sean Bendall

Stanford University

TNFalpha-related apoptosis-inducing ligand (TRAIL) has been shown to specifically target cancer cells, however rampant resistance has curtailed its efficacy as a drug. Cell-to-cell variation has been previously linked to resistance to TRAIL-induced apoptosis. We further investigate non-genetic phenotypic variation as a novel mode of drug resistance. Using mass cytometry, we captured high-dimensional, single-cell signaling states of different cancer types over the course of TRAIL treatment. For the first time, we provide a comprehensive single cell overview of TRAIL signaling dynamics and provide population metrics to quantify heterogeneity within resistance phenotypes. We demonstrate that while all cells respond to TRAIL, a subset of them persist in transient resistant states and do not progress to apoptosis. Our methods show correlation between heterogeneity of response to TRAIL and persistence of non-apoptotic, viable cancer cells in drug. We also show that combinatorial therapies designed to inhibit implicated pathways in conserved resistant states do not eradicate resistance and in fact can induce new states of resistance. This study presents experimental and computational tools to investigate non-genetic phenotypic variation as a novel mode of drug resistance in cancer and demonstrates their utility in understanding resistance to TRAIL-induced apoptosis.

A NOVEL K-NEAREST NEIGHBORS APPROACH TO COMPARE MULTIPLE BIOLOGICAL CONDITIONS IN SINGLE CELL DATA

Tyler J. Burns[1], Garry P. Nolan[2], Nikolay Samusik[2]

[1]Stanford University School of Medicine, Dept. of Cancer Biology; [2]Stanford University School of Medicine, Baxter Laboratory for Stem Cell Biology

High dimensional single-cell data is routinely visualized in two dimensions using dimension reduction algorithms like t-SNE, Principle Components Analysis (PCA), or force-directed graphs. When comparing levels of intracellular proteins in basal versus perturbed cells, clustering must be used to visualize changes in specific markers in a single graph. However, discretizing a dataset does not allow one to understand subtle, rare, and/or continuous biological changes across the original manifold. Herein, we present an algorithm that represents each cell's information content as its average across k-nearest neighbors. This allows for comparisons to be made between biological conditions on a per-cell basis. We use this to produce detailed t-SNE maps depicting biological change, and correlation analysis to enumerate signaling responses to perturbation.

SINGLE-CELL RNA SEQUENCING IN PRIMARY GLIOBLASTOMA: IMPROVING ANALYSIS OF HETEROGENEOUS SAMPLES BY INCORPORATING QUANTIFICATION OF UNCERTAINTY

Wendy Marie Ingram, Debdipto Misra, Nicholas F. Marko, Marylyn Ritchie

Geisinger Health System

Background: Glioblastoma (GBM) is the most common and deadly brain cancer in adults. The associated lethality may be attributable to the intrinsic heterogeneity of micro-invasive tumor cells, some of which are unavoidably left behind following tumor resection. The transcriptomic heterogeneity may contribute to the survival and subsequent proliferation of a small subset of cells that are resistant to radiation and chemotherapy. It has long been hypothesized that investigations into these tumors at a single cell level will allow for better molecular understanding of treatment resistance and the development of novel therapeutic approaches. Recently, advances in single cell capture and sequencing technology have become available and allow for these studies to be conducted. However, there are many technical and computational challenges inherent to single cell transcriptomics that are not addressed by traditional RNA-seq analysis tools. These challenges include uncertainty of technical and biological variance and must be carefully considered in order for biologically and therapeutically relevant conclusions to be reached. Methods: Tumor tissue from two GBM patients undergoing surgical resection as part of standard of care therapy was collected at the time of surgery. We used the Fluidigm C1 microfluidics platform to capture single cells followed by RNA sequencing (RNA-seq) of these cells and a bulk population of ~10,000 cells from each tumor. We compared two different transcriptomic alignment tools, Bowtie and kallisto, and analyzed the single cell transcriptional heterogeneity of cells within and between tumors using the recently developed analysis tools, sleuth. To the best of our knowledge, we are the first to utilize this single cell capture method and perform single cell RNA-seq analysis using the newly developed kallisto and sleuth programs for primary GBM tissue samples. Results: We show that the Fluidigm C1 microfluidics single cell capture method produces high quality transcriptomic material for RNA-seq and may have benefits over alternative methods (e.g. fluorescence-activated cell sorting) such as shorter preparation time. The kallisto-sleuth analysis programs provide improved estimation of gene expression variability and more reliable clustering of single cells by leveraging the unique features of equivalency groups and bootstrap estimates of kallisto. Cluster analysis demonstrates that certain cells from both tumors cluster together and share some common expression patters, but the remaining cells cluster in tumor-specific groups or do not group with other cells. We observe marked intertumor and intratumor transcriptional variability and note that average expression from single cells does not reliably correlate with the bulk cell RNA-seq abundance estimates. Taken together, we have shown that the combination of Fluidigm C1 and the kallisto-sleuth analysis programs prove to be useful and reliable methods to obtain and analyze high quality single cell RNA-seq data for the investigation of primary tumor tissues.

