PACIFIC SYMPOSIUM ON BIOCOMPUTING 2019

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 #2-29 are not assigned poster space.

Abstracts are organized first by session, then the last name of the first author. Presenting authors' names are in **bold** text.

TABLE OF CONTENTS

PROCEEDINGS PAPERS WITH ORAL PRESENTATION

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

SINGLE CELL ANALYSIS – WHAT IS THE FUTURE?

WHEN BIOLOGY GETS PERSONAL: HIDDEN CHALLENGES OF PRIVACY AND ETHICS IN BIOLOGICAL BIG DATA

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

SINGLE CELL ANALYSIS – WHAT IS THE FUTURE?

WHEN BIOLOGY GETS PERSONAL: HIDDEN CHALLENGES OF PRIVACY AND ETHICS IN BIOLOGICAL BIG DATA

POSTER PRESENTATIONS

GENERAL

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

THE EFFECTIVENESS OF MULTITASK LEARNING FOR PHENOTYPING WITH ELECTRONIC HEALTH RECORDS DATA

Daisy Yi Ding[1], Chloe Simpson[1], Stephen Pfohl[1], Dave C. Kale[2], Kenneth Jung[1], Nigam H. Shah[1]

[1]Stanford University, [2]University of Southern California

Electronic phenotyping is the task of ascertaining whether an individual has a medical condition of interest by analyzing their medical record and is foundational in clinical informatics. Increasingly, electronic phenotyping is performed via supervised learning. We investigate the effectiveness of multitask learning for phenotyping using electronic health records (EHR) data. Multitask learning aims to improve model performance on a target task by jointly learning additional auxiliary tasks and has been used in disparate areas of machine learning. However, its utility when applied to EHR data has not been established, and prior work suggests that its benefits are inconsistent. We present experiments that elucidate when multitask learning with neural nets improves performance for phenotyping using EHR data relative to neural nets trained for a single phenotype and to well-tuned baselines. We find that multitask neural nets consistently outperform single-task neural nets for rare phenotypes but underperform for relatively more common phenotypes. The effect size increases as more auxiliary tasks are added. Moreover, multitask learning reduces the sensitivity of neural nets to hyperparameter settings for rare phenotypes. Last, we quantify phenotype complexity and find that neural nets trained with or without multitask learning do not improve on simple baselines unless the phenotypes are sufficiently complex.

ODAL: A ONE-SHOT DISTRIBUTED ALGORITHM TO PERFORM LOGISTIC REGRESSIONS ON ELECTRONIC HEALTH RECORDS DATA FROM MULTIPLE CLINICAL SITES

Rui Duan, Mary Regina Boland, Jason H. Moore, **Yong Chen**

Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania

Electronic Health Records (EHR) contain extensive information on various health outcomes and risk factors, and therefore have been broadly used in healthcare research. Integrating EHR data from multiple clinical sites can accelerate knowledge discovery and risk prediction by providing a larger sample size in a more general population which potentially reduces clinical bias and improves estimation and prediction accuracy. To overcome the barrier of patient-level data sharing, distributed algorithms are developed to conduct statistical analyses across multiple sites through sharing only aggregated information. The current distributed algorithm often requires iterative information evaluation and transferring across sites, which can potentially lead to a high communication cost in practical settings. In this study, we propose a privacy-preserving and communication-efficient distributed algorithm for logistic regression without requiring iterative communications across sites. Our simulation study showed our algorithm reached comparative accuracy comparing to the oracle estimator where data are pooled together. We applied our algorithm to an EHR data from the University of Pennsylvania health system to evaluate the risks of fetal loss due to various medication exposures.

PVC DETECTION USING A CONVOLUTIONAL AUTOENCODER AND RANDOM FOREST CLASSIFIER

Max Gordon, Cranos Williams

North Carolina State University

The accurate detection of premature ventricular contractions (PVCs) in patients is an important task in cardiac care for some patients. In some cases, the usefulness to physicians in detecting PVCs stems from their long-term correlations with dangerous heart conditions. In other cases their potential as a precursor to serious cardiac events may make their detection a useful early warning mechanism. In many of these applications, the long-term nature of the monitoring required and the infrequency of PVCs make manual observation for PVCs impractical. Existing methods of automated PVC detection suffer from drawbacks such as the need to use difficult to extract morphological features, domain-specific features, or large numbers of estimated parameters. In particular, systems using large numbers of trained parameters have the potential to require large amounts of training data and computation and may have issues generalizing due to their potential to overfit. To address some of these drawbacks, we developed a novel PVC detection algorithm based around a convolutional autoencoder to address these weaknesses and validated our method using the MIT-BIH arrhythmia database.

