PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018

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.

Abstracts 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

APPLICATIONS OF GENETICS, GENOMICS AND BIOINFORMATICS IN DRUG DISCOVERY

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

CHARACTERIZATION OF DRUG-INDUCED SPLICING COMPLEXITY IN PROSTATE CANCER CELL LINE USING LONG READ TECHNOLOGY

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

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

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

CELL-SPECIFIC PREDICTION AND APPLICATION OF DRUG-INDUCED GENE EXPRESSION PROFILES

Rachel Hodos[1,2], Ping Zhang[3], Hao-Chih Lee[1], Qiaonan Duan[1], Zichen Wang[1], Neil R. Clark[1], Avi Ma'ayan[1], Fei Wang[3,4], Brian Kidd[1], Jianying Hu[3], David Sontag[5], Joel T. Dudley[1]

[1]Icahn School of Medicine at Mount Sinai, [2]New York University, [3]IBM T. J. Watson Research Center, [4]Cornell University, [5]Massachusetts Institute of Technology

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

LARGE-SCALE INTEGRATION OF HETEROGENEOUS PHARMACOGENOMIC DATA FOR IDENTIFYING DRUG MECHANISM OF ACTION

Yunan Luo, Sheng Wang, Jinfeng Xiao, Jian Peng

University of Illinois at Urbana-Champaign

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

CHEMICAL REACTION VECTOR EMBEDDINGS: TOWARDS PREDICTING DRUG METABOLISM IN THE HUMAN GUT MICROBIOME

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

[1]Stanford University, [2]University of California San Francisco, [3]Food and Drug Administration

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

EXTRACTING A BIOLOGICALLY RELEVANT LATENT SPACE FROM CANCER TRANSCRIPTOMES WITH VARIATIONAL AUTOENCODERS

Gregory P. Way, Casey S. Greene

University of Pennsylvania

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

CHALLENGES OF PATTERN RECOGNITION IN BIOMEDICAL DATA
ORAL PRESENTATION

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

LARGE-SCALE ANALYSIS OF DISEASE PATHWAYS IN THE HUMAN INTERACTOME

Monica Agrawal, Marinka Zitnik, Jure Leskovec

Stanford University

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

MAPPING PATIENT TRAJECTORIES USING LONGITUDINAL EXTRACTION AND DEEP LEARNING IN THE MIMIC-III CRITICAL CARE DATABASE

Brett K. Beaulieu-Jones, Patryk Orzechowski, Jason H. Moore

University of Pennsylvania

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

AUTOMATED DISEASE COHORT SELECTION USING WORD EMBEDDINGS FROM ELECTRONIC HEALTH RECORDS

Benjamin S. Glicksberg, Riccardo Miotto, Kipp W. Johnson, Khader Shameer, Li Li, Rong Chen, Joel T. Dudley

Icahn School of Medicine at Mount Sinai

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

FUNCTIONAL NETWORK COMMUNITY DETECTION CAN DISAGGREGATE AND FILTER MULTIPLE UNDERLYING PATHWAYS IN ENRICHMENT ANALYSES

Lia X. Harrington[1], Gregory P. Way[2], Jennifer A. Doherty[3], Casey S. Greene[2]

[1]Geisel School of Medicine at Dartmouth, [2]University of Pennsylvania, [3]University of Utah

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

CAUSAL INFERENCE ON ELECTRONIC HEALTH RECORDS TO ASSESS BLOOD PRESSURE TREATMENT TARGETS: AN APPLICATION OF THE PARAMETRIC G FORMULA

Kipp W. Johnson[1], Benjamin S. Glicksberg[1], Rachel Hodos[1,2], Khader Shameer[1], Joel T. Dudley[1]

[1]Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai; [2]Courant Institute of Mathematical Sciences, New York University

Hypertension is a major risk factor for ischemic cardiovascular disease and cerebrovascular disease, which are respectively the primary and secondary most common causes of morbidity and mortality across the globe. To alleviate the risks of hypertension, there are a number of effective antihypertensive drugs available. However, the optimal treatment blood pressure goal for antihypertensive therapy remains an area of controversy. The results of the recent Systolic Blood Pressure Intervention Trial (SPRINT) trial, which found benefits for intensive lowering of systolic blood pressure, have been debated for several reasons. We aimed to assess the benefits of treating to four different blood pressure targets and to compare our results to those of SPRINT using a method for causal inference called the parametric g formula. We applied this method to blood pressure measurements obtained from the electronic health records of approximately 200,000 patients who visited the Mount Sinai Hospital in New York, NY. We simulated the effect of four clinically relevant dynamic treatment regimes, assessing the effectiveness of treating to four different blood pressure targets: 150 mmHg, 140 mmHg, 130 mmHg, and 120 mmHg. In contrast to current American Heart Association guidelines and in concordance with SPRINT, we find that targeting 120 mmHg systolic blood pressure is significantly associated with decreased incidence of major adverse cardiovascular events. Causal inference methods applied to electronic methods are a powerful and flexible technique and medicine may benefit from their increased usage.

DATA-DRIVEN ADVICE FOR APPLYING MACHINE LEARNING TO BIOINFORMATICS PROBLEMS

Randal S. Olson, William La Cava, Zairah Mustahsan, Akshay Varik, Jason H. Moore

University of Pennsylvania

As the bioinformatics field grows, it must keep pace not only with new data but with new algorithms. Here we contribute a thorough analysis of 13 state-of-the-art, commonly used machine learning algorithms on a set of 165 publicly available classification problems in order to provide data-driven algorithm recommendations to current researchers. We present a number of statistical and visual comparisons of algorithm performance and quantify the effect of model selection and algorithm tuning for each algorithm and dataset. The analysis culminates in the recommendation of five algorithms with hyperparameters that maximize classifier performance across the tested problems, as well as general guidelines for applying machine learning to supervised classification problems.

HOW POWERFUL ARE SUMMARY-BASED METHODS FOR IDENTIFYING EXPRESSION-TRAIT ASSOCIATIONS UNDER DIFFERENT GENETIC ARCHITECTURES?

Yogasudha C. Veturi, Marylyn D. Ritchie

Biomedical and Translational Informatics Institute, Geisinger

Transcriptome-wide association studies (TWAS) have recently been employed as an approach that can draw upon the advantages of genome-wide association studies (GWAS) and gene expression studies to identify genes associated with complex traits. Unlike standard GWAS, summary level data suffices for TWAS and offers improved statistical power. Two popular TWAS methods include either (a) imputing the cis genetic component of gene expression from smaller sized studies (using multi-SNP prediction or MP) into much larger effective sample sizes afforded by GWAS –- TWAS-MP or (b) using summary-based Mendelian randomization –- TWAS-SMR. Although these methods have been effective at detecting functional variants, it remains unclear how extensive variability in the genetic architecture of complex traits and diseases impacts TWAS results. Our goal was to investigate the different scenarios under which these methods yielded enough power to detect significant expression-trait associations. In this study, we conducted extensive simulations based on 6000 randomly chosen, unrelated Caucasian males from Geisinger's MyCode population to compare the power to detect cis expression-trait associations (within 500 kb of a gene) using the above-described approaches. To test TWAS across varying genetic backgrounds we simulated gene expression and phenotype using different quantitative trait loci per gene and cis-expression /trait heritability under genetic models that differentiate the effect of causality from that of pleiotropy. For each gene, on a training set ranging from 100 to 1000 individuals, we either (a) estimated regression coefficients with gene expression as the response using five different methods: LASSO, elastic net, Bayesian LASSO, Bayesian spike-slab, and Bayesian ridge regression or (b) performed eQTL analysis. We then sampled with replacement 50,000, 150,000, and 300,000 individuals respectively from the testing set of the remaining 5000 individuals and conducted GWAS on each set. Subsequently, we integrated the GWAS summary statistics derived from the testing set with the weights (or eQTLs) derived from the training set to identify expression-trait associations using (a) TWAS-MP (b) TWAS-SMR (c) eQTL-based GWAS, or (d) standalone GWAS. Finally, we examined the power to detect functionally relevant genes using the different approaches under the considered simulation scenarios. In general, we observed great similarities among TWAS-MP methods although the Bayesian methods resulted in improved power in comparison to LASSO and elastic net as the trait architecture grew more complex while training sample sizes and expression heritability remained small. Finally, we observed high power under causality but very low to moderate power under pleiotropy.

DEMOCRATIZING HEALTH DATA FOR TRANSLATIONAL RESEARCH

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

CLINGEN CANCER SOMATIC WORKING GROUP – STANDARDIZING AND DEMOCRATIZING ACCESS TO CANCER MOLECULAR DIAGNOSTIC DATA TO DRIVE TRANSLATIONAL RESEARCH

Subha Madhavan[1], Deborah Ritter[2], Christine Micheel[3], Shruti Rao[1], Angshumoy Roy[2], Dmitriy Sonkin[4], Matthew McCoy[1], Malachi Griffith[5], Obi L. Griffith[5], Peter McGarvey[1], Shashikant Kulkarni[2], on behalf of the Clingen Somatic Working Group

[1]Innovation Center for Biomedical Informatics, Georgetown University, Washington D.C.; [2]Baylor College of Medicine and Texas Children's Hospital, Houston, TX; [3]Vanderbilt University School of Medicine, Nashville, TN; [4]National Cancer Institute, Rockville, MD; [5]The McDonnell Genome Institute, Washington University, St. Louis, MO

A growing number of academic and community clinics are conducting genomic testing to inform treatment decisions for cancer patients. In the last 3-5 years, there has been a rapid increase in clinical use of next generation sequencing (NGS) based cancer molecular diagnostic (MolDx) testing. The increasing availability and decreasing cost of tumor genomic profiling means that physicians can now make treatment decisions armed with patient-specific genetic information. Accumulating research in the cancer biology field indicates that there is significant potential to improve cancer patient outcomes by effectively leveraging this rich source of genomic data in treatment planning. To achieve truly personalized medicine in oncology, it is critical to catalog cancer sequence variants from MolDx testing for their clinical relevance along with treatment information and patient outcomes, and to do so in a way that supports large-scale data aggregation and new hypothesis generation. One critical challenge to encoding variant data is adopting a standard of annotation of those variants that are clinically actionable. Through the NIH-funded Clinical Genome Resource (ClinGen), in collaboration with NLM's ClinVar database and >50 academic and industry based cancer research organizations, we developed the Minimal Variant Level Data (MVLD) framework to standardize reporting and interpretation of drug associated alterations. We are currently involved in collaborative efforts to align the MVLD framework with parallel, complementary sequence variants interpretation clinical guidelines from the Association of Molecular Pathologists (AMP) for clinical labs. In order to truly democratize access to MolDx data for care and research needs, these standards must be harmonized to support sharing of clinical cancer variants. Here we describe the processes and methods developed within the ClinGen's Somatic WG in collaboration with over 60 cancer care and research organizations as well as CLIA-certified, CAP-accredited clinical testing labs to develop standards for cancer variant interpretation and sharing. Keywords: ClinGen, Somatic variants, predictive biomarkers, MVLD, data sharing

A HEURISTIC METHOD FOR SIMULATING OPEN-DATA OF ARBITRARY COMPLEXITY THAT CAN BE USED TO COMPARE AND EVALUATE MACHINE LEARNING METHODS

Jason H. Moore, Maksim Shestov, Peter Schmitt, Randal S. Olson

Institute for Biomedical Informatics, University of Pennsylvania

A central challenge of developing and evaluating artificial intelligence and machine learning methods for regression and classification is access to data that illuminates the strengths and weaknesses of different methods. Open data plays an important role in this process by making it easy for computational researchers to easily access real data for this purpose. Genomics has in some examples taken a leading role in the open data effort starting with DNA microarrays. While real data from experimental and observational studies is necessary for developing computational methods it is not sufficient. This is because it is not possible to know what the ground truth is in real data. This must be accompanied by simulated data where that balance between signal and noise is known and can be directly evaluated. Unfortunately, there is a lack of methods and software for simulating data with the kind of complexity found in real biological and biomedical systems. We present here the Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) method and prototype software for simulating complex biological and biomedical data. Further, we introduce new methods for developing simulation models that generate data that specifically allows discrimination between different machine learning methods.

BEST PRACTICES AND LESSONS LEARNED FROM REUSE OF 4 PATIENT-DERIVED METABOLOMICS DATASETS IN ALZHEIMER'S DISEASE

Jessica D. Tenenbaum, Colette Blach

Duke University

The importance of open data has been increasingly recognized in recent years. Although the sharing and reuse of clinical data for translational research lags behind best practices in biological science, a number of patient-derived datasets exist and have been published enabling translational research spanning multiple scales from molecular to organ level, and from patients to populations. In seeking to replicate metabolomic biomarker results in Alzheimer's disease our team identified three independent cohorts in which to compare findings. Accessing the datasets associated with these cohorts, understanding their content and provenance, and comparing variables between studies was a valuable exercise in exploring the principles of open data in practice. It also helped inform steps taken to make the original datasets available for use by other researchers. In this paper we describe best practices and lessons learned in attempting to identify, access, understand, and analyze these additional datasets to advance research reproducibility, as well as steps taken to facilitate sharing of our own data.

IMAGING GENOMICS

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

DISCRIMINATIVE BAG-OF-CELLS FOR IMAGING-GENOMICS

Benjamin Chidester[1], Minh N. Do[2], Jian Ma[1]

[1]Carnegie Mellon University, [2]University of Illinois at Urbana-Champaign

Connecting genotypes to image phenotypes is crucial for a comprehensive understanding of cancer. To learn such connections, new machine learning approaches must be developed for the better integration of imaging and genomic data. Here we propose a novel approach called Discriminative Bag-of-Cells (DBC) for predicting genomic markers using imaging features, which addresses the challenge of summarizing histopathological images by representing cells with learned discriminative types, or codewords. We also developed a reliable and efficient patch-based nuclear segmentation scheme using convolutional neural networks from which nuclear and cellular features are extracted. Applying DBC on TCGA breast cancer samples to predict basal subtype status yielded a class-balanced accuracy of 70% on a separate test partition of 213 patients. As data sets of imaging and genomic data become increasingly available, we believe DBC will be a useful approach for screening histopathological images for genomic markers. Source code of nuclear segmentation and DBC are available at: https://github.com/bchidest/DBC.

DEEP INTEGRATIVE ANALYSIS FOR SURVIVAL PREDICTION

Chenglong Huang[1], Albert Zhang[2], Guanghua Xiao[3]

[1]Colleyville Heritage High School, [2]Highland Park High School, [3]University of Texas Southwestern Medical Center

Survival prediction is very important in medical treatment. However, recent leading research is challenged by two factors: 1) the datasets usually come with multi-modality; and 2) sample sizes are relatively small. To solve the above challenges, we developed a deep survival learning model to predict patients' survival outcomes by integrating multi-view data. The proposed network contains two sub-networks, one view-specific and one common sub-network. We designated one CNN-based and one FCN-based sub-network to efficiently handle pathological images and molecular profiles, respectively. Our model first explicitly maximizes the correlation among the views and then transfers feature hierarchies from view commonality and specifically fine-tunes on the survival prediction task. We evaluate our method on real lung and brain tumor data sets to demonstrate the effectiveness of the proposed model using data with multiple modalities across different tumor types.

GENOTYPE-PHENOTYPE ASSOCIATION STUDY VIA NEW MULTI-TASK LEARNING MODEL

Zhouyuan Huo[1], Dinggang Shen[2], Heng Huang[1]

[1]University of Pittsburgh, [2]University of North Carolina at Chapel Hill

Research on the associations between genetic variations and imaging phenotypes is developing with the advance in high-throughput genotype and brain image techniques. Regression analysis of single nucleotide polymorphisms (SNPs) and imaging measures as quantitative traits (QTs) has been proposed to identify the quantitative trait loci (QTL) via multi-task learning models. Recent studies consider the interlinked structures within SNPs and imaging QTs through group lasso, e.g. ℓ21-norm, leading to better predictive results and insights of SNPs. However, group sparsity is not enough for representing the correlation between multiple tasks and ℓ21-norm regularization is not robust either. In this paper, we propose a new multi-task learning model to analyze the associations between SNPs and QTs. We suppose that low-rank structure is also beneficial to uncover the correlation between genetic variations and imaging phenotypes. Finally, we conduct regression analysis of SNPs and QTs. Experimental results show that our model is more accurate in prediction than compared methods and presents new insights of SNPs.

CODON BIAS AMONG SYNONYMOUS RARE VARIANTS IS ASSOCIATED WITH ALZHEIMER'S DISEASE IMAGING BIOMARKER

Jason E. Miller[1], Manu K. Shivakumar[2], Shannon L. Risacher[2], Andrew J. Saykin[2], Seunggeun Lee[3], Kwangsik Nho[2], Dokyoon Kim[1,4]

[1]Geisinger Health System, [2]Indiana University School of Medicine, [3]University of Michigan, [4]Pennsylvania State University

Alzheimer's disease (AD) is a neurodegenerative disorder with few biomarkers even though it impacts a relatively large portion of the population and is predicted to affect significantly more individuals in the future. Neuroimaging has been used in concert with genetic information to improve our understanding in relation to how AD arises and how it can be potentially diagnosed. Additionally, evidence suggests synonymous variants can have a functional impact on gene regulatory mechanisms, including those related to AD. Some synonymous codons are preferred over others leading to a codon bias. The bias can arise with respect to codons that are more or less frequently used in the genome. A bias can also result from optimal and non-optimal codons, which have stronger and weaker codon anti-codon interactions, respectively. Although association tests have been utilized before to identify genes associated with AD, it remains unclear how codon bias plays a role and if it can improve rare variant analysis. In this work, rare variants from whole-genome sequencing from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort were binned into genes using BioBin. An association analysis of the genes with AD-related neuroimaging biomarker was performed using SKAT-O. While using all synonymous variants we did not identify any genome- wide significant associations, using only synonymous variants that affected codon frequency we identified several genes as significantly associated with the imaging phenotype. Additionally, significant associations were found using only rare variants that contains an optimal codon in among minor alleles and a non-optimal codon in the major allele. These results suggest that codon bias may play a role in AD and that it can be used to improve detection power in rare variant association analysis.

PRECISION MEDICINE: FROM DIPLOTYPES TO DISPARITIES TOWARDS IMPROVED HEALTH AND THERAPIES

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

SINGLE SUBJECT TRANSCRIPTOME ANALYSIS REPRODUCES SIGNED GENE SET FUNCTIONAL ACTIVATION SIGNALS FROM COHORT ANALYSIS OF MURINE RESPONSE TO HIGH FAT DIET

Joanne Berghout, Qike Li, Nima Pouladi, Jianrong Li, Yves A. Lussier

University of Arizona

Analysis of single-subject transcriptome response data is an unmet need of precision medicine, made challenging by the high dimension, dynamic nature and difficulty in extracting meaningful signals from biological or stochastic noise. We have proposed a method for single subject analysis that uses a mixture model for transcript fold-change clustering from isogenic paired samples, followed by integration of these distributions with Gene Ontology Biological Processes (GO-BP) to reduce dimension and identify functional attributes. We then extended these methods to develop functional signing metrics for gene set process regulation by incorporating biological repressor relationships encoded in GO as negatively_regulates edges. Results revealed reproducible and biologically meaningful signals from analysis of a single subject's response, opening the door to future transcriptomic studies where subject and resource availability are currently limiting. We used inbred mouse strains fed different diets to provide isogenic biological replicates, permitting rigorous validation of our method. We compared significant genotype-specific GO-BP term results for overlap and rank order across three replicates per genotype, and cross-methods to reference standards (limma+FET, SAM+FET, and GSEA). All single-subject analytics findings were robust and highly reproducible (median area under the ROC curve=0.96, n=24 genotypes x 3 replicates), providing confidence and validation of this approach for analyses in single subjects. R code is available online at http://www.lussiergroup.org/publications/PathwayActivity

USING SIMULATION AND OPTIMIZATION APPROACH TO IMPROVE OUTCOME THROUGH WARFARIN PRECISION TREATMENT

Chih-Lin Chi[1], Lu He[2], Kourosh Ravvaz[3], John Weissert[3], Peter J. Tonellato[4,5]

[1]School of Nursing & Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA; [2]Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA; [3]Aurora Health Care, Milwaukee, WI, USA; [4]Department of Biomedical Informatics, Department of Pathology, Harvard Medical School, Boston, MA, USA; [5]Zilber School of Public HealthUniversity of Wisconsin-Milwaukee, Milwaukee, WI, USA

We apply a treatment simulation and optimization approach to develop decision support guidance for warfarin precision treatment plans. Simulation include the use of ~1,500,000 clinical avatars (simulated patients) generated by an integrated data-driven and domain-knowledge based Bayesian Network Modeling approach. Subsequently, we simulate 30-day individual patient response to warfarin treatment of five clinical and genetic treatment plans followed by both individual and sub-population based optimization. Sub-population optimization (compared to individual optimization) provides a cost effective and realistic means of implementation of a precision-driven treatment plan in practical settings. In this project, we use the property of minimal entropy to minimize overall adverse risks for the largest possible patient sub-populations and we temper the results by considering both transparency and ease of implementation. Finally, we discuss the improved outcome of the precision treatment plan based on the sub-population optimized decision support rules.

COALITIONAL GAME THEORY AS A PROMISING APPROACH TO IDENTIFY CANDIDATE AUTISM GENES

Anika Gupta, Min Woo Sun, Kelley M. Paskov, Nate T. Stockham, Jae-Yoon Jung, Dennis P. Wall

Stanford University

Despite mounting evidence for the strong role of genetics in the phenotypic manifestation of Autism Spectrum Disorder (ASD), the specific genes responsible for the variable forms of ASD remain undefined. ASD may be best explained by a combinatorial genetic model with varying epistatic interactions across many small effect mutations. Coalitional or cooperative game theory is a technique that studies the combined effects of groups of players, known as coalitions, seeking to identify players who tend to improve the performance--the relationship to a specific disease phenotype--of any coalition they join. This method has been previously shown to boost biologically informative signal in gene expression data but to-date has not been applied to the search for cooperative mutations among putative ASD genes. We describe our approach to highlight genes relevant to ASD using coalitional game theory on alteration data of 1,965 fully sequenced genomes from 756 multiplex families. Alterations were encoded into binary matrices for ASD (case) and unaffected (control) samples, indicating likely gene-disrupting, inherited mutations in altered genes. To determine individual gene contributions given an ASD phenotype, a “player” metric, referred to as the Shapley value, was calculated for each gene in the case and control cohorts. Sixty seven genes were found to have significantly elevated player scores and likely represent significant contributors to the genetic coordination underlying ASD. Using network and cross-study analysis, we found that these genes are involved in biological pathways known to be affected in the autism cases and that a subset directly interact with several genes known to have strong associations to autism. These findings suggest that coalitional game theory can be applied to large-scale genomic data to identify hidden yet influential players in complex polygenic disorders such as autism.