REGISTRATION OF FLOW CYTOMETRY DATA USING SWIFT CLUSTER TEMPLATES TO REMOVE CHANNEL-SPECIFIC OR CLUSTER-SPECIFIC VARIATION

Jonathan A. Rebhahn[1], Sally A. Quataert[1], Gaurav Sharma[2], Tim R. Mosmann[1]

[1]Center for Vaccine Biology and Immunology, University of Rochester Medical Center;
[2]Department of Electrical and Computer Engineering, University of Rochester

Standardization between flow cytometry experiments performed at different times is difficult because variations in cell parameters can be caused by many factors, including changes in antibody reagents, staining protocols, cell handling, different cytometers, and cytometer settings such as photomultiplier amplification voltages. These variations may overwhelm the genuine biological differences being investigated, such as genetic or disease-specific variations between subjects. Technical variations can be partly reduced by manually adjusting analysis gates, but this is subjective and time-consuming. Previous methods for semi-automated adjustment have relied on histogram peaks or manual gating to identify anchor populations. We have now developed fully-automated methods for registering flow cytometry samples, i.e. normalizing the fluorescence intensity of each cell in all channels. We take advantage of the high-resolution cluster templates derived by clustering reference samples by the SWIFT algorithm. These templates represent Gaussian model descriptions of the multidimensional data. If samples to be registered are at least moderately similar to the target/reference sample, assignment of the test sample to the template results in most cells being assigned to the appropriate cluster, but clusters that have shifted in the test sample then have altered median values in one or more channels. This high-resolution positional information is used for two types of registration: Rigid, or per-channel registration compares cluster locations between the target and the test sample to be registered, and the best-fit registration adjustments are determined for each channel and applied incrementally, reassigning the cells at each step to improve the final fit. This objectively uses positional information from all clusters, regardless of cluster size variation, and successfully corrects global artifacts such as staining or cytometer settings that cause 'batch' differences between assay days. Fluid, or per-cluster registration calculates the registration adjustment required for each cluster in the test sample to overlap fully with its corresponding cluster in the reference sample. This registers clusters more completely, and can remove individual variation (due to e.g. genetic or disease-specific effects). Fluid registration removes most positional information - this is desirable if the main experimental outcome is expected to be variations of the number of cells of different types. This method has been applied to datasets that include changes due to assay dates, flow cytometers, subjects, and sequential blood samples. Most variation occurred between cytometers and assay days, less between subjects, and the least between different bleeds from the same person. Registration substantially improved correlations between cluster medians. The number of cells per cluster also showed increased correlation, suggesting that unmodified samples assigned to the cluster templates sometimes had cells assigned to an inappropriate cluster. Thus the SWIFT cluster-based registration can improve subsequent flow cytometry analysis. Registered samples can be analyzed by a variety of manual or automated procedures.

WORKSHOP: NO BOUNDARY THINKING IN BIOINFORMATICS

POSTER PRESENTATION

ENABLING RICHER DATA INTEGRATION FOR GENOMIC EPIDEMIOLOGY

E. Griffiths[1], D. Dooley[2], C. Bertelli[1], J. Adam[3], F. Bristow[3], T. Matthews[3], A. Petkau[3], M. Courtot[4], J. A. Carriço[5], A. Keddy[6], R. Beiko[6], L. M. Schriml[7], E. Taboada[8], M. Graham[3], G. Van Domselaar[3], W. Hsiao[2], F. Brinkman[1]

[1]SFU, Burnaby, BC, Canada; [2]BC Centre for Disease Control, Vancouver, BC, Canada; [3]PHAC, Winnipeg, MB, Canada; [4]EBI, Hinxton, Cambridge, UK; [5]Univ. of Lisbon, Lisbon, Portugal; [6]Dalhousie Univ., Halifax, NS, Canada; [7]Univ. of Maryland School of Medicine, Baltimore, MD, USA; [8]PHAC, Lethbridge, AB, Canada

One barrier to effectively capitalizing on whole genome sequence data is efficient, robust annotation and integration of associated contextual data (metadata). Whether human, microbial or other organismal genomic sequence, frequently such contextual data is too unorganized, in free text format, to enable effective integration for answering more sophisticated questions. Approaches to help overcome this barrier are illustrated here with the Integrated Rapid Infectious Diseases Analysis (IRIDA.ca) Project and Genomic Epidemiology Ontology (GenEpiO.org) Consortium. Microbial pathogen whole genome sequencing provides the highest resolution molecular "fingerprint" for infectious disease epidemiology and is transforming public health practice – enabling more rapid identification of disease outbreaks, their sources, and potential control measures. However, such microbial genomic data (like human 'omic data) must be combined with epidemiological/clinical/laboratory/other health care data ("contextual data") to be meaningfully interpreted for clinical and public health questions/actions. Furthermore, information must be shared between different agencies to efficiently assess and manage risks to human health across jurisdictions. Currently, terminologies describing public health data cannot be easily mapped across functionally-similar software systems without intricate intervention by specialists, resulting in data exchange systems that are static and fragile. To promote efficient data exchange and intelligence sharing, we propose an intuitive platform for searching, identifying, and verifying the fundamental health care entity elements (ontology terms) to map to institutional application data formats, starting with genomic and public health contextual data. Key innovations are the proposed Genomic Epidemiology Entity Mart (GE2M) that allows users to inspect term definitions, labeling, and database cross references in a user-friendly format, plus a software system allowing different jurisdictions to use the terms suitable for them, essentially choosing from a "shopping cart" of options mapped between jurisdictions/organizations. A very preliminary prototype of this concept has been established as part of the IRIDA.ca project and the GenEpiO Consortium (a consortium of 70 researchers from 15 countries interested in contributing to this effort). We hypothesize that a common and accessible ontology entity mart can be developed, if appropriate tools for interfacing domain experts with this mart are developed – and the mart is first applied to practical microbial genomic epidemiology data sharing needs between select public health systems (with consultation involving a larger consortium). In addition, new genomic data visualization approaches are being developed for integration into the IRIDA software platform, to enable more interactive, flexible visualization of genomic data with different levels or views of contextual data (from finely detailed comparisons of genomic islands and other features between genomes, to examining genomic data in the context of geographical data). IRIDA is being used in Canada's public health agency, and this open source software is also being installed in other countries interested in co-developing this resource and using a federated data sharing approach.

AUTHOR INDEX

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

V

W

X

Y

Z