PLATYPUS: A MULTIPLE–VIEW LEARNING PREDICTIVE FRAMEWORK FOR CANCER DRUG SENSITIVITY PREDICTION

Kiley Graim, Verena Friedl, Kathleen E. Houlahan, Joshua M. Stuart

Dept. of Biomolecular Engineering University of California Santa Cruz, Flatiron Institute and Princeton University, Ontario Institute of Cancer Research and University of Toronto

Cancer is a complex collection of diseases that are to some degree unique to each patient. Precision oncology aims to identify the best drug treatment regime using molecular data on tumor samples. While omics-level data is becoming more widely available for tumor specimens, the datasets upon which computational learning methods can be trained vary in coverage from sample to sample and from data type to data type. Methods that can 'connect the dots' to leverage more of the information provided by these studies could offer major advantages for maximizing predictive potential. We introduce a multi-view machine- learning strategy called PLATYPUS that builds 'views' from multiple data sources that are all used as features for predicting patient outcomes. We show that a learning strategy that finds agreement across the views on unlabeled data increases the performance of the learning methods over any single view. We illustrate the power of the approach by deriving signatures for drug sensitivity in a large cancer cell line database. Code and additional information are available from the PLATYPUS website https://sysbiowiki.soe.ucsc.edu/platypus.

DEEPDOM: PREDICTING PROTEIN DOMAIN BOUNDARY FROM SEQUENCE ALONE USING STACKED BIDIRECTIONAL LSTM

Yuexu Jiang, Duolin Wang, Dong Xu

Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA Email: xudong@missouri.edu

Protein domain boundary prediction is usually an early step to understand protein function and structure. Most of the current computational domain boundary prediction methods suffer from low accuracy and limitation in handling multi-domain types, or even cannot be applied on certain targets such as proteins with discontinuous domain. We developed an ab-initio protein domain predictor using a stacked bidirectional LSTM model in deep learning. Our model is trained by a large amount of protein sequences without using feature engineering such as sequence profiles. Hence, the predictions using our method is much faster than others, and the trained model can be applied to any type of target proteins without constraint. We evaluated DeepDom by a 10-fold cross validation and also by applying it on targets in different categories from CASP 8 and CASP 9. The comparison with other methods has shown that DeepDom outperforms most of the current ab-initio methods and even achieves better results than the top-level template-based method in certain cases. The code of DeepDom and the test data we used in CASP 8, 9 can be accessed through GitHub at https://github.com/yuexujiang/DeepDom.

IMPLEMENTING AND EVALUATING A GAUSSIAN MIXTURE FRAMEWORK FOR IDENTIFYING GENE FUNCTION FROM TNSEQ DATA

Kevin Li[1], Rachel Chen[2], William Lindsey[3], Aaron Best[4], Matthew DeJongh[4], Christopher Henry[5], Nathan Tintle[3]

[1]Columbia University, [2]North Carolina State University, [3]Dordt College, [4]Hope College, [5]Argonne Laboratory

The rapid acceleration of microbial genome sequencing increases opportunities to understand bacterial gene function. Unfortunately, only a small proportion of genes have been studied. Recently, TnSeq has been proposed as a cost-effective, highly reliable approach to predict gene functions as a response to changes in a cell's fitness before-after genomic changes. However, major questions remain about how to best determine whether an observed quantitative change in fitness represents a meaningful change. To address the limitation, we develop a Gaussian mixture model framework for classifying gene function from TnSeq experiments. In order to implement the mixture model, we present the Expectation-Maximization algorithm and a hierarchical Bayesian model sampled using Stan's Hamiltonian Monte-Carlo sampler. We compare these implementations against the frequentist method used in current TnSeq literature. From simulations and real data produced by E.coli TnSeq experiments, we show that the Bayesian implementation of the Gaussian mixture framework provides the most consistent classification results.