CONSIDERATIONS FOR AUTOMATED MACHINE LEARNING IN CLINICAL METABOLIC PROFILING: ALTERED HOMOCYSTEINE PLASMA CONCENTRATION ASSOCIATED WITH METFORMIN EXPOSURE

Alena Orlenko[1], Jason H. Moore[1], Patryk Orzechowski[1], Randal S. Olson[1], Junmei Cairns[2], Pedro J. Caraballo[2], Richard M. Weinshilboum[2], Liewei Wang[2], Matthew K. Breitenstein[1]

[1]University of Pennsylvania, [2]Mayo Clinic

With the maturation of metabolomics science and proliferation of biobanks, clinical metabolic profiling is an increasingly opportunistic frontier for advancing translational clinical research. Automated Machine Learning (AutoML) approaches provide exciting opportunity to guide feature selection in agnostic metabolic profiling endeavors, where potentially thousands of independent data points must be evaluated. In previous research, AutoML using high-dimensional data of varying types has been demonstrably robust, outperforming traditional approaches. However, considerations for application in clinical metabolic profiling remain to be evaluated. Particularly, regarding the robustness of AutoML to identify and adjust for common clinical confounders. In this study, we present a focused case study regarding AutoML considerations for using the Tree-Based Optimization Tool (TPOT) in metabolic profiling of exposure to metformin in a biobank cohort. First, we propose a tandem rank-accuracy measure to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors. Second, while AutoML, using default parameters, demonstrated potential to lack sensitivity to low-effect confounding clinical covariates, we demonstrated residual training and adjustment of metabolite features as an easily applicable approach to ensure AutoML adjustment for potential confounding characteristics. Finally, we present increased homocysteine with long-term exposure to metformin as a potentially novel, non-replicated metabolite association suggested by TPOT; an association not identified in parallel clinical metabolic profiling endeavors. While warranting independent replication, our tandem rank-accuracy measure suggests homocysteine to be the metabolite feature with largest effect, and corresponding priority for further translational clinical research. Residual training and adjustment for a potential confounding effect by BMI only slightly modified the suggested association. Increased homocysteine is thought to be associated with vitamin B12 deficiency – evaluation for potential clinical relevance is suggested. While considerations for clinical metabolic profiling are recommended, including adjustment approaches for clinical confounders, AutoML presents an exciting tool to enhance clinical metabolic profiling and advance translational research endeavors.

ADDRESSING VITAL SIGN ALARM FATIGUE USING PERSONALIZED ALARM THRESHOLDS

Sarah Poole, Nigam Shah

Stanford University

Alarm fatigue, a condition in which clinical staff become desensitized to alarms due to the high frequency of unnecessary alarms, is a major patient safety concern. Alarm fatigue is particularly prevalent in the pediatric setting, due to the high level of variation in vital signs with patient age. Existing studies have shown that the current default pediatric vital sign alarm thresholds are inappropriate, and lead to a larger than necessary alarm load. This study leverages a large database containing over 190 patient-years of heart rate data to accurately identify the 1st and 99th percentiles of an individual's heart rate on their first day of vital sign monitoring. These percentiles are then used as personalized vital sign thresholds, which are evaluated by comparing to non-default alarm thresholds used in practice, and by using the presence of major clinical events to infer alarm labels. Using the proposed personalized thresholds would decrease low and high heart rate alarms by up to 50% and 44% respectively, while maintaining sensitivity of 62% and increasing specificity to 49%. The proposed personalized vital sign alarm thresholds will reduce alarm fatigue, thus contributing to improved patient outcomes, shorter hospital stays, and reduced hospital costs.

EMERGENCE OF PATHWAY-LEVEL COMPOSITE BIOMARKERS FROM CONVERGING GENE SET SIGNALS OF HETEROGENEOUS TRANSCRIPTOMIC RESPONSES

Samir Rachid Zaim, Qike Li, A. Grant Schissler, Yves A. Lussier

The University of Arizona

Recent precision medicine initiatives have led to the expectation of improved clinical decision-making anchored in genomic data science. However, over the last decade, only a handful of new single-gene product biomarkers have been translated to clinical practice (FDA approved) in spite of considerable discovery efforts deployed and a plethora of transcriptomes available in the Gene Expression Omnibus. With this modest outcome of current approaches in mind, we developed a pilot simulation study to demonstrate the untapped benefits of developing disease detection methods for cases where the true signal lies at the pathway level, even if the pathway's gene expression alterations may be heterogeneous across patients. In other words, we relaxed the cross-patient homogeneity assumption from the transcript level (cohort assumptions of deregulated gene expression) to the pathway level (assumptions of deregulated pathway expression). Furthermore, we have expanded previous single-subject (SS) methods into cohort analyses to illustrate the benefit of accounting for an individual's variability in cohort scenarios. We compare SS and cohort-based (CB) techniques under 54 distinct scenarios, each with 1,000 simulations, to demonstrate that the emergence of a pathway-level signal occurs through the summative effect of its altered gene expression, heterogeneous across patients. Studied variables include pathway gene set size, fraction of expressed gene responsive within gene set, fraction of expressed gene responsive up- vs down-regulated, and cohort size. We demonstrated that our SS approach was uniquely suited to detect signals in heterogeneous populations in which individuals have varying levels of baseline risks that are simultaneously confounded by patient-specific "genome -by- environment" interactions (G×E). Area under the precision-recall curve of the SS approach far surpassed that of the CB (1st quartile, median, 3rd quartile: SS = 0.94, 0.96, 0.99; CB= 0.50, 0.52, 0.65). We conclude that single-subject pathway detection methods are uniquely suited for consistently detecting pathway dysregulation by the inclusion of a patient's individual variability.
http://www.lussiergroup.org/publications/PathwayMarker/

ANALYZING METABOLOMICS DATA FOR ASSOCIATION WITH GENOTYPES USING TWO-COMPONENT GAUSSIAN MIXTURE DISTRIBUTIONS

Jason Westra[1,2], Nicholas Hartman[3], Bethany Lake[4], Gregory Shearer[5], Nathan Tintle[1]

[1]Dordt College, [2]Iowa State University, [3]Cornell University, [4]Elon University, [5]The Pennsylvania State University

Standard approaches to evaluate the impact of single nucleotide polymorphisms (SNP) on quantitative phenotypes use linear models. However, these normal-based approaches may not optimally model phenotypes which are better represented by Gaussian mixture distributions (e.g., some metabolomics data). We develop a likelihood ratio test on the mixing proportions of two-component Gaussian mixture distributions and consider more restrictive models to increase power in light of a priori biological knowledge. Data was simulated to validate the improved power of the likelihood ratio test and the restricted likelihood ratio test over a linear model and a log transformed linear model. Then, using real data from the Framingham Heart Study, we analyzed 20,315 SNPs on chromosome 11, demonstrating that the proposed likelihood ratio test identifies SNPs well known to participate in the desaturation of certain fatty acids. Our study both validates the approach of increasing power by using the likelihood ratio test that leverages Gaussian mixture models, and creates a model with improved sensitivity and interpretability.

READING BETWEEN THE GENES: COMPUTATIONAL MODELS TO DISCOVER FUNCTION AND/OR CLINICAL UTILITY FROM NONCODING DNA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

CONVERGENT DOWNSTREAM CANDIDATE MECHANISMS OF INDEPENDENT INTERGENIC POLYMORPHISMS BETWEEN CO-CLASSIFIED DISEASES IMPLICATE EPISTASIS AMONG NONCODING ELEMENTS

Jiali Han[1], Jianrong Li[1], Ikbel Achour[1], Lorenzo Pesce[2], Ian Foster[2], Haiquan Li[3], Yves A. Lussier[3]

[1]Center for Biomedical Informatics and Biostatistics (CB2) and Departments of Medicine and of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA; [2]Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, IL 60637, USA; [3]CB2, BIO5 Institute, UACC, and Dept of Medicine, The University of Arizona, Tucson, AZ 85721, USA

Eighty percent of DNA outside protein coding regions was shown biochemically functional by the ENCODE project, enabling studies of their interactions. Studies have since explored how convergent downstream mechanisms arise from independent genetic risks of one complex disease. However, the cross-talk and epistasis between intergenic risks associated with distinct complex diseases have not been comprehensively characterized. Our recent integrative genomic analysis unveiled downstream biological effectors of disease-specific polymorphisms buried in intergenic regions, and we then validated their genetic synergy and antagonism in distinct GWAS. We extend this approach to characterize convergent downstream candidate mechanisms of distinct intergenic SNPs across distinct diseases within the same clinical classification. We construct a multipartite network consisting of 467 diseases organized in 15 classes, 2,358 disease-associated SNPs, 6,301 SNP-associated mRNAs by eQTL, and mRNA annotations to 4,538 Gene Ontology mechanisms. Functional similarity between two SNPs (similar SNP pairs) is imputed using a nested information theoretic distance model for which p-values are assigned by conservative scale-free permutation of network edges without replacement (node degrees constant). At FDR≤5%, we prioritized 3,870 intergenic SNP pairs associated, among which 755 are associated with distinct diseases sharing the same disease class, implicating 167 intergenic SNPs, 14 classes, 230 mRNAs, and 134 GO terms. Co-classified SNP pairs were more likely to be prioritized as compared to those of distinct classes confirming a noncoding genetic underpinning to clinical classification (odds ratio ~3.8; p≤10E-25). The prioritized pairs were also enriched in regions bound to the same/interacting transcription factors and/or interacting in long-range chromatin interactions suggestive of epistasis (odds ratio ~ 2,500; p≤10E-25). This prioritized network implicates complex epistasis between intergenic polymorphisms of co-classified diseases and offers a roadmap for a novel therapeutic paradigm: repositioning medications that target proteins within downstream mechanisms of intergenic disease-associated SNPs. Supplementary information and software: http://lussiergroup.org/publications/disease_class

NETWORK ANALYSIS OF PSEUDOGENE-GENE RELATIONSHIPS: FROM PSEUDOGENE EVOLUTION TO THEIR FUNCTIONAL POTENTIALS

Travis S. Johnson[1], Sihong Li[1], Johnathan R. Kho[2], Kun Huang[3], Yan Zhang[1]

[1]Ohio State University, [2]Georgia Institute of Technology, [3]Indiana University

Pseudogenes are fossil relatives of genes. Pseudogenes have long been thought of as "junk DNAs", since they do not code proteins in normal tissues. Although most of the human pseudogenes do not have noticeable functions, ~20% of them exhibit transcriptional activity. There has been evidence showing that some pseudogenes adopted functions as lncRNAs and work as regulators of gene expression. Furthermore, pseudogenes can even be "reactivated" in some conditions, such as cancer initiation. Some pseudogenes are transcribed in specific cancer types, and some are even translated into proteins as observed in several cancer cell lines. All the above have shown that pseudogenes could have functional roles or potentials in the genome. Evaluating the relationships between pseudogenes and their gene counterparts could help us reveal the evolutionary path of pseudogenes and associate pseudogenes with functional potentials. It also provides an insight into the regulatory networks involving pseudogenes with transcriptional and even translational activities. In this study, we develop a novel approach integrating graph analysis, sequence alignment and functional analysis to evaluate pseudogene-gene relationships, and apply it to human gene homologs and pseudogenes. We generated a comprehensive set of 445 pseudogene-gene (PGG) families from the original 3,281 gene families (13.56%). Of these 438 (98.4% PGG, 13.3% total) were non-trivial (containing more than one pseudogene). Each PGG family contains multiple genes and pseudogenes with high sequence similarity. For each family, we generate a sequence alignment network and phylogenetic trees recapitulating the evolutionary paths. We find evidence supporting the evolution history of olfactory family (both genes and pseudogenes) in human, which also supports the validity of our analysis method. Next, we evaluate these networks in respect to the gene ontology from which we identify functions enriched in these pseudogene-gene families and infer functional impact of pseudogenes involved in the networks. This demonstrates the application of our PGG network database in the study of pseudogene function in disease context.

LEVERAGING PUTATIVE ENHANCER-PROMOTER INTERACTIONS TO INVESTIGATE TWO-WAY EPISTASIS IN TYPE 2 DIABETES GWAS

Elisabetta Manduchi[1,2], Alessandra Chesi[2], Molly A. Hall[1], Struan F. A. Grant[2], Jason H. Moore[1]

[1]University of Pennsylvania, [2]The Children's Hospital of Philadelphia

We utilized evidence for enhancer-promoter interactions from functional genomics data in order to build biological filters to narrow down the search space for two-way Single Nucleotide Polymorphism (SNP) interactions in Type 2 Diabetes (T2D) Genome Wide Association Studies (GWAS). This has led us to the identification of a reproducible statistically significant SNP pair associated with T2D. As more functional genomics data are being generated that can help identify potentially interacting enhancer-promoter pairs in larger collection of tissues/cells, this approach has implications for investigation of epistasis from GWAS in general.

TEXT MINING AND VISUALIZATION FOR PRECISION MEDICINE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

IMPROVING PRECISION IN CONCEPT NORMALIZATION

Mayla Boguslav, K. Bretonnel Cohen, William A. Baumgartner Jr., Lawrence E. Hunter

Computational Bioscience Program, University of Colorado School of Medicine

Most natural language processing applications exhibit a trade-off between precision and recall. In some use cases for natural language processing, there are reasons to prefer to tilt that trade-off toward high precision. Relying on the Zipfian distribution of false positive results, we describe a strategy for increasing precision, using a variety of both pre-processing and post-processing methods. They draw on both knowledge-based and frequentist approaches to modeling language. Based on an existing high-performance biomedical concept recognition pipeline and a previously published manually annotated corpus, we apply this hybrid rationalist/empiricist strategy to concept normalization for eight different ontologies. Which approaches did and did not improve precision varied widely between the ontologies.

VISAGE: INTEGRATING EXTERNAL KNOWLEDGE INTO ELECTRONIC MEDICAL RECORD VISUALIZATION

Edward W. Huang, Sheng Wang, ChengXiang Zhai

University of Illinois at Urbana-Champaign

In this paper, we present VisAGE, a method that visualizes electronic medical records (EMRs) in a low-dimensional space. Effective visualization of new patients allows doctors to view similar, previously treated patients and to identify the new patients' disease subtypes, reducing the chance of misdiagnosis. However, EMRs are typically incomplete or fragmented, resulting in patients who are missing many available features being placed near unrelated patients in the visualized space. VisAGE integrates several external data sources to enrich EMR databases to solve this issue. We evaluated VisAGE on a dataset of Parkinson's disease patients. We qualitatively and quantitatively show that VisAGE can more effectively cluster patients, which allows doctors to better discover patient subtypes and thus improve patient care.

ANNOTATING GENE SETS BY MINING LARGE LITERATURE COLLECTIONS WITH PROTEIN NETWORKS

Sheng Wang[1], Jianzhu Ma[2], Michael Ku Yu[2], Fan Zheng[2], Edward W. Huang[1], Jiawei Han[1], Jian Peng[1], Trey Ideker[2]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA, [2]School of Medicine, University of California San Diego, San Diego, CA, USA

Analysis of patient genomes and transcriptomes routinely recognizes new gene sets associated with human disease. Here we present an integrative natural language processing system which infers common functions for a gene set through automatic mining of the scientific literature with biological networks. This system links genes with associated literature phrases and combines these links with protein interactions in a single heterogeneous network. Multiscale functional annotations are inferred based on network distances between phrases and genes and then visualized as an ontology of biological concepts. To evaluate this system, we predict functions for gene sets representing known pathways and find that our approach achieves substantial improvement over the conventional text-mining baseline method. Moreover, our system discovers novel annotations for gene sets or pathways without previously known functions. Two case studies demonstrate how the system is used in discovery of new cancer-related pathways with ontological annotations.

APPLICATIONS OF GENETICS, GENOMICS AND BIOINFORMATICS IN DRUG DISCOVERY

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

PREDICTION OF PROTEIN-LIGAND INTERACTIONS FROM PAIRED PROTEIN SEQUENCE MOTIFS AND LIGAND SUBSTRUCTURES

Peyton Greenside, Maureen Hillenmeyer, Anshul Kundaje

Stanford University

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

LOSS-OF-FUNCTION OF NEUROPLASTICITY-RELATED GENES CONFERS RISK FOR HUMAN NEURODEVELOPMENTAL DISORDERS

Milo R. Smith, Benjamin S. Glicksberg, Li Li, Rong Chen, Hirofumi Morishita, Joel T. Dudley

Icahn School of Medicine at Mount Sinai

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

DIFFUSION MAPPING OF DRUG TARGETS ON DISEASE SIGNALING NETWORK ELEMENTS REVEALS DRUG COMBINATION STRATEGIES

Jielin Xu[1], Kelly Regan[1], Siyuan Deng[1], William E. Carson III[2], Philip R.O. Payne[3], Fuhai Li[1]

[1]Deptartment of Biomedical Informatics, The Ohio State University; [2]Comprehensive Cancer Center, The Ohio State University; [3]Institute for Informatics, Washington University in St. Louis

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

CHALLENGES OF PATTERN RECOGNITION IN BIOMEDICAL DATA

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

OWL-NETS: TRANSFORMING OWL REPRESENTATIONS FOR IMPROVED NETWORK INFERENCE

Tiffany J. Callahan[1], William A. Baumgartner Jr.[1], Michael Bada[1], Adrianne L. Stefanski[1], Ignacio Tripodi[2], Elizabeth K. White[1], Lawrence E. Hunter[1]

[1]University of Colorado Denver Anschutz Medical Campus, [2]University of Colorado Boulder

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

AN ULTRA-FAST AND SCALABLE QUANTIFICATION PIPELINE FOR TRANSPOSABLE ELEMENTS FROM NEXT GENERATION SEQUENCING DATA

Hyun-Hwan Jeong, Hari Krishna Yalamanchili, Caiwei Guo, Joshua, M. Shulman, Zhandong Liu

College of Medicine, Jan and Dan Duncan Neurological Research Institute

Transposable elements (TEs) are DNA sequences which are capable of moving from one location to another and represent a large proportion (45%) of the human genome. TEs have functional roles in a variety of biological phenomena such as cancer, neurodegenerative disease, and aging. Rapid development in RNA-sequencing technology has enabled us, for the first time, to study the activity of TE at the systems level. However, efficient TE analysis tools are not yet developed. In this work, we developed SalmonTE, a fast and reliable pipeline for the quantification of TEs from RNA-seq data. We benchmarked our tool against TEtranscripts, a widely used TE quantification method, and three other quantification methods using several RNA-seq datasets from Drosophila melanogaster and human cell-line. We achieved 20 times faster execution speed without compromising the accuracy. This pipeline will enable the biomedical research community to quantify and analyze TEs from large amounts of data and lead to novel TE centric discoveries.

IMPROVING THE EXPLAINABILITY OF RANDOM FOREST CLASSIFIER – USER CENTERED APPROACH

Dragutin Petkovic[1,3], Russ B. Altman[2], Mike Wong[3], Arthur Vigil[4]

[1]Computer Science Department, San Francisco State University (SFSU), 1600 Holloway Ave., San Francisco CA 94132, Petkovic@sfsu.edu; [2]Department of Bioengineering, Stanford University, 443 Via Ortega Drive, Stanford, CA 94305-4145; [3]SFSU Center for Computing for Life Sciences, 1600 Holloway Ave., San Francisco, CA 94132; [4]Twist Bioscience, 455 Mission Bay Boulevard South, San Francisco, CA 94158

Machine Learning (ML) methods are now influencing major decisions about patient care, new medical methods, drug development and their use and importance are rapidly increasing in all areas. However, these ML methods are inherently complex and often difficult to understand and explain resulting in barriers to their adoption and validation. Our work (RFEX) focuses on enhancing Random Forest (RF) classifier explainability by developing easy to interpret explainability summary reports from trained RF classifiers as a way to improve the explainability for (often non-expert) users. RFEX is implemented and extensively tested on Stanford FEATURE data where RF is tasked with predicting functional sites in 3D molecules based on their electrochemical signatures (features). In developing RFEX method we apply user-centered approach driven by explainability questions and requirements collected by discussions with interested practitioners. We performed formal usability testing with 13 expert and non-expert users to verify RFEX usefulness. Analysis of RFEX explainability report and user feedback indicates its usefulness in significantly increasing explainability and user confidence in RF classification on FEATURE data. Notably, RFEX summary reports easily reveal that one needs very few (from 2-6 depending on a model) top ranked features to achieve 90% or better of the accuracy when all 480 features are used. Keywords: Random Forest, Explainability, Interpretability, Stanford FEATURE

TREE-BASED METHODS FOR CHARACTERIZING TUMOR DENSITY HETEROGENEITY

Katherine Shoemaker[1], Brian P. Hobbs[2], Karthik Bharath[3], Chaan S. Ng[2], Veerabhadran Baladandayuthapani[2]

[1]Rice University, [2]MD Anderson Cancer Center, [3]University of Nottingham

Solid lesions emerge within diverse tissue environments making their characterization and diagnosis a challenge. With the advent of cancer radiomics, a variety of techniques have been developed to transform images into quantifiable feature sets producing summary statistics that describe the morphology and texture of solid masses. Relying on empirical distribution summaries as well as grey-level co-occurrence statistics, several approaches have been devised to characterize tissue density heterogeneity. This article proposes a novel decision-tree based approach which quantifies the tissue density heterogeneity of a given lesion through its resultant distribution of tree-structured dissimilarity metrics computed with least common ancestor trees under repeated pixel re-sampling. The methodology, based on statistics derived from Galton-Watson trees, produces metrics that are minimally correlated with existing features, adding new information to the feature space and improving quantitative characterization of the extent to which a CT image conveys heterogeneous density distribution. We demonstrate its practical application through a diagnostic study of adrenal lesions. Integrating the proposed with existing features identifies classifiers of three important lesion types; malignant from benign (AUC = 0.78), functioning from non-functioning (AUC = 0.93) and calcified from non-calcified (AUC of 1).

DEMOCRATIZING HEALTH DATA FOR TRANSLATIONAL RESEARCH

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

IDENTIFYING NATURAL HEALTH PRODUCT AND DIETARY SUPPLEMENT INFORMATION WITHIN ADVERSE EVENT REPORTING SYSTEMS

Vivekanand Sharma, Indra Neil Sarkar

Center for Biomedical Informatics, Brown University

Data on safety and efficacy issues associated with natural health products and dietary supplements (NHP&S) remains largely cloistered within domain specific databases or embedded within general biomedical data sources. A major challenge in leveraging analytic approaches on such data is due to the inefficient ability to retrieve relevant data, which includes a general lack of interoperability among related sources. This study developed a thesaurus of NHP&S ingredient terms that can be used by existing biomedical natural language processing (NLP) tools for extracting information of interest. This process was evaluated relative to intervention name strings sampled from the United States Food and Drug Administration Adverse Event Reporting System (FAERS). A use case was used to demonstrate the potential to utilize FAERS for monitoring NHP&S adverse events. The results from this study provide insights on approaches for identifying additional knowledge from extant repositories of knowledge, and potentially as information that can be included into larger curation efforts.

DEMOCRATIZING DATA SCIENCE THROUGH DATA SCIENCE TRAINING

John Darrell Van Horn[1], Lily Fierro[2], Jeana Kamdar[1], Jonathan Gordon[2], Crystal Stewart[1], Avnish Bhattrai[1], Sumiko Abe[1], Xiaoxiao Lei[1], Caroline O'Driscoll[1], Aakanchha Sinha[2], Priyambada Jain[2], Gully Burns[2], Kristina Lerman[2], José Luis Ambite[2]

[1]USC Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, 2025 Zonal Avenue, SHN, Los Angeles, CA 90033, Phone: 323-442-7246; [2]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

The biomedical sciences have experienced an explosion of data which promises to overwhelm many current practitioners. Without easy access to data science training resources, biomedical researchers may find themselves unable to wrangle their own datasets. In 2014, to address the challenges posed such a data onslaught, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative. To this end, the BD2K Training Coordinating Center (TCC; bigdatau.org) was funded to facilitate both in-person and online learning, and open up the concepts of data science to the widest possible audience. Here, we describe the activities of the BD2K TCC and its focus on the construction of the Educational Resource Discovery Index (ERuDIte), which identifies, collects, describes, and organizes online data science materials from BD2K awardees, open online courses, and videos from scientific lectures and tutorials. ERuDIte now indexes over 9,500 resources. Given the richness of online training materials and the constant evolution of biomedical data science, computational methods applying information retrieval, natural language processing, and machine learning techniques are required - in effect, using data science to inform training in data science. In so doing, the TCC seeks to democratize novel insights and discoveries brought forth via large-scale data science training.

IMAGING GENOMICS

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

HERITABILITY ESTIMATES ON RESTING STATE FMRI DATA USING THE ENIGMA ANALYSIS PIPELINE

Bhim M. Adhikari[1], Neda Jahanshad[2], Dinesh Shukla[1], David C. Glahn[3], John Blangero[4], Richard C. Reynolds[5], Robert W. Cox[5], Els Fieremans[6], Jelle Veraart[6], Dmitry S. Novikov[6], Thomas E. Nichols[7], L. Elliot Hong[1], Paul M. Thompson[2], Peter Kochunov[1]

[1]Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA; [2]Imaging Genetics Center, Stevens Institute for Neuroimaging & Informatics, Keck School of Medicine of USC, Marina del Rey, CA, USA; [3]Department of Psychiatry, Yale University, School of Medicine, New Haven, CT, USA; [4]Genomics Computing Center, University of Texas at Rio Grande Valley, USA; [5]National Institute of Mental Health, Bethesda, MD, USA; [6]Center for Biomedical Imaging, Department of Radiology, New York University School of Medicine, NY, USA; [7]Department of Statistics, University of Warwick, Coventry, CV47AL, UK

Big data initiatives such as the Enhancing NeuroImaging Genetics through Meta-Analysis consortium (ENIGMA), combine data collected by independent studies worldwide to achieve more generalizable estimates of effect sizes and more reliable and reproducible outcomes. Such efforts require harmonized image analyses protocols to extract phenotypes consistently. This harmonization is particularly challenging for resting state fMRI due to the wide variability of acquisition protocols and scanner platforms; this leads to site-to-site variance in quality, resolution and temporal signal-to-noise ratio (tSNR). An effective harmonization should provide optimal measures for data of different qualities. We developed a multi-site rsfMRI analysis pipeline to allow research groups around the world to process rsfMRI scans in a harmonized way, to extract consistent and quantitative measurements of connectivity and to perform coordinated statistical tests. We used the single-modality ENIGMA rsfMRI preprocessing pipeline based on model-free Marchenko-Pastur PCA based denoising to verify and replicate resting state network heritability estimates. We analyzed two independent cohorts, GOBS (Genetics of Brain Structure) and HCP (the Human Connectome Project), which collected data using conventional and connectomics oriented fMRI protocols, respectively. We used seed-based connectivity and dual-regression approaches to show that the rsfMRI signal is consistently heritable across twenty major functional network measures. Heritability values of 20-40% were observed across both cohorts.

MRI TO MGMT: PREDICTING METHYLATION STATUS IN GLIOBLASTOMA PATIENTS USING CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Lichy Han, Maulik R. Kamdar

Program in Biomedical Informatics, Stanford University School of Medicine

Glioblastoma Multiforme (GBM), a malignant brain tumor, is among the most lethal of all cancers. Temozolomide is the primary chemotherapy treatment for patients diagnosed with GBM. The methylation status of the promoter or the enhancer regions of the O6-methylguanine methyltransferase (MGMT) gene may impact the efficacy and sensitivity of temozolomide, and hence may affect overall patient survival. Microscopic genetic changes may manifest as macroscopic morphological changes in the brain tumors that can be detected using magnetic resonance imaging (MRI), which can serve as noninvasive biomarkers for determining methylation of MGMT regulatory regions. In this research, we use a compendium of brain MRI scans of GBM patients collected from The Cancer Imaging Archive (TCIA) combined with methylation data from The Cancer Genome Atlas (TCGA) to predict the methylation state of the MGMT regulatory regions in these patients. Our approach relies on a bi-directional convolutional recurrent neural network architecture (CRNN) that leverages the spatial aspects of these 3-dimensional MRI scans. Our CRNN obtains an accuracy of 67% on the validation data and 62% on the test data, with precision and recall both at 67%, suggesting the existence of MRI features that may complement existing markers for GBM patient stratification and prognosis. We have additionally presented our model via a novel neural network visualization platform, which we have developed to improve interpretability of deep learning MRI-based classification models.

BUILDING TRANS-OMICS EVIDENCE: USING IMAGING AND 'OMICS' TO CHARACTERIZE CANCER PROFILES

Arunima Srivastava[1], Chaitanya Kulkarni[1], Parag Mallick[2], Kun Huang[3], Raghu Machiraju[1]

[1]The Ohio State University, [2]Stanford University, [3]Indiana University School of Medicine

Utilization of single modality data to build predictive models in cancer results in a rather narrow view of most patient profiles. Some clinical facets relate strongly to histology image features, e.g. tumor stages, whereas others are associated with genomic and proteomic variations (e.g. cancer subtypes and disease aggression biomarkers). We hypothesize that there are coherent "trans-omics" features that characterize varied clinical cohorts across multiple sources of data leading to more descriptive and robust disease characterization. In this work, for 105 breast cancer patients from the TCGA (The Cancer Genome Atlas), we consider four clinical attributes (AJCC Stage, Tumor Stage, ER-Status and PAM50 mRNA Subtypes), and build predictive models using three different modalities of data (histopathological images, transcriptomics and proteomics). Following which, we identify critical multi-level features that drive successful classification of patients for the various different cohorts. To build predictors for each data type, we employ widely used "best practice" techniques including CNN-based (convolutional neural network) classifiers for histopathological images and regression models for proteogenomic data. While, as expected, histology images outperformed molecular features while predicting cancer stages, and transcriptomics held superior discriminatory power for ER-Status and PAM50 subtypes, there exist a few cases where all data modalities exhibited comparable performance. Further, we also identified sets of key genes and proteins whose expression and abundance correlate across each clinical cohort including (i) tumor severity and progression (incl. GABARAP), (ii) ER-status (incl. ESR1) and (iii) disease subtypes (incl. FOXC1). Thus, we quantitatively assess the efficacy of different data types to predict critical breast cancer patient attributes and improve disease characterization.

PRECISION MEDICINE: FROM DIPLOTYPES TO DISPARITIES TOWARDS IMPROVED HEALTH AND THERAPIES

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

LOCAL ANCESTRY TRANSITIONS MODIFY SNP-TRAIT ASSOCIATIONS

Alexandra E. Fish[1], Dana C. Crawford[2], John A. Capra[1], William S. Bush[2]

[1]Vanderbilt University, [2]Case Western Reserve University

Genomic maps of local ancestry identify ancestry transitions – points on a chromosome where recent recombination events in admixed individuals have joined two different ancestral haplotypes. These events bring together alleles that evolved within separate continential populations, providing a unique opportunity to evaluate the joint effect of these alleles on health outcomes. In this work, we evaluate the impact of genetic variants in the context of nearby local ancestry transitions within a sample of nearly 10,000 adults of African ancestry with traits derived from electronic health records. Genetic data was located using the Metabochip, and used to derive local ancestry. We develop a model that captures the effect of both single variants and local ancestry, and use it to identify examples where local ancestry transitions significantly interact with nearby variants to influence metabolic traits. In our most compelling example, we find that the minor allele of rs16890640 occuring on a European background with a downstream local ancestry transition to African ancestry results in significantly lower mean corpuscular hemoglobin and volume. This finding represents a new way of discovering genetic interactions, and is supported by molecular data that suggest changes to local ancestry may impact local chromatin looping.

EVALUATION OF PREDIXCAN FOR PRIORITIZING GWAS ASSOCIATIONS AND PREDICTING GENE EXPRESSION

Binglan Li[1], Shefali S. Verma[1,2], Yogasudha C. Veturi[2], Anurag Verma[1,2], Yuki Bradford[2], David W. Haas[3,4], Marylyn D. Ritchie[1,2]

[1]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA; [2]Biomedical and Translational Informatics Institute, Danville, PA, USA; [3]Department of Medicine, Pharmacology, Pathology, Microbiology & Immunology, Vanderbilt University School of Medicine, Nashville, TN, USA; [4]Department of Internal Medicine, Meharry Medical College, Nashville, TN, USA

Genome-wide association studies (GWAS) have been successful in facilitating the understanding of genetic architecture behind human diseases, but this approach faces many challenges. To identify disease-related loci with modest to weak effect size, GWAS requires very large sample sizes, which can be computational burdensome. In addition, the interpretation of discovered associations remains difficult. PrediXcan was developed to help address these issues. With built in SNP-expression models, PrediXcan is able to predict the expression of genes that are regulated by putative expression quantitative trait loci (eQTLs), and these predicted expression levels can then be used to perform gene-based association studies. This approach reduces the multiple testing burden from millions of variants down to several thousand genes. But most importantly, the identified associations can reveal the genes that are under regulation of eQTLs and consequently involved in disease pathogenesis. In this study, two of the most practical functions of PrediXcan were tested: 1) predicting gene expression, and 2) prioritizing GWAS results. We tested the prediction accuracy of PrediXcan by comparing the predicted and observed gene expression levels, and also looked into some potential influential factors and a filter criterion with the aim of improving PrediXcan performance. As for GWAS prioritization, predicted gene expression levels were used to obtain gene-trait associations, and background regions of significant associations were examined to decrease the likelihood of false positives. Our results showed that 1) PrediXcan predicted gene expression levels accurately for some but not all genes; 2) including more putative eQTLs into prediction did not improve the prediction accuracy; and 3) integrating predicted gene expression levels from the two PrediXcan whole blood models did not eliminate false positives. Still, PrediXcan was able to prioritize GWAS associations that were below the genome-wide significance threshold in GWAS, while retaining GWAS significant results. This study suggests several ways to consider PrediXcan's performance that will be of value to eQTL and complex human disease research.

READING BETWEEN THE GENES: COMPUTATIONAL MODELS TO DISCOVER FUNCTION AND/OR CLINICAL UTILITY FROM NONCODING DNA

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

PAN-CANCER ANALYSIS OF EXPRESSED SOMATIC NUCLEOTIDE VARIANTS IN LONG INTERGENIC NON-CODING RNA

Travers Ching[1,2], Lana X. Garmire[1,2]

[1]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa Honolulu, HI 96822, USA; [2]Epidemiology Program, University of Hawaii Cancer Center Honolulu, HI 96813, USA

Long intergenic non-coding RNAs have been shown to play important roles in cancer. However, because lincRNAs are a relatively new class of RNAs compared to protein-coding mRNAs, the mutational landscape of lincRNAs has not been as extensively studied. Here we characterize expressed somatic nucleotide variants within lincRNAs using 12 cancer RNA-Seq datasets in TCGA. We build machine-learning models to discriminate somatic variants from germline variants within lincRNA regions (AUC 0.987). We build another model to differentiate lincRNA somatic mutations from background regions (AUC 0.72) and find several molecular features that are strongly associated with lincRNA mutations, including copy number variation, conservation, substitution type and histone marker features.

TEXT MINING AND VISUALIZATION FOR PRECISION MEDICINE

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

GeneDive: A GENE INTERACTION SEARCH AND VISUALIZATION TOOL TO FACILITATE PRECISION MEDICINE

Paul Previde[1], Brook Thomas[1], Mike Wong[1], Emily K. Mallory[2], Dragutin Petkovic[1], Russ B. Altman[2], Anagha Kulkarni[1]

[1]San Francisco State University, [2]Stanford University

Obtaining relevant information about gene interactions is critical for understanding disease processes and treatment. With the rise in text mining approaches, the volume of such biomedical data is rapidly increasing, thereby creating a new problem for the users of this data: information overload. A tool for efficient querying and visualization of biomedical data that helps researchers understand the underlying biological mechanisms for diseases and drug responses, and ultimately helps patients, is sorely needed. To this end we have developed GeneDive, a web-based information retrieval, filtering, and visualization tool for large volumes of gene interaction data. GeneDive offers various features and modalities that guide the user through the search process to efficiently reach the information of their interest. GeneDive currently processes over three million gene-gene interactions with response times within a few seconds. For over half of the curated gene sets sourced from four prominent databases, more than 80% of the gene set members are recovered by GeneDive. In the near future, GeneDive will seamlessly accommodate other interaction types, such as gene-drug and gene-disease interactions, thus enabling full exploration of topics such as precision medicine. The GeneDive application and information about its underlying system architecture are available at http://www.genedive.net.

APPLICATIONS OF GENETICS, GENOMICS AND BIOINFORMATICS IN DRUG DISCOVERY

POSTER PRESENTATIONS

CELL-SPECIFIC PREDICTION AND APPLICATION OF DRUG-INDUCED GENE EXPRESSION PROFILES

Rachel Hodos[1,2], Ping Zhang[3], Hao-Chih Lee[1], Qiaonan Duan[1], Zichen Wang[1], Neil R. Clark[1], Avi Ma'ayan[1], Fei Wang[3,4], Brian Kidd[1], Jianying Hu[3], David Sontag[5], Joel T. Dudley[1]

[1]Icahn School of Medicine at Mount Sinai, [2]New York University, [3]IBM T. J. Watson Research Center, [4]Cornell University, [5]Massachusetts Institute of Technology

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

SYSTEMATIC DISCOVERY OF GENOMIC MARKERS FOR CLINICAL OUTCOMES THROUGH COMBINED ANALYSIS OF CLINICAL AND GENOMIC DATA

Jinho Kim[1], Hongui Cha[2], Hyun-Tae Shin[2], Boram Lee[2], Jae Won Yun[2], Joon Ho Kang[3], Woong-Yang Park[1]

[1]Samsung Medical Center, [2]Sungkyunkwan University, [3]Sungkyunkwan University School of Medicine

Molecular profiling is a key component of precision medicine for cancer, as it provides targetable gene or pathways to prevent the tumor to grow. In this regard, more and more cancer clinics employ clinical sequencing platform and are accumulating clinicogenomics data. However, it has not been systematically studied how genomic alterations in particular variants in DNA can benefit in predicting clinical outcomes. Here we describe systematic analyses to gain biological insights from a cancer genome databank associated with the clinical information. We established a large databank of clinical and genomic information through our NGS-based clinical sequencing platform, CancerSCAN. We identified novel clinically relevant variant markers which potentially implicated in patient survival and response to chemotherapeutic agents. Finally, we build a multigene model to predict clinical outcome. The model correctly captured clinically relevant somatic variants and was validated using an independent cohort. Our study provides a valuable resource to realize precision oncology.

IDENTIFICATION OF A PREDICTIVE GENE SIGNATURE FOR DIFFERENTIATING THE EFFECTS OF CIGARETTE SMOKING

Gang Liu[1], Justin Li[2], G.L. Prasad[1]

[1]RAI Services Company, P.O. Box 1487, Winston-Salem, NC 27102, USA;
[2]AccuraScience, 5721 Merle Hay Road, Johnston, IA 50131, USA

Background: Chronic cigarette smoking adversely impacts multiple organs and is a major risk factor for several diseases such as cancer, cardiovascular diseases, and chronic pulmonary obstructive disease (COPD). Because smoking-related diseases often develop over a long period, it is useful to investigate the effects of smoking in healthy individuals to understand the pre-clinical changes that lead to disease states. Those early molecular events could be further developed into biomarkers that are indicative of the adverse effects of smoking. Several classes of different tobacco products, including electronic cigarettes (E-cigs), are currently marketed in the USA, and their impact on consumers has not yet been fully understood. Given that there is no epidemiology data available for these new classes of tobacco products, an understanding of the early molecular and cellular changes in healthy consumers could help to differentiate the effects of cigarettes and other classes of tobacco products. Toward that end, in this study, we aim to develop predictive gene signatures that can be used to differentiate smokers from non-tobacco consumers.

Methods: The data we used for identification of gene signatures were derived from blood-based genome-wide expression profiles from 40 smokers and 40 non-tobacco consumers enrolled in a cross-sectional biomarker study. We systematically evaluated the performance of several machine learning algorithms. These algorithms are combinations of four classification methods, including Support Vector Machine (SVM), and four feature selection methods including Recursive Feature Elimination (RFE). Each gene expression signature model was constructed using a two-layer cross-validation scheme. They were evaluated using accuracy and Mathew's correlation coefficient (MCC), which are performance evaluation metrics widely used in machine learning techniques.

Results: Our results suggest that SVM combined with RFE outperforms the 15 other algorithms we have tested. This led to identification of a 32-gene signature with high sensitivity and specificity. In addition, this new gene signature achieves excellent validation results (accuracy: 0.87, MCC: 0.7) when evaluated using another independent microarray dataset from smokers and non-smokers. The genes in the 32-gene signature include previously reported gene biomarkers such as GPR15, SASH1, and LRRN3, and also consist of additional novel genes associated with inflammation, liver injury, and arachidonic acid metabolism. We are currently working to further refine and validate this gene signature using other publically-available smoking-related gene expression datasets and the polymerase chain reaction-based assay.

Conclusions: We have described a high-performing 32-gene signature that enables prediction of molecular changes in healthy smokers. This gene signature could aid in differentiating the effects of additional classes of tobacco products such as E-cigs.

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

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

Dart NeuroScience

We are interested in discovering 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 diseases. We implemented 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.extremememorychallenge.com) – to screen through an unlimited number of subjects to find individuals with exceptional memory consolidation abilities. A subset of subjects were validated by a battery of secondary memory tasks and provided saliva samples from which we can isolate DNA for GWAS. To date, 26,348 participants from 187 nations have been screened (with 16,486 completing both sessions). The sample is primarily Caucasian (58%), post-secondary school-educated (64%), average age of 34 years old, and equal numbers of each gender. The average forgetting rate across sessions was 10%. The secondary screening involved memory, IQ, attentional control, and personality measures. Analyses are underway to determine the relationship between exceptional memory and genetics.