RES2S2AM: DEEP RESIDUAL NETWORK-BASED MODEL FOR IDENTIFYING FUNCTIONAL NONCODING SNPS IN TRAIT-ASSOCIATED REGIONS

Zheng Liu, Yao Yao, Qi Wei, Benjamin Weeder, Stephen A. Ramsey

Oregon State University

Noncoding single nucleotide polymorphisms (SNPs) and their target genes are important components of the heritability of diseases and other polygenic traits. Identifying these SNPs and target genes could potentially reveal new molecular mechanisms and advance precision medicine. For polygenic traits, genome-wide association studies (GWAS) are preferred tools for identifying trait-associated regions. However, identifying causal noncoding SNPs within such regions is a difficult problem in computational biology. The DNA sequence context of a noncoding SNP is well-established as an important source of information that is beneficial for discriminating functional from nonfunctional noncoding SNPs. We describe the use of a deep residual network (ResNet)-based model—entitled Res2s2aM—that fuses flanking DNA sequence information with additional SNP annotation information to discriminate functional from nonfunctional noncoding SNPs. On a ground-truth set of disease-associated SNPs compiled from the Genome-wide Repository of Associations between SNPs and Phenotypes (GRASP) database, Res2s2aM improves the prediction accuracy of functional SNPs significantly in comparison to models based only on sequence information as well as a leading tool for post-GWAS noncoding SNP prioritization (RegulomeDB).

BI-DIRECTIONAL RECURRENT NEURAL NETWORK MODELS FOR GEOGRAPHIC LOCATION EXTRACTION IN BIOMEDICAL LITERATURE

Arjun Magge[1], Davy Weissenbacher[2], Abeed Sarker[2], Matthew Scotch[1], Graciela Gonzalez-Hernandez[2]

[1]Arizona State University, [2]University of Pennsylvania

Phylogeography research involving virus spread and tree reconstruction relies on accurate geographic locations of infected hosts. Insufficient level of geographic information in nucleotide sequence repositories such as GenBank motivates the use of natural language processing methods for extracting geographic location names (toponyms) in the scientific article associated with the sequence, and disambiguating the locations to their co-ordinates. In this paper, we present an extensive study of multiple recurrent neural network architectures for the task of extracting geographic locations and their effective contribution to the disambiguation task using population heuristics. The methods presented in this paper achieve a strict detection F-1 score of 0.94, disambiguation accuracy of 91% and an overall resolution F-1 score of 0.88 that are significantly higher than previously developed methods, improving our capability to find the location of infected hosts and enrich metadata information.

COMPUTATIONAL KIR COPY NUMBER DISCOVERY REVEALS INTERACTION BETWEEN INHIBITORY RECEPTOR BURDEN AND SURVIVAL

Rachel M. Pyke[1], Raphael Genolet[2], Alexandre Harari[2], George Coukos[2], David Gfeller[2], Hannah Carter[1]

[1]University of California - San Diego, [2]Ludwig Institute for Cancer Research - University of Lausanne

Natural killer (NK) cells have increasingly become a target of interest for immunotherapies1. NK cells express killer immunoglobulin-like receptors (KIRs), which play a vital role in immune response to tumors by detecting cellular abnormalities. The genomic region encoding the 16 KIR genes displays high polymorphic variability in human populations, making it difficult to resolve individual genotypes based on next generation sequencing data. As a result, the impact of polymorphic KIR variation on cancer phenotypes has been understudied. Currently, labor-intensive, experimental techniques are used to determine an individual's KIR gene copy number profile. Here, we develop an algorithm to determine the germline copy number of KIR genes from whole exome sequencing data and apply it to a cohort of nearly 5000 cancer patients. We use a k-mer based approach to capture sequences unique to specific genes, count their occurrences in the set of reads derived from an individual and compare the individual's k-mer distribution to that of the population. Copy number results demonstrate high concordance with population copy number expectations. Our method reveals that the burden of inhibitory KIR genes is associated with survival in two tumor types, highlighting the potential importance of KIR variation in understanding tumor development and response to immunotherapy.

SEMANTIC WORKFLOWS FOR BENCHMARK CHALLENGES: ENHANCING COMPARABILITY, REUSABILITY AND REPRODUCIBILITY

Arunima Srivastava[1], Ravali Adusumilli[2], Hunter Boyce[2], Daniel Garijo[3], Varun Ratnakar[3], Rajiv Mayani[3], Thomas Yu[4], Raghu Machiraju[1], Yolanda Gil[3], Parag Mallick[2]

[1]The Ohio State University, [2]Stanford University, [3]University of Southern California, [4]Sage Bionetworks