EXTRACTING A BIOLOGICALLY RELEVANT LATENT SPACE FROM CANCER TRANSCRIPTOMES WITH VARIATIONAL AUTOENCODERS

Gregory P. Way, Casey S. Greene

Genomics and Computational Biology Graduate Program, Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104 USA

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

CHALLENGES OF PATTERN RECOGNITION IN BIOMEDICAL DATA
ORAL PRESENTATION

POSTER PRESENTATIONS

LARGE-SCALE ANALYSIS OF DISEASE PATHWAYS IN THE HUMAN INTERACTOME

Monica Agrawal[1], Marinka Zitnik[1], Jure Leskovec[1,2]

[1]Department of Computer Science, Stanford University; [2]Chan Zuckerberg Biohub, San Francisco, CA

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

PROFILING OF SOMATIC ALTERATIONS IN BRCA1-LIKE BREAST TUMORS

Youdinghuan Chen[1,2,3], Yue Wang[3,4], Lucas A. Salas[1], Todd W. Miller[3,7], Jonathan D. Marotti[5], Nicole P. Jenkins[2], Arminja N. Kettenbach[2,3,7], Chao Cheng[3,4,7], Brock C. Christensen[1,3,7]

[1]Department of Epidemiology, [2]Department of Biochemistry and Cell Biology, [3]Department of Molecular and Systems Biology, [4]Department of Genetics, [5]Department of Pathology and Laboratory Medicine, [6]Department of Biomedical Data Science at Geisel School of Medicine, Dartmouth, Lebanon, NH 03756;
[7]Norris Cotton Cancer Center, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756

Germline or somatic mutation in BRCA1 is associated with an increased risk of breast cancer and more aggressive tumor subtypes. BRCA1-deficient tumor cells have defective homologous recombination (HR) DNA repair, exhibiting genome instability and aneuploidy. HR deficiency can also arise in tumors in the absence of BRCA1 mutation. An HR-deficient, BRCA1-like phenotype has been referred to as “BRCAness.” BRCA1-like cancers exhibit worse prognosis but are selectively sensitive to chemotherapeutic treatments (e.g. platinum-based alkylating agents). However, the molecular landscapes of BRCA1-like breast tumors remain largely unknown in part because they are less common in the general population. By applying a copy number-based classifier, we observed that >30% of The Cancer Genome Atlas (TCGA) breast tumors are BRCA1-like even though only ~3% tumors analyzed carry a BRCA1 mutation or promoter hypermethylation. Separately, a differential analysis controlling for hormone receptor status, subject age, tumor stage and purity revealed a significant increase in DNA methyltransferase 1 (DNMT1) protein expression in BRCA1-like tumors. In addition, differentially methylated gene sets in BRCA1-like tumors indicated a strong enrichment in developmental signaling and a moderate involvement in gene transcription. Profiling of concomitant somatic alteration landscapes in BRCA1-like breast tumors provides alternative strategies to identify this subset of tumors and insights into novel potential therapeutic approaches.

USING ARTIFICIAL INTELLIGENCE IN DIGITAL PATHOLOGY TO CLASSIFY MELANOCYTIC LESIONS

Steven N. Hart, W. Flotte, A.P. Norgan, K.K. Shah, Z.R. Buchan, K.B. Geiersbach, T. Mounajjed, T.J. Flotte

Mayo Clinic, 200 First St. SW, Rochester, MN 55901

Examination of hematoxylin and eosin staining (H&E) stained slides by light microscopy has been the cornerstone of histopathology for over a century. During microscopic examination, a pathologist uses salient clinical information, pattern matching and feature recognition (shape, color, structure, etc.) to render a diagnosis. Recently, whole-slide image (WSI) scanners have made it possible to fully digitize pathology slides. In addition to enabling long term slide preservation and facilitating slide sharing for collaboration or second opinions, digitization of pathology slides allows for the development and utilization Artificial Intelligence (AI)-driven diagnostic tools. We conducted a pilot study to test the ability an AI convolutional neural network (CNN) to distinguish between two types of melanocytic lesions, Conventional and Spitz nevi. We sought to determine the added value of pathologist-assisted training by comparing training effectiveness of complete slide analysis versus training on pathologist selected image patches. Images were classified by a deep CNN using Google's TensorFlow framework. We found significant improvement in classification accuracy when the model was trained from the pathologist-curated image set. These data provide strong evidence for the continued development of AI-driven diagnostic tools in digital pathology, and highlights the added value of domain experts when building AI workflows. Future directions of this work include expanding the number melanocytic lesions recognized by this tool, and enhancing its clinical performance through incorporation of molecular, demographic, and outcomes data.

A MACHINE LEARNING APPROACH TO STUDY COMMON GENE EXPRESSION PATTERNS

Mingze He[1,2], Carolyn J. Lawrence-Dill[1,2,3]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA, 50011; [2]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, USA 50011; [3]Department of Agronomy, Iowa State University, Ames, Iowa, USA 50011

Gene expression landscape changes according to certain circumstances, such as stress responses. The main difficulties in predicting common expression patterns among groups of genes lay in locating reliable gene markers and developing novel statistical approaches. We firstly build a shared gene ontology (GO) correlated grouping database by natural language processing (NLP). Further, we test and apply a mixture of supervised and unsupervised machine learning algorithms to compare principal components of expression patterns across species. We found several surprising common expression patterns between maize genes and human tumor cell lines if G-quadruplex (G4) used as gene classifier. Especially, response to reactive oxygen species (ROS) related G4 carrying genes show a significant clustering of maize under cold and UV stress with human tumor cell lines. This result implies that G4 regulate nearby genes under similar stress situation.

GENERAL

POSTER PRESENTATIONS

DATABASE-FREE METAGENOMIC ANALYSIS WITH AKRONYMER

Gabriel Al-Ghalith[1], Abigail Johnson[2], Pajau Vangay[1], Dan Knights[3]

[1]Bioinformatics and Computational Biology - University of Minnesota; [2]The Biotechnology Institute - University of Minnesota; [3]Department of Computer Science and Engineering - University of Minnesota

Microbiome research is characterized by the comparison of microbial community census data inferred from biological samples. To create these censuses, metagenomic DNA is typically clustered, aligned, or otherwise annotated to form a set of features with which to evaluate and compare microbial communities. These features may take different forms. Amplicon-based studies may use reference-based approaches and/or clustering of similar reads to distill a representative set of features such as operational taxonomic units. Shotgun-based approaches can result in finer-grained, less biased taxonomic resolution, but often rely on reference databases or classifiers trained on known microbial entities. While taxonomy and other database annotations are useful for interpretation, they may mask useful sequence-level information for comparing samples to each other. In particular, whenever there is not enough sequence data from particular organisms in the reference database (or raw reads) to identify them reliably, information about these organisms can be lost or misattributed. This causes many environments to be difficult or even impossible to compare with current methods. Further, clustering or reference-based analyses are typically computationally demanding. We present a complementary (or alternative) strategy for microbiome comparison in the software aKronyMer. It uses a novel, probability adjusted deterministic k-mer distance metric and ultrafast non-heuristic Nei-Saitou-based tree clustering algorithm to rapidly calculate alpha diversity, beta diversity, and sample inter-relatedness trees with either amplicon or shotgun sequence data directly without a database. It is robust to low-depth sequencing, it recovers person-specific signatures with fewer than 100,000 shotgun reads per sample in a dataset of 34 healthy individuals, and it recapitulates other expected trends in public datasets. Additionally, aKronyMer can be used to infer phylogenetic trees from amplicon data in seconds on a laptop, create a whole-genome phylogenomic tree from all ~100,000 RefSeq microbial genomes in a few hours on a desktop, denoise reads during processing, and in other potential applications.

SOFTWARE COMPARISON FOR PREPROCESSING GC/LC-MS-BASED METABOLOMICS DATA

Julian Aldana[1], Monica Cala Molina[1], Martha Zuluaga[2]

[1]Department of Chemistry Grupo de Investigación en Química Analítica y Bioanalítica (GABIO), Universidad de los Andes Bogotá DC, Colombia; [2]Department of Chemistry Grupo de Investigación en Cromatografía y Técnicas Afines (GICTA) Universidad de Caldas Manizales Colombia

Metabolomics data preprocessing is the first step from raw instrument output to biological inference, and it is crucial for the discovery of metabolic signatures related to a particular physiopathological state of an organism. Moreover, data handling of gas chromatography/mass spectrometry (GC/MS) and liquid chromatography/mass spectrometry (LC/MS) datasets are challenging due to its size, complexity and noise. Therefore, data preprocessing is performed as a multi-step task that involves: filtering, peak detection, deconvolution, and alignment, which can be carried out using a wide variety of algorithms and software packages. Given the lack of a singular preprocessing software as a benchmark, the goal of this study is to compare the performance in the preprocessing of GC/LC-MS data between open source platforms (MZmine 2, XCMS online and MetaboAnalyst 3.0) and commercial software (MassHunter Profinder 8.0 and Metaboliteplot). For this purpose, data sets were collected from the analysis of replicate samples from a plasma pooling, and we follow a workflow process in each software adjusting the parameters in a similar way to allow the comparison. Then, the data generated was analyzed to determine the number of features, coefficient of variation and peak area. As a result, significant differences were determined in the quantitative performance of the preprocessing evaluated packages for both GC and LC-MS data sets. Finally, this comparison allowed us to evaluate the magnitude of preprocessing effect in the final output in MS -based metabolomic data, and how the results of different software can be compared each other.

GATEKEEPER: A NEW HARDWARE ARCHITECTURE FOR ACCELERATING PRE-ALIGNMENT IN DNA SHORT READ MAPPING

Mohammed Alser[1], Hasan Hassan[2], Hongyi Xin[3], Oğuz Ergin[4], Onur Mutlu[2], Can Alkan[1]

[1]Bilkent University, [2]ETH Zurich, [3]Carnegie Mellon University, [4]TOBB University of Economics and Technology

Motivation: Until today, it remains challenging to sequence the entire DNA molecule as a whole. In the era of high throughput DNA sequencing (HTS) technologies, genomes are sequenced relatively quickly but result in an excessive number of small DNA segments (called short reads and are about 75-300 basepairs long). Resulting reads do not have any information about which part of genome they come from; hence the biggest challenge in genome analysis is to determine the origin of each of the billions of short reads within a reference genome to construct the donor's complete genome. Identifying the potential origin of each read, called alignment, typically performed using quadratic-time dynamic programming algorithms. These optimal alignment algorithms are unavoidable and essential for providing accurate information about the quality of the alignment. In recent works [1-4], researchers observed that the majority of candidate locations in the reference genome do not align with a given read due to high dissimilarity. Calculating the alignment of such incorrect candidate locations wastes the execution time and incur significant computational burden. Therefore, it is crucial to develop a fast and effective heuristic method that can detect incorrect candidate locations and eliminate them before invoking computationally costly alignment algorithms. Results: We propose GateKeeper, a new hardware accelerator that functions as a pre-alignment step that quickly filters out most incorrect candidate locations. GateKeeper is the first design to accelerate pre-alignment using Field-Programmable Gate Arrays (FPGAs), which can perform pre-alignment much faster than software. When implemented on a single FPGA chip, GateKeeper maintains high accuracy (on average >96%) while providing, on average, 90-fold and 130-fold speedup over the state-of-the-art software pre-alignment techniques, Adjacency Filter and Shifted Hamming Distance (SHD), respectively. The addition of GateKeeper as a pre-alignment step can reduce the verification time of the mrFAST mapper by a factor of 10. Availability: GateKeeper is open-source and freely available online at https://github.com/BilkentCompGen/GateKeeper. An extended version of this work appears in [1].

References: [1] Alser, M., et al., GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping. Bioinformatics, 2017. 33(21): p. 3355-3363. [2] Xin, H., et al., Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter to Accelerate Alignment Verification in Read Mapping. Bioinformatics, 2015. 31(10): p. 1553-1560. [3] Xin, H., et al., Accelerating read mapping with FastHASH. BMC genomics, 2013. 14(Suppl 1): p. S13. [4] Kim, J., et al., Genome Read In-Memory (GRIM) Filter: Fast Location Filtering in DNA Read Mapping using Emerging Memory Technologies, to appear in BMC Genomics, 2018.

MODELING THE ENHANCER ACTIVITY THROUGH THE COMBINATION OF EPIGENETIC FACTORS

MinGyun Bae, Taeyeop Lee, Jaeho Oh, Jun Hyeong Lee, Jung Kyoon Choi

Department of Bio Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

Epigenome maps allow us to predict thousands of putative regulatory regions such as promoter, insulators and enhancers in various cell lines through in vivo epigenomic signatures and are widely used for studying gene regulation of developmental process and disease. Especially, super- enhancers, which consist of clusters of active enhancers predicted from H3K27ac signal, are known to regulate near genes that are important in controlling and defining cell identity. However, the combination of transcription factors for regulating enhancer activity is not studied yet. In this study, we used massively parallel reporter assay (MPRA) data which measure the quantitative activity of regulatory regions to identify enhancers. Through 5-nucleotide resolution tiling of overlapping MPRA constructs with a probabilistic graphical model, we estimated the high resolution activity spanning 15000 putative regulatory regions in K562 and HepG2 cell line. According to the ratio of activity at boundary and center of regulatory region, we identified thousands of enhancers candidates. Using these enhancers, we developed a random forest model to identify the epigenetic differences using about 300 histone modifications and transcription factors in encyclopedia of DNA elements (ENCODE). Through the performance test by area under curve (AUC), we confirmed that our model accurately predicted the enhancers. In conclusion, we identified enhancers through high-throughput reporter assay and found the epigenetic features through random forest modelling.

FREQUENCY AND PROPERTIES OF MOSAIC SOMATIC MUTATIONS IN A NORMAL DEVELOPING BRAIN

Taejeong Bae[1], Jessica Mariani[2], Livia Tomasini[2], Bo Zhou[3], Alexander E. Urban[3], Alexej Abyzov[1], Flora M. Vaccarino[2]

[1]Mayo Clinic, [2]Yale University, [3]Stanford University

As mounting evidence indicates, each cell in the human body has its own genome, a phenomenon called somatic mosaicism. Few studies have been conducted to understand post-zygotic accumulation of mutations in cells of the healthy human body. Starting from single cells, directly obtained from three fetal brains, we established 31 separate colonies of neuronal progenitor cells, and carried out whole-genome sequencing on DNA from each colony. The clonal nature of these colonies allows a high-resolution analysis of the genomes of the founder progenitor cells without being confounded by the artifacts of in vitro single cell whole genome amplification. Across the three brains we detected 200 to 400 non-germline SNVs per clone. Validation experiments (with PCR, digital droplet PCR, and capture deep sequencing) revealed high specificity (>95%) and sensitivity (>80%) of the SNVs as well as confirmed the presence of over a hundred of SNVs in the original brain tissues, thereby proving that the detected SNVs represent genuine mosaic variants present in neuronal progenitors. The per-cell number of mosaic SNVs increased linearly with brain age allowing us to estimate the mutation rate at about 8.6 SNVs per cell division. Dozens of SNVs were genotyped in multiple different regions of a brain and even in blood, suggesting that they have occurred prior to gastrulation. Using these SNVs, we reconstructed cell lineages for the first five post-zygotic cleavages and calculated a mutation rate of ~1.3 SNVs per division per daughter cell. Comparison of mutation spectra revealed a shift towards oxidative damage-related mutations in neurogenesis. Both neurogenesis and early embryogenesis exhibit drastically more mutagenesis than adulthood. On a coarse-grained scale mosaic SNVs were distributed uniformly across the genome and were enriched in mutational signatures observed in medulloblastoma, neuroblastoma, as well as in a signature observed in all cancers and in de novo variants and which, as we previously hypothesized, is a hallmark of normal cell proliferation. Correlations with histone marks further strengthened the similarity of mosaic mutations in normal fetal brain with somatic mutations reported for brain cancers. On a smaller scale SNVs were mostly benign, showed no association with any GO category and tended to avoid DNAse hypersensitive sites. These findings reveal a large degree of somatic mosaicism in the developing human brain, link de novo and cancer mutations to normal mosaicism and set a baseline for mosaic genome variation related to human brain development and function.

CYCLONOVO: DE NOVO SEQUENCING ALGORITHM DISCOVERS NOVEL CYCLIC PEPTIDE NATURAL PRODUCTS IN SUNFLOWER AND CYANOBACTERIA USING TANDEM MASS SPECTROMETRY DATA

Bahar Behsaz[1], Hosein Mohimani[2], Alexey Gurevich[3], Andrey Prjibelski[3], Mark F. Fisher[4], Larry Smarr[2], Pieter C. Dorrestein[5], Joshua S. Mylne[4], Pavel A. Pevzner[2]

[1]Bioinformatics and Systems Biology Program, University of California at San Diego, La Jolla, USA; [2]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, USA; [3]Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St Petersburg, Russia; [4]The University of Western Australia, School of Molecular Sciences and ARC Centre of Excellence in Plant Energy Biology, Crawley, Australia; [5]Department of Pharmacology, University of California at San Diego, La Jolla, USA

Cyclopeptides represent an important class of natural products with an unparalleled track record in pharmacology: many antibiotics, antitumor agents, and immunosuppressors, are cyclopeptides. While billions of tandem mass spectra of natural products have been deposited to Global Natural Products Social (GNPS) molecular network, the discovery of novel cyclopeptides from this gold mine of spectral data remains challenging. As the result, only a small fraction of spectra in the GNPS molecular network have been identified so far. To address this bottleneck, we developed CycloNovo algorithm for de novo cyclopeptide sequencing based on the concept of the de Bruijn graphs, the workhorse of modern genome sequencing algorithms. Given a spectral dataset, CycloNovo first identifies a subset of this dataset that may represent cyclic and branch-cyclic peptides by analyzing spectral-convolution of each spectrum. Afterward, it attempts to de novo sequence each spectrum of putative cyclic or branch-cyclic peptides. CycloNovo pipeline includes (i) computing the spectral convolution of each spectrum, and extracting the set of masses that represent putative amino acids in the unknown PNP, (ii) computing compositions of masses that matches the precursor mass of the spectrum, (iii) constructing potential 5-mers for each composition with high score against the spectrum, (iv) constructing a de Bruijn graph with those 5-mers, (v) traversing the de Bruijn graph and generating candidate sequences, and (vi) computing the Peptide-Spectrum-Match (PSM) score for each candidate sequence. CycloNovo revealed many still unknown cyclopeptides (hundreds of novel cyclopeptide families) illustrating that currently known cyclopeptides represent just a small fraction of cyclopeptides whose spectra are already deposited into public databases such as GNPS. CycloNovo addresses the challenge of analyzing the "dark matter of cyclopeptidome" by applying de Bruijn graphs to cyclopeptide sequencing. It correctly sequenced many known cyclopeptides in a blind experiment and reconstructed novel cyclopeptides originated from plants and cyanobacteria that were further validated using RNA-seq data and genome mining, the first cyclopeptides discovered in a completely automated de novo fashion. Our analysis of human microbiome is the first demonstration that numerous bioactive cyclopeptides from consumed plants remain stable in the proteolytic human gut environment and thus are expected to interact with human microbiome. In addition, it revealed a large number of still unknown cyclopeptides in the human gut that are either a part of the human diet or are products of the human gut's microbiome.

FUNCTIONAL ANNOTATION OF GENOMIC VARIANTS IN STUDIES OF LATE-ONSET ALZHEIMER'S DISEASE

Mariusz Butkiewicz, Jonathan L. Haines, William S. Bush

Institute for Computational Biology and Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH USA

Annotation of genomic variants is an increasingly important and complex part of the analysis of sequence-based genomic analyses. Computational predictions of variant function are routinely incorporated into gene-based analyses of rare-variants, though to date most studies use limited information for assessing variant function that is often agnostic of the disease being studied. In this work, we outline an annotation process motivated by the Alzheimer's Disease Sequencing Project, and illustrate the impact of including tissue-specific transcript sets and sources of gene regulatory information, and assess the potential impact of changing genomic builds on the annotation process. While these factors only impact a small proportion of total variant annotations (~5%), they influence the potential analysis of a large fraction of genes (~25%). Variant annotation is available for bulk download, and individual variant annotations are also available via the NIAGADS GenomicsDB.

OCTAD: AN OPEN CANCER THERAPEUTIC DISCOVERY WORKSPACE IN THE ERA OF PRECISION MEDICINE

Bin Chen, Benjamin S. Glicksberg, William Zeng, Yuying Chen, Ke Liu

Institute for Computational Health Sciences, University of California, San Francisco, 550 16th Street, San Francisco, California 94143, USA

Rapidly decreasing costs of RNA sequencing have enabled large-scale profiling of cancer tumor samples with precisely defined clinical and molecular features (e.g., Low grade IDH1 mutant Glioma) . Identifying drugs targeting a specific subset of cancer patients, particularly those that do not respond to conventional treatments, is critically important for translational research. Many studies have demonstrated the utility of a systems-based approach that connects cancers to efficacious drugs through gene expression signatures to prioritize drugs from a large drug library. From our previous work on liver cancer, Ewing's Sarcoma, and Basal cell carcinoma, we have shown that the success of this approach is made possible by critical procedures, such as quality control of tumor samples, selection of appropriate reference tissues, evaluation of disease signatures, and weighting cancer cell lines. There is a plethora of relevant datasets and analysis modules that are publicly available, yet are isolated in distinct silos, making it tedious to implement this approach in translational research. As such, we present the current protocol, which we envision as a best practice to prioritize drugs for further experimental evaluation, primarily based on open transcriptomic datasets and the free open-source R language and Bioconductor packages.   In this project, we retrieved patient tumor samples based on specified clinical and/or molecular features from the Genomic Data Commons Data Portal using an API. We then created a gene expression signature for these samples through employing normalized RNA-Seq counts processed in the UCSC Xena project, where all RNA-Seq samples from TCGA, TARGET, and GTEx were aligned and normalized using the same pipeline. We evaluated the quality of samples based on their purity and correlation with cancer cell lines. The reference tissue samples were selected based on their profile similarity with GTEx samples. We evaluated each disease signature via a cross-validation approach. We then created drug signatures using a similar procedure from large-scale, open access platforms, namely the LINCS L1000 library, which consists of over 20,000 compounds. Our pipeline can then compute and assess the reversal potency between the disease signature and each drug signature. The drugs that present high reversal potency are prioritized as drug hits. Finally, we performed enrichment analysis of drug hits to identify compelling enriched targets and pathways. For our pilot study, we use IDH1 mutant Oligodendroglioma as a case study, where the efficacy of over 300 LINCS compounds was measured in three relevant cell lines. We have shown that our prediction corroborate with the experimental data.

DEEP LEARNING PREDICTS TUBERCULOSIS DRUG RESISTANCE STATUS FROM WHOLE-GENOME SEQUENCING DATA

Michael L. Chen, Isaac S. Kohane, Andrew L. Beam, Maha Farhat

Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Background The diagnosis of multidrug resistant and extensively drug resistant tuberculosis is a global health priority. There is a pressing need for a rapid and comprehensive drug susceptibility test that can circumvent the limited scope of conventional methods and the associated long wait times. We sought to implement the first deep learning framework as a predictive diagnostic tool for Mycobacterium tuberculosis (MTB) drug resistance. Methods Using a large public data set of 3,601 MTB strains that underwent targeted or whole genome sequencing and conventional drug resistance phenotyping, we built the first-of-its-kind multitask wide and deep neural network (WDNN) architecture to predict phenotypic drug resistance to 11 anti-tubercular drugs. We compared performance of the proposed WDNN to regularized logistic regression and random forest models using five-fold cross validation. We conducted permutation tests for evaluating feature importance and a t-distributed stochastic neighborhood embedding (t-SNE) to visualize the high dimensional model output on the full dataset. Results The multitask WDNN achieved state-of-the-art predictive performance compared to regularized logistic regression and random forest: the average sensitivities and specificities, respectively, for all 11 drugs were 87.1% and 93.7% (multitask WDNN), 85.4% and 93.8% (random forest), and 82.2% and 93.9% (regularized logistic regression). The multitask WDNN achieved a higher sum of specificity and sensitivity for 9 of the 11 drugs compared to both the random forest and regularized logistic regression. We show considerable performance gains in our current multitask WDNN with respect to our previously reported random forest model, noting improvements of up to 54.0% in the sum of specificity and sensitivity. Patterns in susceptibility status emerged between drugs after applying t-SNE that correlate well with what is known about the order of MTB drug resistance acquisition. Novel t-SNE findings included major cluster differences between pyrazinamide and other first-line drugs and increased amounts of resistance clusters for capreomycin compared to other second-line drugs. Notable findings in the feature importance analyses included expected shared resistance-associated mutations between drugs and provided new insight potential mechanistic relationships. Capreomycin exclusively shared 10 features with first-line drugs, highlighting potential avenues for future research into the diagnostic similarities between capreomycin and other subtypes of anti-tubercular drugs. Conclusion Our proposed architecture provides a unified model of drug resistance across 11 anti-tubercular drugs and shows considerable performance gains over simper methods. Deep learning has a clear role in improving identification of drug resistant MTB strains and holds promise in bringing sequencing technologies closer to the bedside.

DESIGNING PREDICTION MODEL FOR HYPERURICEMIA WITH VARIOUS MACHINE LEARNING TOOLS USING HEALTH CHECK-UP EHR DATABASE

Eun Kyung Choe[1], Sang Woo Lee[2]

[1]Department of Surgery, Seoul National University College of Medicine; [2]Network Division, Samsung Electronics

Hyperuricemia is an elevated uric acid level in blood. It can lead to gout and nephrolithiasis but also has been implicated as an indicator for disease like metabolic syndrome, diabetes mellitus, cardiovascular disease, and chronic renal disease. The aim of the present study is to design a prediction model for hyperuricemia using EHR database from health check-up using various machine learning tools. From 2005 to 2015, self-paid people had comprehensive health check-up. Input factors were age, gender, body mass index (BMI), blood pressure, waist circumference, white blood cell count, hemoglobin, glucose level, cholesterol, GOT/GPT, GGT, creatinine, triglyceride, urine albumin, smoking/alcohol habit, and diabetes/hypertension/dyslipidemia medication status, which are the factors covered by national health insurance. Output factor was uric acid level which is not included in the national health check-up. All of the data were extracted from the EHR database and text mining was performed. We designed a prediction model for hyperuricemia using machine learning tools such as linear regression model (LR), support vector model (SVM), classification tree model (CT) and neural network model (NN). Machine learning was performed by MATLAB R2016b (The Mathworks, Natick, MA). The prediction power of each models were evaluated by calculating the area under the curve (AUC), sensitivity, specificity and accuracy. Total 55,227 persons were included in the analysis. The median age was 52 years (range 21-95 years) and 53.5% of persons were males. There were 10,586 (19.2%) persons who had uric acid level in hyperuricemia. BMI was higher in hyperuricemia group (25.2+/-3.0 vs. normal uric acid group, 23.4+/-2.9, p<0.001) and there were more alcohol drinking habits in hyperuricemia group (67.8% vs. normal uric acid group, 52.4%, p<0.001). Sorting the results by the accuracy of each machine learning models, the CT showed the highest accuracy of 0.954 (AUC = 0.886 ; sensitivity = 0.792 ; specificity = 0.981) compared to SVM of 0.892 (AUC = 0.630 ; sensitivity = 0.261 ; specificity = 0.999), NN of 0.859 (AUC = 0.770 ; sensitivity = 0.09 ; specificity = 0.991) and LR of 0.857 (AUC = 0.761 ; sensitivity=0.033; specificity=0.997). This study used a health check-up EHR database to predict a disease status (hyperuricemia) using various machine learning tools. Since the amount of EHR database are increasing rapidly, the data included in the database could be used as biomarkers to predict disease status or high risk conditions by modeling a prediction model using machine learning tools. But since the optimal analysis tool or analyzing protocol is not well established and the over-fitting problem is yet not solved, more training and researches in various set of populations should been done in future study for replication.

RICK: RNA INTERACTIVE COMPUTING KIT

Galina A. Erikson, Ling Huang, Maxim Shokhirev

Salk Institute for Biological Studies

The advent of massively parallel sequencing of RNA (RNA-Seq) enables fast and inexpensive global measurement of thousands of genes across biological perturbations involving drug treatment, genetic mutations, and time series. To facilitate comparison, many tools have been developed, however most of these tools require extensive programming and bioinformatics knowledge: little is available for the scientist that wants to analyze their own RNA-seq data but lacks bioinformatics expertise. The RNA Interactive Computing Kit (RICK) aims to fill this gap by providing an interactive web workspace designed to facilitate RNA-Seq analysis and visualization. RICK accepts as input a file with raw read counts for each transcript and sample and performs sample clustering), visualizes the global gene expression with heatmaps, runs principal component analysis and prepares print ready figures. Users can add and remove samples and regenerate new figures on the fly. For differential genes expression users have the option to use: edgeR, Deseq2 or the combination of all and filter the results based on adjusted p-value and fold change. RICK is able to use the DE results from the previous step to identify the significantly altered KEGG pathways or enriched GO terms using the gage or GOseq pathway analysis package with visualization. Users also have the option to upload their customized gene/background gene list to do a DAVID-like analysis. RICK supports RNA-Seq based research by providing a workflow that requires no bioinformatics skills. RICK is freely available at rick.salk.edu.

PRIVATE INFORMATION LEAKAGE IN FUNCTIONAL GENOMICS EXPERIMENTS: QUANTIFICATION AND LINKING

Gamze Gursoy, Mark Gerstein

Program in Computational Biology and Bioinformatics Yale University

The success of the ENCODE (Encyclopedia of DNA Elements) project opened the doors to a deeper understanding of the functional genome through genome-wide experimental assays. Although identifying individuals using DNA variants from whole genome or exome sequencing data is a major privacy and security concern, no study on genomic privacy has focused on the quantity of information in functional genomic experiments such as ChIP-Seq, RNA-Seq and Hi-C, since the majority of this data is partial and biased. Here, we quantify the amount of leaked genotype information in different functional genomic assays at varying coverages. We show that sequencing data from functional genomics assays provides enough private information to be able to link these samples to a panel of individuals with known genotypes.

CARPE D.I.E.M: A DATA INTEGRATION EXPECTATION MAP FOR THE POTENTIAL OF MULTI-`OMICS INTEGRATION IN COMPLEX DISEASE

Tia Tate Hudson, ClarLynda Williams-DeVane

North Carolina Central University, Durham, NC, USA

Advances in high throughput technologies and the availability of multi-`omics data present the opportunity for more holistic understandings of biological regulation in complex diseases and disparities. The complexity and disparate nature of various diseases requires the development of equally complex models with multiple layers of biological information. This however, requires the integration of biological, computational, and statistical domains. Currently, nonetheless, there exist major gaps in the availability and knowledge amongst the three domains. Typically, biologist experience problems with processing and analyzing biological data; therefore, seeking data scientist for more customized analysis. In contrast, some data scientists lack a thorough understanding of the regulation and complex interactions of various systems giving rise to varying complex phenotypes. This generally results in less comprehensive analysis and an overall narrow understanding of complex disease phenotypes, which can only be thoroughly understood when various levels of `omic interactions are considered as a whole. Thus, developing the most comprehensive biological models must consider the multiple appropriate layers of genomic, epigenomic, transcriptomic, proteomic, and metabolomic regulation, as well as the potential role environmental and social factors play at each `omic level. Historically, diverse data types have been considered independently while combinations of two or more data types have been utilized less frequently. Singular analysis of independent `omic contributions of disease often neglect the intricate interactions among the distinct levels giving rise to these complex traits. Although environmental and social factors have a major role in the disparate nature of diverse diseases, many diseases result from mutual alterations in assorted pathways and biological processes, including gene mutations, epigenetic changes, and modifications in gene regulation. Therefore, the various phenotypes in diverse disease represent a major example of the need for integrated biological models for complex trait analysis. In this study, we present the Data Integration Expectation Map (D.I.E.M), where we explore the scientific value of integrating various `omic data combinations that can reveal mechanisms of biological regulation in disease disparities. Our goal is to convey the potential for integration of genomic, epigenomic, transcriptomic, proteomic, and metabolomic data for improving our understanding of the complexity and nature of disparity in complex disease traits. In doing so, this map will address the holes in the various domains necessary for integrated data analysis and interpretation. D.I.E.M will also reveal the expected outcomes for each `omic data type and the various combinations that may or may not divulge a holistic view into complex disease phenotypes. With that, we expect to gain a greater understanding of physiological processes contributing to disparities as well as the role each `omic interaction plays in screening, diagnosis, and prognosis of disease.

IMPROVING GENE FUSION DETECTION ACCURACY WITH FUSION CONTIG REALIGNMENT IN TARGETED TUMOR SEQUENCING

Jin Hyun Ju, Xiao Chen, June Snedecor, Han-Yu Chuang, Ben Mishkanian, Sven Bilke

Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA

Gene fusions have been identified as driver mutations in multiple cancer types, and a number of drugs targeting specific fusions have been developed as treatment options. Therefore, the ability to identify fusions from tumor samples has become critical for the selection of appropriate treatments for patients. Previously, gene fusions have been detected by targeted approaches such as polymerase chain reaction (PCR) or Fluorescent In Situ Hybridization (FISH). These methods not only require prior knowledge of the fusion, but are also labor intensive and not efficient. Newer methods utilizing RNA sequencing (RNA-seq) that are able to detect multiple types of gene fusions with no prior knowledge required have been introduced with the emergence of next-generation sequencing technology. One critical challenge in using RNA-seq data for gene fusion detection is false positive findings introduced by aligner specific biases or regions with sequence similarity in the genome. This problem becomes more apparent in clinical settings where the abundance of fusion transcripts can be limited by the composition and heterogeneity of the tumor sample. To avoid the critical risk of failing to detect a potentially treatable gene fusion, imposing a stringent detection threshold becomes difficult in these situations leading to the inclusion of fusions based on relatively low read evidence. To address this problem, we describe a novel fusion filtering method based on fusion contig realignment that is designed to identify spurious false positive fusions. Our method can be used together with any assembly-based fusion calling method that constructs a contig sequence for each reported fusion. The first step is to realign the fusion contigs with Basic Local Alignment Search Tool (BLAST), which is relatively more flexible in finding alternative alignment results with high sequence similarity. Subsequently, we determine whether a specific fusion call can be supported by evidence found in BLAST alignments. Specifically, we aim to filter out fusions that can be explained by regions originating from a single gene or genomic region, or have weak support on either side of the fusion in BLAST alignments. In our preliminary analysis of 1171 fusion calls in 322 samples, 111 out of 161 false positive calls (68%) were filtered out while no calls from the total of 1010 true positives were filtered out.

SPARSE REGRESSION FOR NETWORK GRAPHS AND ITS APPLICATION TO GENE NETWORKS OF THE BRAIN

Hideko Kawakubo, Yusuke Matsui, Teppei Shimamura

Nagoya Universityi

Recent rare variant analyses of single nucleotide variations (SNVs) and copy number variations (CNVs) has identified dozens of candidate genes that may contribute to neurogenetic disorders such as autism and schizophrenia. However, it is unclear whether and how these disease-causing genes are associated with cellular mechanisms in brain. This problem is a challenging task, since the brain contains hundreds of distinct cell types, each of which has unique morphologies, projections, and functions, and thus disease-causing genes may contribute to different behavioral abnormalities of distinct cell types in the nervous system. In order to identify candidate cell types of the brain related to a complex genetic disorder, we propose a statistical method, called graph oriented sparse learning (GOSPEL).

GRIM-FILTER: FAST SEED LOCATION FILTERING IN DNA READ MAPPING USING PROCESSING-IN-MEMORY TECHNOLOGIES

Jeremie S. Kim[1,2], Damla S. Cali[1], Hongyi Xin[3], Donghyuk Lee[1,4], Saugata Ghose[1], Mohammed Alser[5], Hasan Hassan[2,6], Oğuz Ergin[6], Can Alkan[5], Onur Mutlu[2,1]

[1]ECE Department, Carnegie Mellon University; [2]CS Department, ETH Zurich; [3]CS Department, Carnegie Mellon University; [4]NVIDIA Research; [5]CE Department, Bilkent University; [6]CE Department, TOBB University of Economics and Technology

Seed location filtering is critical in DNA read mapping, a process where billions of DNA fragments (reads) sampled from a donor are mapped onto a reference genome in order to identify the genomic variants of the donor. State-of-the-art read mappers determine the original location of a read sequence within a reference genome in 3 generalized steps. A read mapper 1) quickly generates possible mapping locations for seeds (i.e., smaller segments) within a read, 2) extracts the reference sequence at each of the mapping locations, and 3) determines the similarity score between the read and its associated reference sequences with a computationally-expensive algorithm (i.e., sequence alignment). With the similarity scores across all possible locations, the read mapper can determine the original location of the read sequence. The differences between the read sequence and the matching reference sequence indicate the genomic variants of the donor, which can be further analyzed for preventative care or diagnosis. A seed location filter (e.g., FastHASH [2], SHD [3], GateKeeper [4]) comes into play before sequence alignment (step 3) and reduces the number of unnecessary alignments. A seed location filter efficiently determines whether a candidate mapping location would result in an incorrect mapping before performing the computationally-expensive sequence alignment step for that location. In the ideal case, a seed location filter would discard all poorly matching locations prior to alignment such that there is no wasted computation on unnecessary alignments. We propose a novel seed location filtering algorithm, GRIM-Filter, optimized to exploit 3D-stacked memory systems that integrate computation within a logic layer stacked under memory layers, to perform processing-in-memory (PIM). GRIM-Filter quickly filters seed locations by 1) introducing a new representation of coarse-grained segments of the reference genome, and 2) using massively-parallel in-memory operations to identify read presence within each coarse-grained segment. Our evaluations show that for a sequence alignment error tolerance of 0.05, GRIM-Filter 1) reduces the false negative rate of filtering by 5.59x--6.41x, compared to the best previous seed location filtering algorithm, and 2) provides an end-to-end read mapper speedup of 1.81x--3.65x, compared to a state-of-the-art read mapper employing the best previous seed location filtering algorithm [2]. This work will appear at the 16th Asia Pacific Bioinformatics Conference in January 2018 [1]. The preliminary version of the full article is at https://arxiv.org/pdf/1711.01177.pdf. [1] Kim, Jeremie S, et al. "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies." to appear in BMC Genomics (2018). [2] Xin, Hongyi, et al. “Accelerating read mapping with FastHASH.” BMC Genomics (2013). [3] Xin, Hongyi, et al. “Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping.” Bioinformatics (2015). [4] Alser, Mohammed, et al. "GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping." Bioinformatics (2017).

MULTI-CLASS CLASSIFICATION STRATEGY FOR SUPPORT VECTOR MACHINES USING WEIGHTED VOTING AND VOTING DROP

Sungho Kim, Taehun Kim

Yeungnam University, DGIST

A novel multi-class strategy for Support Vector Machines (SVMs) was developed to perform multi-class classification, such as One Versus One, One Versus All and Dynamic Acyclic Graph. These strategies do not reflect the distance between the hyper-plane that separates two classes and input data. This is not reasonable when the input data is placed near the hyper-plane. The proposed weighted voting resolves this problem by weighting the voting values according to the distance from the boundary and the enhanced performance of the SVMs with the proposed voting drop. The proposed Weighted Voting is based on the voting method. The voting method is carried out by accumulating votes, then choosing the most voted class. The proposed Weighted Voting method is a weighting of the voting value by reflecting the distance from the boundary and margin. Second proposed Voting Drop method is about how to accumulate votes. The novel voting method accumulates every vote but this manner can be a problem because there are redundantly responding SVMs. Because the SVM is a binary classifier, each SVM learns only about two classes. Therefore, a SVM does not have discernment for the non-learned classes. This is why when a SVM predicts data belonging to a non-learned class, the SVM responds redundantly. This irrelevant SVM causes an incorrect vote that makes the decision confused. To resolve this problem, the Voting Drop method drops the redundant votes by removing the irrelevant SVM. This algorithm finds the irrelevant SVM, then dropping the votes caused by the irrelevant SVM. The way to find an irrelevant SVM is to find a least voted class because a least voted class can be thought of as an irrelevant class to input data. As shown in the experiments, evenly reflecting the distance from the hyper-plane and the discernment of the hyper-plane and removing the redundant SVM`s voting leads to higher performance. The proposed methods can be used for a range of classification tasks.

GENOME-WIDE ANALYSIS OF TRANSCRIPTIONAL AND CYTOKINE RESPONSE VARIABILITY IN ACTIVATED HUMAN IMMUNE CELLS

Sarah Kim-Hellmuth[1,2], Matthias Bechheim[3], Benno Pütz[2], Pejman Mohammadi[1,4], Johannes Schumacher[5], Veit Hornung[3,6], Bertram Müller-Myhsok[2], Tuuli Lappalainen[1,4]

[1]New York Genome Center, New York, NY, USA; [2]Max-Planck-Institute of Psychiatry, Munich, Germany; [3]Institute of Molecular Medicine, University of Bonn, Bonn, Germany; [4]Department of Systems Biology, Columbia University, New York, NY, USA; [5]Institute of Human Genetics, University of Bonn, Bonn, Germany; [6]Gene Center and Department of Biochemistry, Ludwig-Maximilians-University Munich, Munich, Germany

The immune system plays a major role in human health and disease. Understanding variability of immune responses on the population level and how it relates to susceptibility to diseases is vital. In this study, we aimed to characterize the genetic contribution to interindividual variability of immune response using genome-wide association and functional genomics approaches. For this purpose, we studied genetic associations to cellular (gene expression) and molecular (cytokine) phenotypes in primary human cells activated with diverse microbial ligands. We isolated monocytes of 134 individuals and stimulated them with three bacterial and viral components (LPS, MDP, and ppp-dsRNA). We performed transcriptome profiling at three time points (0 min/90 min/6 h) and genome-wide SNP-genotyping. In addition, we profiled five cytokines produced by peripheral blood mononuclear cells activated by five components from the same individuals to perform a genome-wide association study. Comparing expression quantitative trait loci (eQTLs) under baseline and upon immune stimulation revealed 417 immune response specific eQTLs (reQTLs). We characterized the dynamics of genetic regulation on early and late immune response, and observed an enrichment of reQTLs in distal cis-regulatory elements. Analysis of signs of recent positive selection and the direction of the effect of the derived allele of reQTLs on immune response suggested an evolutionary trend towards enhanced immune response. Furthermore, multivariate GWAS analysis of cytokine responses to diverse stimuli revealed 159 genome-wide significant loci; however, only a small number of these could be reliably linked to potentially causal eQTLs in monocytes. Finally, given the central role of inflammation in many diseases, we examined reQTLs as a potential mechanism underlying genetic associations to complex diseases. We uncovered novel reQTL effects in multiple GWAS loci, and showed a stronger enrichment of response than constant eQTLs in GWAS signals of several autoimmune diseases. These results indicate a substantial, disease-specific role of environmental interactions with microbial ligands in genetic risk to complex autoimmune diseases. While tissue-specificity of molecular effects of GWAS variants is increasingly appreciated, our results suggest that innate immune stimulation is a key cellular state to consider in future eQTL studies as well as in targeted functional follow-up of GWAS loci.

PREDICTING FATIGUE SEVERITY IN ONCOLOGY PATIENTS ONE WEEK FOLLOWING CHEMOTHERAPY

Kord M. Kober, Xiao Hu, Bruce A. Cooper, Steven M. Paul, Christine Miaskowski

University of California San Francisco

Effective symptom management is a critical component of cancer treatment. Computational tools that predict the course and severity of these symptoms have the potential to assist oncology clinicians to personalize the patient's treatment regimen more efficiently and provide more aggressive and timely interventions. Cancer-related fatigue (CRF) is the most common symptom associated with cancer and its treatments. CRF has a negative impact on the patients' ability to tolerate treatments and on their quality of life. One of the limitations to effective treatment of CRF is the availability of a valid and reliable model to predict the severity of CRF. The objective of this pilot study was to generate a predictive model for fatigue severity 1 week after chemotherapy (CTX) administration (T2) using 28 demographic and clinical characteristics that were collected just prior to CTX administration (T1) in a sample of 1042 cancer patients undergoing CTX. In this pilot study, we used support vector regression (SVR) with a polynomial kernel to predict the severity of the evening fatigue between two different time points during a cycle of CTX. Patients with missing data were removed, leaving a total of 689 for this analysis. Training and testing groups consisted of 518 and 171 patients, respectively. We used 10-times 10-fold cross-validation root-mean-square error (RMSE) to assess the fit of the predictive model. Our model achieved an RMSE/mean of 0.269. The five predictors with the highest importance were: evening fatigue at T1, morning fatigue at T1, attentional function, sleep disturbance, and performance status. The five predictors with the lowest importance were: living alone, caregiver to adult, and level of education, cycle length, and number of metastatic sites. Overall, clinical characteristics associated with cancer and its treatment, including cancer diagnosis, had low importance in the model. These findings suggest that the experience and mechanisms of CRF may be general and not cancer specific. This type of predictive model can be used to identify high risk patients, educate patients about their symptom experience, and improve the timing of pre-emptive and personalized symptom management interventions. These results suggest that the integration of demographic and clinical data can enhance clinical prognostic prediction, which will contribute to the development of precision cancer medicine. Our methods are generalizable to other types of symptoms.