Benchmark challenges, such as the Critical Assessment of Structure Prediction (CASP) and Dialogue for Reverse Engineering Assessments and Methods (DREAM) have been instrumental in driving the development of bioinformatics methods. Typically, challenges are posted, and then competitors perform a prediction based upon blinded test data. Challengers then submit their answers to a central server where they are scored. Recent efforts to automate these challenges have been enabled by systems in which challengers submit Docker containers, a unit of software that packages up code and all of its dependencies, to be run on the cloud. Despite their incredible value for providing an unbiased test-bed for the bioinformatics community, there remain opportunities to further enhance the potential impact of benchmark challenges. Specifically, current approaches only evaluate end-to-end performance; it is nearly impossible to directly compare methodologies or parameters. Furthermore, the scientific community cannot easily reuse challengers' approaches, due to lack of specifics, ambiguity in tools and parameters as well as problems in sharing and maintenance. Lastly, the intuition behind why particular steps are used is not captured, as the proposed workflows are not explicitly defined, making it cumbersome to understand the flow and utilization of data. Here we introduce an approach to overcome these limitations based upon the WINGS semantic workflow system. Specifically, WINGS enables researchers to submit complete semantic workflows as challenge submissions. By submitting entries as workflows, it then becomes possible to compare not just the results and performance of a challenger, but also the methodology employed. This is particularly important when dozens of challenge entries may use nearly identical tools, but with only subtle changes in parameters (and radical differences in results). WINGS uses a component driven workflow design and offers intelligent parameter and data selection by reasoning about data characteristics. This proves to be especially critical in bioinformatics workflows where using default or incorrect parameter values is prone to drastically altering results. Different challenge entries may be readily compared through the use of abstract workflows, which also facilitate reuse. WINGS is housed on a cloud based setup, which stores data, dependencies and workflows for easy sharing and utility. It also has the ability to scale workflow executions using distributed computing through the Pegasus workflow execution system. We demonstrate the application of this architecture to the DREAM proteogenomic challenge.

REMOVING CONFOUNDING FACTORS ASSOCIATED WEIGHTS IN DEEP NEURAL NETWORKS IMPROVES THE PREDICTION ACCURACY FOR HEALTHCARE APPLICATIONS

Haohan Wang[1], Zhenglin Wu[2], Eric P. Xing[3]

[1]Carnegie Mellon University, [2]University of Illinois Urbana-Champaign, [3]Carnegie Mellon University

The proliferation of healthcare data has brought the opportunities of applying data-driven approaches, such as machine learning methods, to assist diagnosis. Recently, many deep learning methods have been shown with impressive successes in predicting disease status with raw input data. However, the ``black-box'' nature of deep learning and the high-reliability requirement of biomedical applications have created new challenges regarding the existence of confounding factors. In this paper, with a brief argument that inappropriate handling of confounding factors will lead to models' sub-optimal performance in real-world applications, we present an efficient method that can remove the influences of confounding factors such as age or gender to improve the across-cohort prediction accuracy of neural networks. One distinct advantage of our method is that it only requires minimal changes of the baseline model's architecture so that it can be plugged into most of the existing neural networks. We conduct experiments across CT-scan, MRA, and EEG brain wave with convolutional neural networks and LSTM to verify the efficiency of our method.

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

AN OPTIMAL POLICY FOR PATIENT LABORATORY TESTS IN INTENSIVE CARE UNITS

Li-Fang Cheng, **Niranjani Prasad**, Barbara E. Engelhardt

Princeton University

Laboratory testing is an integral tool in the management of patient care in hospitals, particularly in intensive care units (ICUs). There exists an inherent trade-off in the selection and timing of lab tests between considerations of the expected utility in clinical decision-making of a given test at a specific time, and the associated cost or risk it poses to the patient. In this work, we introduce a framework that learns policies for ordering lab tests which optimizes for this trade-off. Our approach uses batch off-policy reinforcement learning with a composite reward function based on clinical imperatives, applied to data that include examples of clinicians ordering labs for patients. To this end, we develop and extend principles of Pareto optimality to improve the selection of actions based on multiple reward function components while respecting typical procedural considerations and prioritization of clinical goals in the ICU. Our experiments show that we can estimate a policy that reduces the frequency of lab tests and optimizes timing to minimize information redundancy. We also find that the estimated policies typically suggest ordering lab tests well ahead of critical onsets---such as mechanical ventilation or dialysis---that depend on the lab results. We evaluate our approach by quantifying how these policies may initiate earlier onset of treatment.