SINGLE-MOLECULE PROTEIN IDENTIFICATION BY SUB-NANOPORE SENSORS

Mikhail Kolmogorov[1], Eamonn Kennedy[2], Zhuxin Dong[2], Gregory Timp[2], Pavel A. Pevzner[1]

[1]Department of Computer Science and Engineering, University of California San Diego, USA; [2]Electrical Engineering and Biological Science, University of Notre Dame, USA

Recent advances in top-down mass spectrometry enabled identification of intact proteins, but this technology still faces challenges. For example, top-down mass spectrometry suffers from a lack of sensitivity since the ion counts for a single fragmentation event are often low. In contrast, nanopore technology is exquisitely sensitive to single intact molecules, but it has only been successfully applied to DNA sequencing, so far. Here, we explore the potential of sub-nanopores for single-molecule protein identification (SMPI) and describe an algorithm for identification of the electrical current blockade signal (nanospectrum) resulting from the translocation of a denaturated, linearly charged protein through a sub-nanopore. The analysis of identification p-values suggests that the current technology is already sufficient for matching nanospectra against small protein databases, e.g., protein identification in bacterial proteomes.

GENE EXPRESSION PROFILE OF OSTEOARTHRITIS AFFECTED FINGER JOINTS

Milica Krunic[1], Klaus Bobacz[2], Arndt von Haeseler[3]

[1]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria; [2]Department of Internal Medicine III, Division of Rheumatology, Medical University of Vienna, Vienna, Austria; [3]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

Osteoarthritis (OA) is a joint disease, which can affect any joint. However, the most frequent non-weight bearing joints affected by OA are hand joints. The most common clinical presentation of hand OA is pain and loss of hand strength, which restricts the ability of people to perform daily activities. Multiple factors can contribute to the development of the hand OA, of which the most frequently observed are: age, gender, genetics, obesity, occupation, and repetitive joint usage. OA in proximal interphalangeal (PIP) and distal interphalangeal (DIP) joints is considered to be the most common cause of hand pain nowadays. To our best knowledge, there is no published research, which in details addresses unclear genetic etiology of the finger OA. Since cartilage is one of the most commonly defected tissue in OA, the aim of our study was to explore gene expression profile of chondrocites sampled from two finger joints: PIP and DIP, and to investigate which pathways and gene ontology terms were altered in patients affected by this disease.

DISCOVERY AND PRIORITIZATION OF DE NOVO MUTATIONS IN AUTISM SPECTRUM DISORDER

Taeyeop Lee, Jaeho Oh, MinGyun Bae, Jun Hyeong Lee, Jung Kyoon Choi

Department of Bio Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired social-interaction, and restricted and repetitive behaviors. Previous studies have reported that the genetic contribution or heritability is as high as 80% in ASD. In order to elucidate the genetic architecture of ASD, many researchers performed extensive studies and discovered some significant findings. Currently, hundreds of different genes have been unveiled, mostly through identification of related rare variants. Rare genetic variants, both inherited and de novo, are proposed to be causal in ~30% of ASD patients. In comparison, common genetic variants also are estimated to contribute to approximately 50% of ASD etiology. However, no specific common risk variant has been found to date, possibly due to insufficient sample size. Here, we report a whole genome sequencing study of ASD patients to discover and characterize de novo mutations in Asian population. By sequencing 101 autism trios and unaffected siblings, we located causal variants in 74 candidate genes. The variants included not only loss of function and missense variants, but also intronic and intergenic non-coding variants. The candidate gene set showed significant overlap with known autism, intellectual disability, and chromatin related gene set. Furthermore, to prioritize the non-coding de novo mutations, we developed a deep learning framework based on >2,000 functional features. The features included DNase I hypersensitive sites, histone modification profiles, disease pathways, and transcription factor binding sites, where the nonlinear combinations of the features indicate the causal probability of a non-coding variant. The performance of the model was evaluated with area under curve (AUC) and F1 score. Our results suggest that de novo variants are related to important ASD risk genes, and that noncoding de novo variants have a non-zero effect in ASD.

CROSSTALKER: AN OPEN NETWORK AND PATHWAY ANALYSIS PLATFORM

Sean Maxwell, Mark R. Chance

Case Western Reserve UnIversity and Neo Proteomics, Cleveland, Ohio

Introduction: Network analysis methods have become commonplace research tools due to their proven ability to interrogate and organize lists of molecular targets of interest identified by basic statistics alone, and use of network analysis to refine classifier feature sets has been shown to provide superior performance compared to targets identified singly. We introduce Crosstalker as a freeware platform for academic use that is web based and incorporates multiple public interaction and gene set databases to perform network analysis, enrichment testing and visualization in a modern HTML5+JS interface. The use of open databases and algorithms coupled to convenient user choices allows cross comparison of findings and permits easy replication of results by any laboratory improving reproducibility and rigor. Methods: Lists of seed molecules are mapped onto a reference interaction network selected by the user and a random walk with restarts (RWR) is performed using the seed molecules as the restart nodes. The RWR scores are adjusted to z-scores using Monte-Carlo estimated score distributions for each node in the interaction network, and we have optimized the Monte-Carlo estimation parameters using analytic methods and computational testing. Assuming the z-scores follow a normal distribution, the adjusted scores are used to select nodes that have a $p < 0.001$ chance of achieving the same or higher RWR score by chance as they do from the user input. The resulting molecules are tested for enrichments against user selected gene set databases and used to induce result subnetworks from the reference network. The induced subnetworks are visualized with options to annotate nodes (molecules) and edges (interactions). Results: Computational experiments using inputs generated by combining annotated sets of functionally related molecules with unrelated “noise molecules” showed that adjusting proximity scores by null-distribution improved predictions of functionally related molecules over rank-only methods when the inputs contained more noise molecules than annotated molecules. Choices of multiple interaction networks (like BioGRID, BioPlex or COXPRESdb) enable testing of different hypotheses within the same interface, such as co-expression or direct/indirect physical interactions of related molecules. The optimized algorithms used by the computational portion of the software facilitate analysis times under 1 minute, minimizing wait times and maximizing the number of concurrent users the system can support. Novel Aspect: Analytically verified Monte-Carlo estimation parameters. Multiple options for interaction networks and gene sets. Web-based with options to export results and data in open (JSON, CSV) and binary (XLSX) formats. Free for academic use

SIGNATURES OF NON–SMALL-CELL LUNG CANCER RELAPSE PATIENTS: DIFFERENTIAL EXPRESSION ANALYSIS AND GENE NETWORK ANALYSIS

Abigail E. Moore[1], Brandon Zheng[2], Patricia M. Watson[3], Robert C. Wilson[3], Dennis K. Watson[3], Paul E. Anderson[4]

[1]Department of Natural Science, Hampshire College, Amherst, MA 01002, USA; [2]Department of Biology, Bard College, Annandale-On-Hudson, NY 12504, USA; [3]Department of Pathology and Laboratory Medicine, Medical University of South Carolina, Charleston, NC 29425, USA; [4]Department of Computer Science, College of Charleston, Charleston, SC 29424, USA

Background Lung cancer is both the second most represented cancer diagnosis and the leading cause of cancer death within the United States. Despite the high occurrence of non–small-cell lung cancer (NSCLC), 30% to 55% of patients relapse after curative resection, and the 5-year relative survival rate is 15% to 21%. The high costs of cancer medication and cancer drug failures are impacted by biomarker programs, which help select patients who may benefit from a given drug. Methods NSCLC RNA samples were taken from 38 patients, and clinical outcomes were determined by the American College of Surgery Oncology Group. Of these patients, 20 were diagnosed as disease-free, and 18 as relapse patients within 3 years of surgical resection. RNA-Seq libraries were paired-end sequenced on HiScanSQ and HiSeq 2500 systems. Read quality was determined by FastQC, and adapters and low-quality reads were trimmed with Trimmomatic. Trimmed paired-end reads were aligned to the human genome (HG38, UCSC) with RSEM. Aligned reads were input into the R/Bioconductor EBSeq package to perform median normalization and differential expression analysis. Differentially expressed genes were analyzed for over-representation of protein complexes, gene ontology terms and pathways via ConcensusPathDB. Results Empirical Bayesian methods identified 122 differentially expressed genes (FDR < 0.05). Many lung cancer-related genes were recognized, such as BAMBI, CPS1, CD70, SHISA3, and WNT11. Also identified were novel genes with upregulated expression in relapse patients: LILRA2, ALOX12, TSPAN-11, and CADM3, which are involved in immune response, arachidonic acid metabolism, cell surface receptor signaling, and cell-cell adhesion, respectively. Novel genes with downregulated expression in relapse patients were identified, including MCCC1, MRGPRF, PRR4, and SLC7A14, which are associated with biotin metabolism, signal transduction, cell adhesion, and negative regulation of phosphatase activity, respectively. A hypergeometric test revealed over-representation of gene ontology terms for biological processes related to cancer development: positive regulation of cell proliferation (p = 4.66e-06), lipoxygenase pathway (p = 6.95e-05), and beta-amyloid metabolic process (p = 0.000531). Only one protein complex-based set was over-represented: G protein-coupled receptor ligand. Accordingly, six GPCR-related pathways were over-represented (p-values from 6.77e-05 to 0.000196). Over-representation of other cancer-related pathways were found and include prostaglandin synthesis and regulation (p = 8.8e-05), fluoxetine metabolism pathway (p = 0.000217), and arachidonic acid metabolism (p = 0.000243). Conclusions Identifying NSCLC patients at risk of recurrence is crucial in cancer research. Our analyses identified 122 differentially expressed genes among disease-free and relapse NSCLC patients, including known lung cancer-related genes and new candidate biomarker genes that are involved in the diverse processes related to NSCLC development. Future research in alternative splicing and the development of a predictive model based on our results could support a new method for identifying individual recurrence risk.

RANKING BIOLOGICAL FEATURES BY DIFFERENTIAL ABUNDANCE

Soumyashant Nayak, Nicholas Lahens, Eun Ji Kim, Gregory Grant

University of Pennsylvania

We often want to rank features by their differential abundance between two populations. In RNA-Seq for example, we obtain quantified values for tens of thousands of genes across a wide spectrum of expression intensities. A naive ranking by fold-change leads to several issues. One of them is the division-by-zero issue which happens when the change is from 0 to a positive quantity. This problem is usually dealt with by using a pseudo-count of 1. Fold changes from smaller numbers however can tend to dominate the top of ranking lists in case of discrete data like RNA-Seq. Therefore, one might wonder whether a change from 1 to 2 (fold change of 2) is to be considered more significant than a change from 100 to 190 (fold change of 1.9). We systematically study this issue at both theoretical and empirical levels. We conclude that in RNA-Seq data there is an optimal value of the pseudo-count which yields the best significance comparisons. We formulate the necessary foundational mathematics in terms of a philosophical axiomatic framework to enable the systematic exploration of the ranking problem. Additionally we demonstrate how the use of pseudo-counts actually integrates fold-change and difference and this observation can be used to obtain the advantages of both methods, while minimizing the disadvantages.

SYSTEMATIC ANALYSIS OF OBESITY ASSOCIATED VARIATIONS THROUGH MACHINE LEARNING BASED ON GENOMICS AND EPIGENOMICS

Jaeho Oh, Jun Hyeong Lee, Taeyeop Lee, MinGyun Bae, Jung Kyoon Choi

Department of Bio Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

Obesity, one of the major global health concerns, is a metabolic disorder resulting from both behavioral and heritable causes. Various solutions, such as diet, exericse, surgery and drug therapies, have been proposed but these failed to provide long-term effects. Many researchers performed genome-wide association studies (GWAS)to identify disease-associated genomic regions, but interpretation of the data poses great challenge. Numerous GWAS analysis studies report that FTO is the region most closely associated with obesity, but the mechanism remains unresolved. According to one recent paper, 'outside variants', defined as SNPs that are in weak LD with GWAS risk SNPs and influence target gene's regulatory circuitry in combination, should be further invesitgated. 'Outside variant' approach suggest that not only statistically significant GWAS SNPs but also other SNPs may be biologically meaningful. To develop an obesity-related model and unravel the mechanism through 'outside variant' approach, we used the imputed GWAS data of 14,122 subject with BMI information. To select functional epigenetic region, we used histone modification ChIP-seq data from adipocytes and obesity-associated tissues and extracted SNP set that is highly related to FTO. By performing regression between SNPs and FTO SNPs, we found SNPs with high explanatory-power for obesity in the functional epigenetic-region. Our results suggest that the 'outside variant' analysis, along with several epigenetic data, is a novel approach to discover a set of SNPs, including SNPs that appear statistically insignificant, that affect obesity.

SPARSE REGRESSION MODELING OF DRUG RESPONSE WITH A LOCALIZED ESTIMATION FRAMEWORK

Teppei Shimamura, Hideko Kawakubo, Hyunha Nam, Yusuke Matsui

Division of Systems Biology, Nagoya University Graduate School of Medicine, Japan

A major challenge in pharmacogenomic studies is differences in the clinical characterization of patients and their reactions, which makes it difficult to identify clinically meaningful gene-drug interactions and predict drug response for each patient. In this study, we consider a localized regression model for each sample to predict a drug response with a set of main effects and second-order interactions for oncogenic alterations for patients. We propose a sparse modeling of interactions with localized estimation framework (SMILE) for this task. We take a regularization approach to inducing strong hierarchy in the sense that an interaction coefficient can have a non-zero estimate only if both of corresponding main effect coefficients are non-zero. We incorporate two different constraints into the group lasso and the lasso within the framework of local likelihood, to determine the type of structure such as strong hierarchy and enhance sparsity on the interaction coefficients, which enable to generate an interpretable localized interaction model for each sample. It can be formulated as the solution to a convex optimization problem, which we use the alternating direction method of multipliers (ADMM) method for solving SMILE. We then demonstrate the performance of our proposed method in a simulation study and on a pharmacogenomic data set.

PDBMAP: A PIPELINE AND DATABASE FOR MAPPING GENETIC VARIATION INTO PROTEIN STRUCTURES AND HOMOLOGY MODELS

R. Michael Sivley[1], John A. Capra[2], William S. Bush[3]

[1]Department of Biomedical Informatics, Vanderbilt Genetics Institute, Vanderbilt University; [2]Department of Biological Sciences, Vanderbilt Genetics Institute, Vanderbilt University; [3]Department of Population and Quantitative Health Sciences, Institute for Computational Biology, Case Western Reserve University

Rare genetic variants identified from sequencing studies are often grouped by genes, functional domains, and other annotations to increase power in trait association tests and identify shared phenotypic effects. However, association tests rarely consider variants' orientation in their functional context—three-dimensional (3D) protein structures. Various tools have been developed for visualizing specific variants in the context of individual protein structures; however, these tools do not support a complete, systematic mapping of variants in identified in sequencing studies into all available solved and computationally predicted protein structures. We describe PDBMap, a computational pipeline to efficiently map human genetic variation generated by sequencing studies into the structome. We also present the complete mapping of missense variants from the 1000 Genomes Project, Genome Aggregation Database (gnomAD, N=3,010,061), Catalogue of Somatic Mutations in Cancer (COSMIC, N=1,104,417), ClinVar (N=56,235), and the Alzheimer's Disease Sequencing Project (ADSP, N=891,849) into solved protein structures from the Protein Data Bank (N=31,688) and computationally predicted homology models from ModBase (N=186,802). Source code is available from https://github.com/capralab/pdbmap and downloads are available at http://astrid.icompbio.net .

REPETITIVE RNA AND GENOMIC INSTABILITY IN HIGH-GRADE SEROUS OVARIAN CANCER PROGRESSION AND DEVELOPMENT

James R. Torpy[1], Nenad Bartonicek[1], David D. L. Bowtell[2], Marcel E. Dinger[1]

[1]Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst 2010, Sydney, Australia; [2]Peter MacCallum Cancer Centre, East Melbourne, Victoria 3002, Australia

Ovarian cancer is a highly complex disease with a range of different histological subtypes. This highly lethal disease is estimated to be the fifth most common cause of death from cancer in females, with a five-year relative survival rate of 46.2%. High-grade serous ovarian cancer (HGSOC), characterized by widespread genomic instability, accounts for 70-80% of ovarian cancer deaths, and survival rates have not improved significantly for the last few decades. Furthermore, the underlying cause of around 1/3 of HGSOC cases cannot be explained. Evidence suggests that RNA derived from repetitive regions of the genome plays a role in genomic instability and development of cancers such as high-grade serous ovarian cancer, and may play a role in the unexplained HGSOC cases. Aberrant expression of centromere-derived RNA causes dysfunctional chromosomal segregation during mitosis and aneuploidy. Telomere-derived RNA maintains telomeres, preventing chromosomal fusion, breakage and subsequent rearrangement of the chromosomes. Retrotransposable elements such as LINE1s and Alus insert into different genomic locations, disrupting sequences and causing rearrangements such as duplications, inversions and translocations. We have analysed over 120 HGSOC case and control RNA-sequencing data sets of primary samples from the Australian Ovarian Cancer Study, comparing differences in expression of repetitive RNA transcripts across multiple HGSOC subtypes and controls. We found a range of differentially expressed repetitive RNA species including LINE1, Alu and centromere-derived RNA which may be contributing to genomic instability in these tumours. In order to investigate the potential causes of the differences in repeat RNA levels, their expression was correlated with expression of a range of methyltransferases such as DNMT1 and DNMT3A-C that are known to regulate methylation at repetitive heterochromatin, controlling RNA expression from these regions. Expression of RNAi-associated factors such as Dicer was also assessed as these factors can contribute to repetitive RNA regulation.

DIMENSION REDUCTION OF GENOME-WIDE SEQUENCING DATA BASED ON LINKAGE DISEQUILIBRIUM STRUCTURE

Yun Joo Yoo[1], Suh-Ryung Kim[1], Sun Ah Kim[2], Shelley B. Bull[3]

[1]Department of Mathematics Education, Seoul National University; [2]Department of Statistics, Seoul National University; [3]Prosserman Centre for Health Research, The Lunenfeld-Tanenbaum Research Institute

Genetic association analysis using high-density genome-wide sequencing data consisting of single nucleotide polymorphism (SNP) genotypes can benefit from various dimension reduction strategies for several reasons. First, genome-wide significance level for individual SNP tests should be determined considering the correlation structure of genotype data. Adjustment for Type I error inflation due to multiple hypothesis testing can be sought based on the dimension reduction methods. Second, increased Type I error may be reduced as the number of variables in the analysis decreases by dimension reduction. Third, the computational burden can be reduced as the complexity of the analysis model is reduced. Fourth, the power of association test can be gained by combining multiple signals in a group as a result of the dimension reduction strategy. We developed a genome partitioning method by clustering SNPs into blocks based on linkage disequilibrium structure. The algorithm uses a graph modeling of communities of highly correlated SNPs and applies a clique partitioning algorithm to the graph to partition SNPs into blocks. We applied the algorithm to 1000 Genomes Project data, and obtained 162K, 173K, 334K blocks including singleton blocks in the autosomal regions of 22 chromosomes for Asian, European, and African data respectively. The average LD measure r^2 (the Pearson correlation coefficient of two additively coded genotype variables) values within blocks are 0.465, 0.437 and 0.329 for Asian, European, and African data whereas the average r^2 values between consecutive blocks are 0.156, 0.145, and 0.098 for three populations. We evaluated the Type I error and the power gain from these partitions for several multi-SNP association tests using the simulated data based on 1000 Genomes Project data. Compared to other clustering methods, several tests using local dimension reduction strategies combined with genome-wide dimension reduction showed better power than other methods. We also developed a local dimension reduction method for genome-wide sequencing data especially targeting the multi-collinearity issue of dense SNP genotype data to be analyzed by multiple regression analysis. This method clusters SNPs in multi-collinearity by examining the variance inflation factor (VIF), and replaces such group by principal components. The algorithm proceeds iteratively until all VIF values are under a threshold value. When we compared the power between the analysis based on original data and the analysis based on the dimension reduced data using VIF evaluation, we observed the power gain in quadratic-type tests such as Wald test.

THE MULTIPLE GENE ISOFORM TEST

Yao Yu, Chad D. Huff

Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Gene-based association tests aggregate multiple variants in a gene to evaluate statistical evidence for rare variant association. Typically, these tests include variants from all coding exons in a gene, irrespective of gene isoform. For genes with multiple isoforms, this is often approximately equivalent to a test of the largest isoform, which is not necessarily optimal. Because smaller isoforms tend to be enriched for the core functional domains of a gene, they may also be enriched for pathogenic variants or larger variant effect sizes. To address the opportunities presented by isoform-specific patterns of disease susceptibility, we introduce the Multiple Gene Isoform Test (MGIT). MGIT employs a permutation approach to test each isoform of a gene, summarizing the contribution of each transcript to calculate a single gene-level p-value, without the need to explicitly model correlation between transcripts. MGIT can be applied in conjunction with any gene-based association test to assess gene-level significance and to identify isoforms that may be enriched for protein domains impacting disease risk. To demonstrate the utility of MGIT, we report results from a gene-based association test (VAAST) involving 783 breast cancer cases, 322 skin cutaneous melanoma cases, and 3,607 controls of European ancestry. For two established cancer genes, we observed a two-fold and three-fold reduction in p-value with MGIT relative to a whole-gene test, for MITF in melanoma and BRCA1 in breast cancer, respectively. In contrast, for other established cancer genes, we observed either no change in p-value (RAD51B and BRCA2 in breast cancer and MC1R, MTAP, and BRCA2 in melanoma) or a modest attenuation of association signal (CHEK2 in breast cancer). In the case of BRCA1, the difference in the MGIT association signal was primarily driven by rare, predicted damaging missense variants, which exhibited large differences in effect size between the smallest and largest isoforms. MGIT is implemented in the software package XPAT, with support for VAAST, SKAT-O, and 27 additional gene-based association tests.

IMAGING GENOMICS

POSTER PRESENTATIONS

GENETIC ANALYSIS OF CEREBRAL BLOOD FLOW IMAGING PHENOTYPES IN ALZHEIMER'S DISEASE

Xiaohui Yao[1], Shannon L. Risacher[2], Kwangsik Nho[2], Andrew J. Saykin[2], Heng Huang[3], Ze Wang[4], Li Shen[2]

[1]School of Informatics and Computing, Indiana University, Indianapolis; [2]Department of Radiology and Imaging Sciences, Indiana University School of Medicine; [3]Department of Electrical and Computer Engineering, University of Pittsburgh; [4]Department of Radiology, Lewis Katz School of Medicine, Temple University

Cerebral blood flow (CBF) provides a means to assess the neuronal and neurovascular consequences of Alzheimer's disease (AD) pathology. Both AD specific and non-specific CBF changes may be driven by unique or common genetic factors. To identify genetic variants associated with AD pathogenesis, we performed a targeted analysis to examine association between 4,033 SNPs of 24 AD candidate genes and CBF phenotypes measured by arterial spin labeling (ASL) magnetic resonance imaging (MRI) in four brain regions of interest (ROIs) including left angular, right angular, left temporal and right temporal gyri. Participants include 258 non-Hispanic Caucasian subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Targeted genetic association analysis of CBF on each ROI was tested using linear regression under an additive genetic model in PLINK, where age, gender and APOE ε4 status were included as covariates. Post-hoc analysis used Bonferroni correction for adjusting both the genetic and CBF measures. GATES was used to calculate gene-level p-values. The additive effects of the identified genetic variants from the above association analysis were also assessed at each voxel using SPM12 under one-way ANOVA test with age, gender and APOE ε4 status as covariates. The single nucleotide polymorphism (SNP) level analysis identified a novel locus in INPP5D (inositol polyphosphate-5-phosphatase D) significantly associated with left angular gyrus (L-AG) CBF. In gene-based analysis, both INPP5D and CD2AP (CD2 associated protein) were associated with L-AG CBF. The discovered INPP5D locus explained 8.29% variance of left angular CBF after adjusting for age, gender and APOE ε4 status. Further analyses on an independent subset of the ADNI samples (N=906) revealed that the minor allele of the locus was associated with lower cerebrospinal fluid t-tau/Aβ1-42 ratio. INPP5D functions as a negative regulator in immune system and a number of inflammatory responses, and has been found related to inhibit TREM2 signaling. The identified CBF risk factor has the potential to provide novel insights for better revealing the complex molecular mechanisms of AD. It warrants further investigation whether the risk factor is associated with the AD pathophysiology, the vascular pathophysiology, and/or their interaction.

PBRM1 MUTATIONS ARE ASSOCIATED WITH TISSUE MORPHOLOGICAL CHANGES IN KIDNEY CANCER

Jun Cheng[1], Jie Zhang[2], Zhi Han[2], Liang Cheng[2], Qianjin Feng[1], Kun Huang[2]

[1]Southern Medical University, [2]Indiana University School of Medicine

Background: Clear cell renal cell carcinoma (CCRC) is the most common kidney cancer. With the accumulation of large scale genomic data, genes with mutations that are common to CCRC patients have been identified. For instance, VHL has mutations in almost 49.9% of the CCRC patients in The Cancer Genome Atlas (TCGA) project followed by PBRM1, MUC4 and SETD2. While some of these genes have been established as driver genes for CCRC (e.g., VHL and SETD2), the functional implications of their mutations are still being characterized. Previous studies often focused on the effects of the mutations on molecular levels such as gene/microRNA expression and DNA methylation. In this study we aim to characterize the morphological changes at cellular and tissue levels associated with these mutations. Methods: Mutational status and histopathological imaging data for 448 CCRC patients were obtained from TCGA through the NCI Genomic Data Commons. There are six genes with mutations in more than 7% of the patients, they are VHL, PBRM1, MUC4, SETD2, BAP1, and MTOR. The imaging features were then extracted using computational pipeline we have previously developed. Our pipeline consists of three steps: nucleus segmentation, cell-level feature extraction, and aggregating cell-level features into patient-level features. Ten types of cell-level features were extracted including nuclear area (area), lengths of major and minor axes of cell nucleus and their ratio (major, minor, and ratio), mean pixel values of nucleus in RGB three channels respectively (rMean, gMean, and bMean), and mean, maximum, and minimum distances (distMean, distMax, and distMin) to neighboring nuclei in Delaunay triangulation graph. At last, all cell-level features from the same patient were aggregated into patient-level features using a bag-of-visual-words model with K-means (K=10) algorithm for learning words. Five additional parameters were calculated for each type of features - mean, standard deviation, skewness, kurtosis, and entropy. Thus there are 150 image features in total. For each selected gene, the features were compared between patients with and without mutations using Mann-Whitney-U tests. Results: While there are imaging features with p-value less than 0.05 for every gene, multiple test compensation (BH FDR) suggested that only PBRM1 mutations are associated with significantly different imaging features (69 features with q-value<0.05). Among them 'distMax_bin2', 'distMin_bin3', 'ratio_bin9' show significantly increases in the mutation group while 'distMean_std', 'major_std' and 'ratio_std' show significant decreases. Discussion and Conclusion: The above results suggest that tumor cells in the patients with PBRM1 mutations are more compact and their nuclei shapes are more homogeneous and closer to a round shape. These results are consistent with visual inspection and previous report that PBRM1 mutation leads to decrease of extracellular matrix gene expression and thus a reduction of stroma.

IMAGE GENOMICS OF INTRA-TUMOR HETEROGENEITY USING DEEP NEURAL NETWORKS

Hui Qu[1], Subhajyoti De[2], Dimitris Metaxas[1]

[1]Rutgers University, [2]Cancer Institute of New Jersey

Intra-tumor heterogeneity i.e. genetic, molecular, and phenotypic differences between tumor cells within a single tumor is a major challenge for clinical management of cancer patients, contributing to therapeutic failure, disease relapses and drug resistance. While recent findings suggest that there is extensive intra-tumor genetic heterogeneity in all major cancer types, it remains to be understood how that relates to intra-tumor heterogeneity at the pathway- and cell phenotype level. We have developed an innovative computational framework based on neural networks to identify cellular features from histological slides and then associate them with genomic and pathway-level features in a multi-scale model, before applying it to a cohort of 469 bladder cancer samples which has genomic, transcriptomic, pathway, and histological imaging data. In brief, our method first uses a Tumor Segmentation Network (TSN) and Nuclei Segmentation Network (NCN) to identify tumor cells regions and tumor nuclei in the histological slides. For tumor segmentation, we firstly extracted tumor and normal patches from the whole slide images of 40 patients, then trained a TSN to classify any patch into tumor or normal. Given any other whole slide image, the trained model can identify all tumor patches, which forms the tumor regions after morphological operations. The segmented tumor regions and nuclei are then used to compute q-statistic, and also alpha and beta diversity measures which reflect extent of local and regional intra-tumor phenotypic heterogeneity. Benchmarking against pathologically curated estimates indicates that this approach has high accuracy in identifying tumor cell features in a heterogeneous tumor. We then integrate imaging and genomics data to predict aspects of phenotypic heterogeneity based on cancer-related mutations and gene expression using uni- and multivariate approaches such as Relation Network (RN). Our preliminary results are consistent with biological knowledge. For example, we estimated the number of subclones in each tumor based on mutation data, and observed that indeed the samples with a high number of subclones have high phenotypic heterogeneity scores. We also estimated mRNA expression level of Ki67, a marker of cell growth and observed that the samples with higher q-statistic also had higher Ki67 expression, suggesting that certain patterns of intra-tumor heterogeneity correlate with tumor cell growth rates. Multi-scale analysis integrating genetic, pathway- and phenotypic heterogeneity will provide fundamental insights into "functional" variability within and across cancers, helping to refine precision medicine approaches to improve clinical management of cancer patients.

THE NEUROIMAGING INFORMATICS TOOLS AND RESOURCES COLLABORATORY (NITRC) AND ITS IMAGING GENOMICS DOMAIN

Li Shen[1], David Kennedy[2], Christian Haselgrove[2], Abby Paulson[3], Nina Preuss[3], Robert Buccigrossi[3], Matthew Travers[3], Albert Crowley[3], and The NITRC Team[3]

[1]Department of Radiology and Imaging Sciences, Indiana University School of Medicine; [2]Department of Psychiatry, University of Massachusetts Medical School; [3]TCG, Inc.

Aim of Investigation: Neuroimaging Informatics Tools and Resources Collaboratory (NITRC) is a neuroinformatics knowledge environment for MR, PET/SPECT, CT, EEG/MEG, optical imaging, clinical neuroinformatics, computational neuroscience, and imaging genomics tools and resources. We encourage researchers to list their Imaging Genomics tools at the NITRC website www.nitrc.org. Methods: Initiated in 2006 through the NIH Blueprint for Neuroscience Research, NITRC's mission is to foster a user-friendly knowledge environment for the neuroinformatics community. In 2012, NITRC added Imaging Genomics to its broadened scientific scope. By continuing to identify existing software tools and resources valuable to this community, NITRC's goal is to support its researchers dedicated to enhancing, adopting, distributing, and contributing to the evolution of neuroinformatics analysis software, data, and compute resources. Results: Located on the web at www.nitrc.org, the Resources Registry (NITRC-R) promotes software tools and resources, vocabularies, test data, and databases, thereby extending the impact of previously funded, neuroimaging informatics contributions to a broader community. NITRC-R gives researchers greater and more efficient access to the tools and resources they need, better categorizing and organizing existing tools and resources, facilitating interactions between researchers and developers, and promoting better use through enhanced documentation and tutorials—all while directing the most recent upgrades, forums, and updates. As of 11/2017, over 970 public resources are listed on NITRC-R, where the Imaging Genomics domain includes 60 resources such as ADNI, TCGA, ENIGMA, UK Biobank, and others. NITRC-Image Repository (NITRC-IR) makes 8,285 imaging sessions publicly available at no charge, and NITRC Computational Environment (NITRC-CE) provides cloud-based computation services downloadable to your machines or via commercial cloud providers such as Amazon Web Services and Microsoft Azure. Conclusions: In summary, NITRC is now an established knowledge environment for the neuroimaging community where tools and resources are presented in a coherent and synergistic environment. With its expanded scope into imaging genomics, NITRC aims to become a trusted source for identification of resources in this highly active and promising domain bridging advanced neuroimaging and genomics. We encourage the imaging genomics research community to continue providing valuable resources, design and content feedback and to utilize these resources in support of data sharing requirements, software dissemination and cost-effective computational performance. Acknowledgements: Funded by the NIH Blueprint for Neuroscience Research, NIBIB, NIDA, NIMH, and NINDS.

IDENTIFYING THE GIST OF CNNS: FINDING INTERPRETABLE SIGNATURES OF HISTOLOGY IMAGE MODELS BUILT USING NEURAL NETWORKS

Arunima Srivastava[1], Chaitanya Kulkarni[1], Kun Huang[2], Parag Mallick[3], Raghu Machiraju[1]

[1]The Ohio State University, [2]Indiana University School of Medicine, [3]Stanford University

Convolutional Neural Networks (CNNs) have gained steady popularity as the selected method of histology image analysis and subsequent disease modeling. Since CNNs are purely data driven learning models, they have an edge over morphology driven (pre-selected) tissue image features that may be biased and difficult to generalize. Morphological features, namely tissue texture, structure, nuclei size and shape, presence of fibroblasts and lymphocytes etc., might not be comprehensive enough for different datasets, but they do provide an inherently interpretable characterization of the histology. While CNNs and their subsequent features prove to be powerful classifiers, they fail to provide an explanation for this classification, as the features are ONLY interpretable by the CNNs themselves. Translating "under the hood" activities of a CNN would endeavor to make it more generalizable while the final model will not only be able to effectively classify whole slide tissue images, it will also have the potential to educate us on the nuances of the histological data. This work aims to use both types of interpretable (morphological) and powerful but un-interpretable (CNN based) features to derive a signature for successful CNN models, which help relate them to known biological attributes and shed light on components that are critical to the various subtypes under investigation. We use a stratified breast cancer histology classification dataset from the BioImaging (2015) Challenge that contains sample images from four different kinds of breast tissue (Normal, Benign lesion, In-situ carcinoma and Invasive carcinoma). By following a two-pronged approach of modeling the same dataset using CNNs (using the GoogLeNet architecture) and morphological features (using CellProfiler - a biological image analytics tool), it was possible to infer an interpretable signature of features utilized by the CNN. We additionally explore the possibility of combining these two techniques to extract a more powerful and precise classification. This work summarizes the need for understanding the widely trusted models built using deep learning, and adds a layer of biological context to a technique that functioned as a classification only approach till now.

PRECISION MEDICINE: FROM DIPLOTYPES TO DISPARITIES TOWARDS IMPROVED HEALTH AND THERAPIES

POSTER PRESENTATIONS

EXPLORING THE POTENTIAL OF EXOME SEQUENCING IN NEWBORN SCREENING

Steven E. Brenner[1], Aashish N. Adhikari[1], Yaqiong Wang[1], Robert J. Currier[2], Renata C. Gallagher[3], Robert L. Nussbaum[4], Yangyun Zou[1], Uma Sunderam[5], Joseph Sheih[3], Flavia Chen[3], Mark Kvale[3], Sean D. Mooney[6], Raj Srinivasan[5], Barbara A. Koenig[3], Pui Kwok[3], Jennifer M. Puck[3], The NBSeq Project

[1]University of California - Berkeley, [2]California Department of Public Health, [3]University of California - San Francisco, [4]Invitae, [5]Tata Consultancy Services, [6]University of Washington

The NBSeq project is evaluating effectiveness of whole exome sequencing (WES) for detecting inborn errors of metabolism (IEM) for newborn screening (NBS). De-identified archived dried blood spots from MS/MS true positive and false positive cases previously identified in the California NBS were studied. 18 out of 137 affected individuals lacked two rare potentially damaging single nucleotide variants or short indels in genes responsible for their Mendelian disorders. The sensitivity of causal mutation detection in 137 Phase I NBSeq exomes varied across disorders; all affected PKU cases were predicted correctly, but several cases of other IEMs were missed. In some cases, exomes also confidently identified disorders different from the metabolic center diagnoses, suggesting that sequencing information would have been valuable for proper clinical diagnoses in those cases. Deeper analysis of the data was undertaken to assess sources of discrepancy between sequencing results, MS/MS call, and clinical diagnosis. Copy number variation (CNV) calling tools were evaluated on NBSeq exomes for ability to resolve some of these exome false negatives. CNV tools can both miss CNVs in exomes and report them spuriously. We optimized tools for our data and filtered out genes (PRODH, HCFC1, ETFA) harboring common CNVs (identified from CNV calls on the 1000 genomes project exomes). This identified deletions in the correct genes for 4 of the 32 exome false negatives using XHMM: 2 isovaleric acidemia cases, 1 methylmalonic acidemia case and 1 OTC deficiency case. We also systematically reviewed every variant in 78 metabolic disorder genes annotated by HGMD or ClinVar as pathogenic or likely pathogenic with 1000 genomes MAF > 0.1%. Our re-assessment of the primary literature for 59 such variants found that only 18 were reportable (many still VUS) and the rest we excluded from the pipeline. Literature review also helped identify 8 cases diagnosed with short-chain acyl-CoA dehydrogenase (SCAD) deficiency but not flagged by exomes. All 8 individuals harbored a common (1000 Genomes MAF: 18.2%) ACADS allele (c.625A>G) present in several NBSeq exomes, which sometimes confers a partial biochemical phenotype but not clinical disease. For assessment, we treated these individuals as unaffected. Incorporation of CNV detection and variant curation into our analysis pipeline improved overall sensitivity from 77.9% to 87.6% on the 137 affected Phase I NBSeq samples. This updated pipeline will be run on additional NBSeq exomes to assess the potential role for WES in NBS. While still not sufficiently specific alone for screening of most IEMs, WES can facilitate timely and more precise case resolution.

A METHOD FOR IMPROVED VARIANT CALLING AT HOMOPOLYMER MARGINS (AND ELSEWHERE)

J. Buckley, M. Hiemenz, J. Biegel, T. Triche, A. Ryutov, D. Maglinte, D. Ostrow, X. Gai

Center for Personalized Medicine, Children's Hospital of Los Angeles

All sequencing technologies are subject to read errors which, in the context of variant calling (particularly low variant-allele-frequency (VAF) variant calling), can yield miscalls. Read errors are most problematic when genomic context (such as proximity to homopolymers) influences the error rate The Center of Personalized Medicine at Children's Hospital of Los Angeles (CHLA) recently collaborated with Thermo-Fisher (TF) in development of a clinical pediatric cancer panel for somatic variant detection (OncoKidsTM), using TF's Ion Torrent sequencing platform. The test needed to identify variants in tumor sub-clones and in samples with an admixture of tumor and normal cells, both situations that can yield low VAFs. Our challenge was to optimize variant calling at homopolymer margins, and other genomic loci with a high background error rate (noise). The TF approach was to identify problematic loci and to either limit base calls to reads from one strand (when errors clustered mostly on the other strand), or 'blacklist' the locus altogether. While this approach was conservative, avoiding most false positives, it resulted in unacceptable false negative rates, particularly for InDels. Given the deep coverage (over 1000x in many regions), it seemed likely that a more nuanced approach might yield accurate calls, even in the presence of substantial noise. This presentation outlines an algorithm (Local Adjustment for Background, or LAB) developed at CHLA that uses a reference data set (filtering out true positives) to establish the noise distribution at each locus. The noise distribution varies greatly across the panel genes, from essentially error-free loci to loci in which the majority of reads show a spurious base substitution or InDel. While proximity to a homopolymer is a strong determinant of noise, non-homopolymer regions can also have high noise and many homopolymers yield relatively clean data. Variant calls are made through comparison of the observed VAF with the locus-specific VAF distribution in the reference. Optionally, the reference set can be limited to samples of the same type as the test sample (e.g. FFPE). Adjustments may be made for samples with globally increased error rates. In regions of complex InDel patterns, a statistical model tests for shifts in these patterns, indicative of a true variant. An important component is a GUI that provides a visual representation of the basis for a call, and options such as strand-specific analysis. Application to samples with known SNVs and InDels (Acrometrix 'ground truth' samples) resulted in improvement in InDel calls from 65% to 100%. The presentation will describe the calling pipeline, with illustrative examples, and present comparative performance data.

EFFICIENT SURVIVAL MULTIFACTOR DIMENSIONALITY REDUCTION METHOD FOR DETECTING GENE-GENE INTERACTION

Jiang Gui, Xuemei Ji, Christopher I. Amos

Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth, Lebanon, NH 03756

The problem of identifying SNP-SNP interactions in case-control studies has been studied extensively and a number of new techniques have been developed. Little progress has been made, however in the analysis of SNP-SNP interactions in relation to censored survival data. We present an extension of the two class multifactor dimensionality reduction (MDR) algorithm that enables detection and characterization of epistatic SNP-SNP interactions in the context of survival outcome. The proposed an Efficient Survival MDR (ES-MDR) method handles censored data by modifying MDR's constructive induction algorithm to use logrank Test.

We applied ES-MDR to genetic data of over 470,000 SNPs from the OncoArray Consortium. We use onset age of lung cancer and case-control (n=27,312) status as the survival outcome and divided data into training and testing sets. We also adjust for subject's age, gender and smoking status. From training set, we identified interation between SNPs from BRCA1 and IL17RC genes as the top model that is assciated with lung cancer onset age. This result is validated in the testing set. ES-MDR is capable of detecting interaction models with weak main effects. These epistatic models tend to be dropped by traditional regression approaches.

BIOINFORMATICS PROCESSING STRATEGIES FOR EFFICIENT SEQUENCING DATA STORAGE USING GVCF BANDING

Nicholas B. Larson, Shannon K. McDonnell, Iain F. Horton, Saurabh Baheti, Jeanette E. Eckel-Passow, Steven N. Hart

Mayo Clinic

An emerging challenge in the era of next-generation sequencing (NGS) is efficient data storage practices, particularly for file formats that accommodate ad hoc construction of analysis-ready datasets. The Variant Call Format (VCF) is the predominant file type used for storing and analyzing NGS-based genetic variant information. However, it presents multiple practical limitations when merging individual files for multi-sample representations. Recent development of the gVCF file format by GATK addresses many of these concerns by characterizing same-as-reference segments of the genome as interval entries defined by a shared genotype quality (GQ) score. Current default settings to generate this intermediate file format result in a new data entry at each basepair position the GQ shifts, presenting cost-benefit considerations of improved and computationally efficient multi-sample genotyping at the expense of large intermediate files. However, additional options allow for contiguous entries to be merged if they fall within a predefined GQ bin, a process known as banding. We hypothesized that substantial gVCF filesize reduction could be attained for whole-genome sequencing (WGS) through the use of coarse GQ banding options; although the impact of this approach on output quality of multi-sample variant calling is currently unknown. To investigate the properties of gVCF banding on genotyping integrity, we processed 50 WGS samples as well as 50 whole-exome sequencing (WES) samples from the Mayo Clinic Biobank under a variety of GQ banding settings (default, intervals of 10, {0,20,60}, {0,20}). These single-sample gVCF files were subsequently merged and joint genotyped under varying combinations of banding options, separately by sequencing application, and output genotypes for chromosome 22 were compared for concordance with results using complete information (i.e., no banding). Overall, WGS samples exhibited substantially smaller gVCF files, with {0,20} banding resulting in a mean filesize reduction of 87% (range: 84-90%) relative to default settings. Genotype concordance exceeded 99.9% under all comparisons, while we additionally observed more variable positions emitted as coarser bin definitions were applied. Comparable findings were observed for WES data. Our results highlight impressive improvements in NGS variant call data storage efficiency gained by coarse banding options for gVCF output, with minimal impact on accompanying genotyping quality.