CROWDVARIANT: A CROWDSOURCING APPROACH TO CLASSIFY COPY NUMBER VARIANTS

Peyton Greenside[1], Justin Zook[2], Marc Salit[3], Ryan Poplin[4], Madeleine Cule[5], Mark DePristo[4]

[1]Stanford University, [2]National Institute of Standards and Technologies (NIST), [3]National Institute of Standards and Technologies (NIST)/Joint Initiative for Metrology in Biology (JIMB), [4]Google Inc./Verily Life Sciences, [5]Calico/Verily Life Sciences

Copy number variants (CNVs) are an important type of genetic variation that play a causal role in many diseases. The ability to identify high quality CNVs is of substantial clinical relevance. However, CNVs are notoriously difficult to identify accurately from array-based methods and next-generation sequencing (NGS) data, particularly for small (< 10kbp) CNVs. Manual curation by experts widely remains the gold standard but cannot scale with the pace of sequencing, particularly in fast-growing clinical applications. We present the first proof-of-principle study demonstrating high throughput manual curation of putative CNVs by non-experts. We developed a crowdsourcing framework, called CrowdVariant, that leverages Google's high-throughput crowdsourcing platform to create a high confidence set of deletions for NA24385 (NIST HG002/RM 8391), an Ashkenazim reference sample developed in partnership with the Genome In A Bottle (GIAB) Consortium. We show that non-experts tend to agree both with each other and with experts on putative CNVs. We show that crowdsourced non-expert classifications can be used to accurately assign copy number status to putative CNV calls and identify 1,781 high confidence deletions in a reference sample. Multiple lines of evidence suggest these calls are a substantial improvement over existing CNV callsets and can also be useful in benchmarking and improving CNV calling algorithms. Our crowdsourcing methodology takes the first step toward showing the clinical potential for manual curation of CNVs at scale and can further guide other crowdsourcing genomics applications.

A REPOSITORY OF MICROBIAL MARKER GENES RELATED TO HUMAN HEALTH AND DISEASES FOR HOST PHENOTYPE PREDICTION USING MICROBIOME DATA

Wontack Han, Yuzhen Ye

Indiana University

The microbiome research is going through an evolutionary transition from focusing on the characterization of reference microbiomes associated with different environments/hosts to the translational applications, including using microbiome for disease diagnosis, improving the efficacy of cancer treatments, and prevention of diseases (e.g., using probiotics). Microbial markers have been identified from microbiome data derived from cohorts of patients with different diseases, treatment responsiveness, etc, and often predictors based on these markers were built for predicting host phenotype given a microbiome dataset (e.g., to predict if a person has type 2 diabetes given his or her microbiome data). Unfortunately, these microbial markers and predictors are often not published so are not reusable by others. In this paper, we report the curation of a repository of microbial marker genes and predictors built from these markers for microbiome-based prediction of host phenotype, and a computational pipeline called Mi2P (from Microbiome to Phenotype) for using the repository. As an initial effort, we focus on microbial marker genes related to two diseases, type 2 diabetes and liver cirrhosis, and immunotherapy efficacy for two types of cancer, non-small-cell lung cancer (NSCLC) and renal cell carcinoma (RCC). We characterized the marker genes from metagenomic data using our recently developed subtractive assembly approach. We showed that predictors built from these microbial marker genes can provide fast and reasonably accurate prediction of host phenotype given microbiome data. As understanding and making use of microbiome data (our second genome) is becoming vital as we move forward in this age of precision health and precision medicine, we believe that such a repository will be useful for enabling translational applications of microbiome data.

AICM: A GENUINE FRAMEWORK FOR CORRECTING INCONSISTENCY BETWEEN LARGE PHARMACOGENOMICS DATASETS

Zhiyue Tom Hu[1], Yuting Ye[1], Patrick A. Newbury[2], Haiyan Huang[2,3,4], Bin Chen[5]

[1]University of California Berkeley, Department of Biostatistics; [1]University of California Berkeley, Department of Biostatistics; [2]University of California Berkeley, Department of Pediatrics and Human Development; [3]Michigan State University, Department of Statistics, [4]University of California Berkeley, Department of Pharmacology and Toxicology; [5]Michigan State University

The inconsistency of open pharmacogenomics datasets produced by different studies limits the usage of such datasets in many tasks, such as biomarker discovery. Investigation of multiple pharmacogenomics datasets confirmed that the pairwise sensitivity data correlation between drugs, or rows, across different studies (drug-wise) is relatively low, while the pairwise sensitivity data correlation between cell-lines, or columns, across different studies (cell-wise) is considerably strong. This common interesting observation across multiple pharmacogenomics datasets suggests the existence of subtle consistency among the different studies (i.e., strong cell-wise correlation). However, significant noises are also shown (i.e., weak drug-wise correlation) and have prevented researchers from comfortably using the data directly. Motivated by this observation, we propose a novel framework for addressing the inconsistency between large-scale pharmacogenomics data sets. Our method can significantly boost the drug-wise correlation and can be easily applied to re-summarized and normalized datasets proposed by others. We also investigate our algorithm based on many different criteria to demonstrate that the corrected datasets are not only consistent, but also biologically meaningful. Eventually, we propose to extend our main algorithm into a framework, so that in the future when more datasets become publicly available, our framework can hopefully offer a "ground-truth" guidance for references.