IDENTIFICATION OF A NOVEL TSC2 MUTATION IN A PATIENT WITH TUBEROUS SCLEROSIS COMPLEX

Jae-Hyung Lee[1], Su-Kyeong Hwang[2], Jung-eun Yang[3], Chae-Seok Lim[3], Jin-A Lee[4], Kyungmin Lee[5], Bong-Kiun Kaang[3], Yong-Seok Lee[6]

[1]Kyung Hee University, [2]Kyungpook National University Hospital, [3]Seoul National University, [4]Hannam University, [5]Kyungpook National University Graduate School of Medicine, [6]Seoul National University College of Medicine

Tuberous sclerosis complex (TSC) is a neurocutaneous disorder characterized by multiple symptoms including neuropsychological deficits such as seizures, intellectual disability, and autism. TSC is inherited in an autosomal dominant pattern and is caused by mutations in either the TSC1 or TSC2 genes, which result in the hyperactivation of the mammalian target of rapamycin (mTOR) signaling pathway. In this study, we identified a novel small deletion mutation in TSC2 by performing whole exome sequencing in a Korean patient, who exhibited multiple TSC-associated symptoms including frequent seizures, intellectual disability, language delays, and social problems. In addition, we validated the functional significance of the novel mutation by examining the effect of the deletion mutant on mTOR pathway activation. Recent studies have suggested that mTOR inhibitors such as rapamycin can be effective to treat TSC-associated deficits in rodent models of TSC. Accordingly, we found that everolimus treatment has beneficial effects on SEGA size and autism related behaviors in the patient.

CONSIDERATIONS FOR AUTOMATED MACHINE LEARNING IN CLINICAL METABOLIC PROFILING: ALTERED HOMOCYSTEINE PLASMA CONCENTRATION ASSOCIATED WITH METFORMIN EXPOSURE

Alena Orlenko[1], Jason H. Moore[1], Patryk Orzechowski[1,2], Randal S. Olson[1], Junmei Cairns[3], Pedro J. Caraballo[3], Richard M. Weinshilboum[3], Liewei Wang[3], Matthew K. Breitenstein[1]

[1]University of Pennsylvania; [2]AGH University of Science and Technology, Krakow, Poland; [3]Mayo Clinic

With the maturation of metabolomics science and proliferation of biobanks, clinical metabolic profiling is an increasingly opportunistic frontier for advancing translational clinical research. Automated Machine Learning (AutoML) approaches provide exciting opportunity to guide feature selection in agnostic metabolic profiling endeavors, where potentially thousands of independent data points must be evaluated. In previous research, AutoML using high-dimensional data of varying types has been demonstrably robust, outperforming traditional approaches. However, considerations for application in clinical metabolic profiling remain to be evaluated. Particularly, regarding the robustness of AutoML to identify and adjust for common clinical confounders. In this study, we present a focused case study regarding AutoML considerations for using the Tree-Based Optimization Tool (TPOT) in metabolic profiling of exposure to metformin in a biobank cohort. First, we propose a tandem rank-accuracy measure to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors. Second, while AutoML, using default parameters, demonstrated potential to lack sensitivity to low-effect confounding clinical covariates, we demonstrated residual training and adjustment of metabolite features as an easily applicable approach to ensure AutoML adjustment for potential confounding characteristics. Finally, we present increased homocysteine with long-term exposure to metformin as a potentially novel, non-replicated metabolite association suggested by TPOT; an association not identified in parallel clinical metabolic profiling endeavors. While warranting independent replication, our tandem rank-accuracy measure suggests homocysteine to be the metabolite feature with largest effect, and corresponding priority for further translational clinical research. Residual training and adjustment for a potential confounding effect by BMI only slightly modified the suggested association. Increased homocysteine is thought to be associated with vitamin B12 deficiency – evaluation for potential clinical relevance is suggested. While considerations for clinical metabolic profiling are recommended, including adjustment approaches for clinical confounders, AutoML presents an exciting tool to enhance clinical metabolic profiling and advance translational research endeavors.

PharmGKB: NEW WEBSITE RELEASE 2017

Michelle Whirl-Carrillo[1], Ryan M. Whaley[1], Mark Woon[1], Katrin Sangkuhi[1], Li Gong[1], Julia Barbarino[1], Caroline Thorn[1], Rachel Huddart[1], Maria Alvarellos[1], Jill Robinson[1], Russ B. Altman[2], Teri E. Klein[3]

[1]Department of Biomedical Data Science, Stanford University; [2]Department of Bioengineering, Medicine and Genetics, Stanford University; [3]Department of Biomedical Data Science and Medicine, Stanford University

With PharmGKB is the largest publicly available resource for pharmacogenomics (PGx) discovery and implementation. Its mission is to collect, curate, integrate and disseminate knowledge about how human genetic variation influences drug response. The PharmGKB website allows users to select and view information via search, filter and browse options. Data is also available by direct download through the website and through the PharmGKB API.

PharmGKB launched a new and improved user interface in September 2017. The new website offers benefits such as a display that works on mobile and small screen devices, improved searching and filtering capabilities, and faster page load speeds. While the look of PharmGKB has changed, all the content that was available previously is still available, including:

- 5500 annotated genetic variants
- 14,000 curated peer-reviewed PGx articles
- 125 evidence-based pharmacokinetic and pharmacodynamics pathways
- 60 reviews of key PGx genes (very important pharmacogenes)
- 450 curated drug labels
- 90 gene-drug pairs with curated genotype-based drug dosing guidelines

The website features an online tutorial that users can access by following the screen prompts. For more information, please visit PharmGKB at http://www.pharmgkb.org.

READING BETWEEN THE GENES: COMPUTATIONAL MODELS TO DISCOVER FUNCTION AND/OR CLINICAL UTILITY FROM NONCODING DNA

POSTER PRESENTATIONS

NETWORK ANALYSIS OF PSEUDOGENE-GENE RELATIONSHIPS: FROM PSEUDOGENE EVOLUTION TO THEIR FUNCTIONAL POTENTIALS

Travis S. Johnson[1], Sihong Li[1], Johnathan R. Kho[2], Kun Huang[3], Yan Zhang[1]

[1]Ohio State University, [2]Georgia Institute of Technology, [3]Indiana University

Pseudogenes are fossil relatives of genes. Pseudogenes have long been thought of as "junk DNAs", since they do not code proteins in normal tissues. Although most of the human pseudogenes do not have noticeable functions, ~20% of them exhibit transcriptional activity. There has been evidence showing that some pseudogenes adopted functions as lncRNAs and work as regulators of gene expression. Furthermore, pseudogenes can even be "reactivated" in some conditions, such as cancer initiation. Some pseudogenes are transcribed in specific cancer types, and some are even translated into proteins as observed in several cancer cell lines. All the above have shown that pseudogenes could have functional roles or potentials in the genome. Evaluating the relationships between pseudogenes and their gene counterparts could help us reveal the evolutionary path of pseudogenes and associate pseudogenes with functional potentials. It also provides an insight into the regulatory networks involving pseudogenes with transcriptional and even translational activities. In this study, we develop a novel approach integrating graph analysis, sequence alignment and functional analysis to evaluate pseudogene-gene relationships, and apply it to human gene homologs and pseudogenes. We generated a comprehensive set of 445 pseudogene-gene (PGG) families from the original 3,281 gene families (13.56%). Of these 438 (98.4% PGG, 13.3% total) were non-trivial (containing more than one pseudogene). Each PGG family contains multiple genes and pseudogenes with high sequence similarity. For each family, we generate a sequence alignment network and phylogenetic trees recapitulating the evolutionary paths. We find evidence supporting the evolution history of olfactory family (both genes and pseudogenes) in human, which also supports the validity of our analysis method. Next, we evaluate these networks in respect to the gene ontology from which we identify functions enriched in these pseudogene-gene families and infer functional impact of pseudogenes involved in the networks. This demonstrates the application of our PGG network database in the study of pseudogene function in disease context.

RANDOM WALKS ON MUTUAL MICRORNA-TARGET GENE INTERACTION NETWORK IMPROVE THE PREDICTION OF DISEASE-ASSOCIATED MICRORNAS

Duc-Hau Le[1], Lieven Verbeke[2], Le Hoang Son[3], Dinh-Toi Chu[4], Van-Huy Pham[5]

[1]Vinmec Research Institute of Stem Cell and Gene Technology, 458 Minh Khai, Hai Ba Trung, Hanoi, Vietnam; [2]Department of Information Technology, Ghent University - imec, Ghent, Belgium; [3]VNU University of Science, Vietnam National University, Hanoi, Vietnam; [4]Faculty of Biology, Hanoi National University of Education, Hanoi, Vietnam; [5]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Background MicroRNAs (miRNAs) have been shown to play an important role in pathological initiation, progression and maintenance. Because identification in the laboratory of disease-related miRNAs is not straightforward, numerous network-based methods have been developed to predict novel miRNAs in silico. Homogeneous networks (in which every node is a miRNA) based on the targets shared between miRNAs have been widely used to predict their role in disease phenotypes. Although such homogeneous networks can predict potential disease-associated miRNAs, they do not consider the roles of the target genes of the miRNAs. Here, we introduce a novel method based on a heterogeneous network that not only considers miRNAs but also the corresponding target genes in the network model. Results Instead of constructing homogeneous miRNA networks, we built heterogeneous miRNA networks consisting of both miRNAs and their target genes, using databases of known miRNA-target gene interactions. In addition, as recent studies demonstrated reciprocal regulatory relations between miRNAs and their target genes, we considered these heterogeneous miRNA networks to be undirected, assuming mutual miRNA-target interactions. Next, we introduced a novel method (RWRMTN) operating on these mutual heterogeneous miRNA networks to rank candidate disease-related miRNAs using a random walk with restart (RWR) based algorithm. Using both known disease-associated miRNAs and their target genes as seed nodes, the method can identify additional miRNAs involved in the disease phenotype. Experiments indicated that RWRMTN outperformed two existing state-of-the-art methods: RWRMDA, a network-based method that also uses a RWR on homogeneous (rather than heterogeneous) miRNA networks, and RLSMDA, a machine learning-based method. Interestingly, we could relate this performance gain to the emergence of “disease modules” in the heterogeneous miRNA networks used as input for the algorithm. Moreover, we could demonstrate that RWRMTN is stable, performing well when using both experimentally validated and predicted miRNA-target gene interaction data for network construction. Finally, using RWRMTN, we identified 76 novel miRNAs associated with 23 disease phenotypes which were present in a recent database of known disease-miRNA associations. Conclusions Summarizing, using random walks on mutual miRNA-target networks improves the prediction of novel disease-associated miRNAs because of the existence of “disease modules” in these networks.

TEXT MINING AND VISUALIZATION FOR PRECISION MEDICINE

POSTER PRESENTATIONS

MINING ELECTRONIC HEALTH RECORDS FOR PATIENT-CENTERED OUTCOMES TO GUIDE TREATMENT PATHWAY DECISIONS FOLLOWING PROSTATE CANCER DIAGNOSIS

Selen Bozkurt[1], Jung In Park[2], Daniel L. Rubin[3], James D. Brooks[4], Tina Hernandez-Boussard[5]

[1]Akdeniz University Faculty of Medicine Department of Biostatistics and Medical Informatics Antalya, Turkey; [2]Stanford University Department of Medicine (Biomedical Informatics); [3]Stanford University Department of Radiology; [4]Stanford University Department of Urology; [5]Stanford University Department of Medicine (Biomedical Informatics)

Electronic health records (EHRs) have potential for novel discovery of patient-centered outcomes that can be used to improve health care delivery. However, a significant amount of data stored in EHRs is hidden in clinical narratives as unstructured text. For prostate cancer patients, these clinic narratives contain a large amount of information. Previous work suggests that structured data regarding dysfunctions after treatment for prostate cancer are not consistently captured in the EHR and thus cannot be reliably extracted for clinical and research purposes. Therefore, in this preliminary study we propose a rule-based natural language processing pipeline to extract patient-centered outcomes related to the presence of urinary, bowel and erectile dysfunction following treatment of prostate cancer from the free text of the EHR notes. We developed a lexicon of terms related to urinary, bowel or erectile dysfunctions based on domain knowledge, prior experience in the field, and review of medical notes. A reference standard of 100 randomly selected documents for each outcome from inpatient admissions was annotated by a research nurse to identify all related concepts as: present, negated, historical, and discussed risk. We developed a rule-based natural language processing (NLP) pipeline which uses dictionary mapping combined with ConText algorithm. We trained our NLP pipeline using 1,336 documents and tested on 20 documents to determine agreement with the human reference standard and standard precision, recall and overall accuracy rates were used as metrics to quantify the automatic annotation performance. The precision, recall, and accuracy scores for the urinary incontinence annotations against the reference standard output created by a domain expert was 62.5%, 100% and 76.9%, respectively. For most of the misclassified cases, which annotated as presence of urinary incontinence by the NLP algorithm but not by the expert, it is seen that medication information included in the term dictionary caused ambiguity regarding phenotype classification. For the erectile dysfunction annotations, precision was 100%, recall was 75% and overall accuracy was 90%. On the other hand, since any bowel dysfunction was reported in the randomly selected test set, evaluation metrics were not calculated. In this preliminary study, we have shown that it is possible to identify the patient-centered outcomes from the free text of EHRs using natural language processing. Using EHRs to assess patient-centered outcomes promotes population-based assessments of these valued yet difficult to assess outcomes and will enable detailed sensitivity and subgroup analysis. Such results will allow clinicians to individualize care for their patients. The results will also provide desperately needed evidence-based criteria for patient-centered outcomes. These criteria can be used in research studies, in clinical practice, and to develop practice guidelines. Future work will create larger number of well-annotated data sets and combine our rule-based approach with machine learning techniques.

GDMINER: A BIO TEXT MINING SYSTEM FOR GENE-DISEASE RELATION ANALYSIS

Soo Jun Park[1], Jihyun Kim[2], Soo Young Cho[2], Charny Park[2], Young Seek Lee[3]

[1]Electronics and Telecommunications Research Institute, [2]National Cancer Center, [3]Hanyang University

Researchers of Biology and Medicine often visit PubMed to find literatures for their studies. While the keyword search in PubMed may be a popular tool to retrieve information, it is limiting as it only provides a small number of results. The keyword search does not allow the user to sift through decades worth of research and extract all corresponding studies as needed. This poster presentation will provide solutions through a bio text mining system called GDMiner that identifies biological entities, extracts the relationship from those entities, and discovers associations between genes and diseases. When GDMiner collects abstracts from PubMed (PubMed collector), an automatic naming entity sorts the information into 40 biological categories (Entity Recognizer). GDMiner then extracts relations from the biomedical categories (Relation Extractor) by using natural language processing techniques, like Part-of-Speech (POS) tagging and syntactic parsing. The display features graphs and tables showing the extracted relations. For example, a gene-disease association data query can be mined by analyzing the relations between genes and diseases. The system consists of the following three parts: PubMed collector, relation extractor and relation analyzer. The PubMed collector asks abstracts with a query given by a user and fetches them. The relation extractor divides abstracts into sentences and recognizes biomedical named entities in sentences. Then, the relation analyzer extracts relational events among recognized entities. Relations are extracted by syntactic analysis not by co-occurrence information. Our system parses sentences syntactically in forms of the PennTreebank syntactic tags and extract relations by analyzing parsing results. Our rules are simple and small because the syntactic tag set have fewer number of tags than the POS tag set, but not limited to relation types. The relation viewer accumulates extracted relations and visualizes in graphs and tables. If the number of nodes in the generated relationship network is small, it is easy for the user to easily find the relationship between desired bio objects (named entities). However, if the size of the generated network is very large, it is very difficult to find the relations. Our system help user to find the relation between the desired bio objects by creating a small size sub-network using the search and filtering function. There is a rapidly growing interest in properly utilizing biomedicine literature within the research community and the rate in which the biomedicine literature is accumulating is accelerating worldwide. The importance of not only preserving data, but also the way in which researchers extract information is necessary in aiding future biological studies and discoveries. Implementing an automated system is necessary in keeping up with the growth and providing accuracy in finding analogous information to a researcher's search.

WORKSHOP

MACHINE LEARNING AND DEEP ANALYTICS FOR BIOCOMPUTING: CALL FOR BETTER EXPLAINABILITY

POSTER PRESENTATIONS

METHODS FOR EXAMINING DATA QUALITY IN HEALTHCARE INTEGRATED DATA REPOSITORIES

Vojtech Huser[1], Michael G. Kahn[2], Jeffrey S. Brown[3], Ramkiran Gouripeddi[4]

[1]National Library of Medicine, National Instittutes of Health 8600 Rockville Pk, Bld 38a Bethesda, MD, 20852, USA Email: vojtech.huser@nih.gov; [2]Department of Pediatrics, University of Colorado 13001 East 17th Place MS-F563 Aurora, CO 80045 USA Email: Michael.Kahn@ucdenver.edu; [3]Department of Ppopulation Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute 401 Park Drive, Suite 401 East Boston, MA 02215 USA Email: jeff_brown@hphc.orgjeff_brown@hphc.org; [4]University of Utah, School of Medicine Salt Lake City, 84102, Utah, USA Email: ram.gouripeddi@utah.edu

This paper summarizes content of the workshop focused on data quality. The first speaker (VH) described data quality infrastructure and data quality evaluation methods currently in place within the Observational Data Science and Informatics (OHDSI) consortium. The speaker described in detail a data quality tool called Achilles Heel and latest development for extending this tool. Interim results of an ongoing Data Quality study within the OHDSI consortium were also presented. The second speaker (MK) described lessons learned and new data quality checks developed by the PEDsNet pediatric research network. The last two speakers (JB, RG) described tools developed by the Sentinel Initiative and University of Utah's service oriented framework. The workshop discussed at the end and throughout how data quality assessment can be advanced by combining the best features of each network.

MULTI-CLASS CLASSIFICATION STRATEGY FOR SUPPORT VECTOR MACHINES USING WEIGHTED VOTING AND VOTING DROP

Sungho Kim, Taehun Kim

Yeungnam University, DGIST

A novel multi-class strategy for Support Vector Machines (SVMs) was developed to perform multi-class classification, such as One Versus One, One Versus All and Dynamic Acyclic Graph. These strategies do not reflect the distance between the hyper-plane that separates two classes and input data. This is not reasonable when the input data is placed near the hyper-plane. The proposed weighted voting resolves this problem by weighting the voting values according to the distance from the boundary and the enhanced performance of the SVMs with the proposed voting drop. The proposed Weighted Voting is based on the voting method. The voting method is carried out by accumulating votes, then choosing the most voted class. The proposed Weighted Voting method is a weighting of the voting value by reflecting the distance from the boundary and margin. Second proposed Voting Drop method is about how to accumulate votes. The novel voting method accumulates every vote but this manner can be a problem because there are redundantly responding SVMs. Because the SVM is a binary classifier, each SVM learns only about two classes. Therefore, a SVM does not have discernment for the non-learned classes. This is why when a SVM predicts data belonging to a non-learned class, the SVM responds redundantly. This irrelevant SVM causes an incorrect vote that makes the decision confused. To resolve this problem, the Voting Drop method drops the redundant votes by removing the irrelevant SVM. This algorithm finds the irrelevant SVM, then dropping the votes caused by the irrelevant SVM. The way to find an irrelevant SVM is to find a least voted class because a least voted class can be thought of as an irrelevant class to input data. As shown in the experiments, evenly reflecting the distance from the hyper-plane and the discernment of the hyper-plane and removing the redundant SVM`s voting leads to higher performance. The proposed methods can be used for a range of classification tasks.

A TOPOLOGY-BASED APPROACH TO QUANTIFY NETWORK PERTURBATION SCORES FOR ASSESSMENT OF DIFFERENT TOBACCO PRODUCT CLASSES

Quynh T. Tran[1], Lee Larcombe[2], Subhashini Arimilli[3], G.L. Prasad[1]

[1]Reynolds American Inc. Services Company - Winston Salem NC - USA 27105; [2]Applied Exomics Ltd - Stevenage UK SG1 2FX; [3]Wake Forest Baptist Health - Winston Salem NC USA 27104

Background: Chronic cigarette smoking is known to cause immune suppresion, which in turn contributes to increased susceptibility to cancer. However, there is limited information on the effects of non-combustible tobacco products, such as moist snuff. To better understand the molecular changes that result from consumption of different tobacco products, global profiling techniques have been extensively utilized. A limitation of such approaches is that differential gene expression alone may be insufficient to identify both the source of perturbation and the extent to which perturbations propagate through a network of interacting genes. Systems biology tools support the analyses and integration of complex datasets, and provide a holistic view of the underlying biological changes. Hence, we implemented a network-based analysis tool to elucidate molecular changes that arise from the use of different tobacco products.
Methods: We developed an analytical approach to quantify and visualize gene-level perturbation scores of a pre-identified network. This approach differentiates biological effects of multiple treatments, using genome-scale expression data and considering interactome-wide effects. We utilized a microarray gene expression dataset of peripheral blood mononuclear cells treated with aqueous extracts of whole smoke conditioned medium (WS-CM) and smokeless tobacco extract (STE) prepared from 3R4F cigarettes and 2S3 moist snuff reference tobacco products, respectively, at baseline and after stimulation with toll-like receptor (TLR) agonists. The analytical pipeline takes normalized gene expression values and performs the following steps: 1) generates gene-level network scores using a weighted topology approach considering both the gene expression data and the full human interactome information available in IntAct (a literature curated molecular interaction database); 2) derives gene-level perturbation scores for each treatment condition compared to its baseline; and 3) calculates a single impact score for each exposure condition and creates a network graph to be visualized using CytoScape.
Results: The pipeline was applied to calculate impact scores under each stimulation and each treatment condition for an inflammatory response network, signaling through a triggered receptor expressed on myeloid cells 1 (TREM1). Samples stimulated with TLR agonists had higher scores or more perturbation compared to non-stimulated samples. Those exposed to higher WS-CM doses received higher scores compared to lower doses of WS-CM. Samples exposed to STE received a lower score suggesting STE treatment perturbed TREM1 network to a lower degree than WS-CM. On the other hand, the classical differential gene expression analysis did not identify significant changes in gene expression for STE treated samples stimulated with TLR agonists, compared to untreated cells. Conclusions: In summary, this network scoring methodology suggests that, under these conditions, STE exerts less perturbation on select immune networks compared to combustible tobacco products. These scores potentially serve as tools to differentiate the biological effects resulting from different tobacco classes.

AUTHOR INDEX