INTEGRATING RNA EXPRESSION AND VISUAL FEATURES FOR IMMUNE INFILTRATE PREDICTION

Derek Reiman[1], Lingdao Sha[1], Irvin Ho[1], Timothy Tan[2], Denise Lau[1], **Aly A. Khan**[3]

[1]Tempus Labs, [2]Northwestern University, [3]Toyota Technological Institute at Chicago

Patient responses to cancer immunotherapy are shaped by their unique genomic landscape and tumor microenvironment. Clinical advances in immunotherapy are changing the treatment landscape by enhancing a patient's immune response to eliminate cancer cells. While this provides potentially beneficial treatment options for many patients, only a minority of these patients respond to immunotherapy. In this work, we examined RNA-seq data and digital pathology images from individual patient tumors to more accurately characterize the tumor-immune microenvironment. Several studies implicate an inflamed microenvironment and increased percentage of tumor infiltrating immune cells with better response to specific immunotherapies in certain cancer types. We developed NEXT (Neural-based models for integrating gene EXpression and visual Texture features) to more accurately model immune infiltration in solid tumors. To demonstrate the utility of the NEXT framework, we predicted immune infiltrates across four different cancer types and evaluated our predictions against expert pathology review. Our analyses demonstrate that integration of imaging features improves prediction of the immune infiltrate. Of note, this effect was preferentially observed for B cells and CD8 T cells. In sum, our work effectively integrates both RNA-seq and imaging data in a clinical setting and provides a more reliable and accurate prediction of the immune composition in individual patient tumors.

OUTGROUP MACHINE LEARNING APPROACH IDENTIFIES SINGLE NUCLEOTIDE VARIANTS IN NONCODING DNA ASSOCIATED WITH AUTISM SPECTRUM DISORDER

Maya Varma, Kelley Marie Paskov, Jae-Yoon Jung, Brianna Sierra Chrisman, Nate Tyler Stockham, Peter Yigitcan Washington, Dennis Paul Wall

Stanford University

Autism spectrum disorder (ASD) is a heritable neurodevelopmental disorder affecting 1 in 59 children. While noncoding genetic variation has been shown to play a major role in many complex disorders, the contribution of these regions to ASD susceptibility remains unclear. Genetic analyses of ASD typically use unaffected family members as controls; however, we hypothesize that this method does not effectively elevate variant signal in the noncoding region due to family members having subclinical phenotypes arising from common genetic mechanisms. In this study, we use a separate, unrelated outgroup of individuals with progressive supranuclear palsy (PSP), a neurodegenerative condition with no known etiological overlap with ASD, as a control population. We use whole genome sequencing data from a large cohort of 2182 children with ASD and 379 controls with PSP, sequenced at the same facility with the same machines and variant calling pipeline, in order to investigate the role of noncoding variation in the ASD phenotype. We analyze seven major types of noncoding variants: microRNAs, human accelerated regions, hypersensitive sites, transcription factor binding sites, DNA repeat sequences, simple repeat sequences, and CpG islands. After identifying and removing batch effects between the two groups, we trained an l1-regularized logistic regression classifier to predict ASD status from each set of variants. The classifier trained on simple repeat sequences performed well on a held-out test set (AUC-ROC = 0.960); this classifier was also able to differentiate ASD cases from controls when applied to a completely independent dataset (AUC-ROC = 0.960). This suggests that variation in simple repeat regions is predictive of the ASD phenotype and may contribute to ASD risk. Our results show the importance of the noncoding region and the utility of independent control groups in effectively linking genetic variation to disease phenotype for complex disorders.

PRECISION DRUG REPURPOSING VIA CONVERGENT EQTL-BASED MOLECULES AND PATHWAY TARGETING INDEPENDENT DISEASE-ASSOCIATED POLYMORPHISMS

Francesca Vitali[1,2], Joanne Berghout[1, 2,3], Jungwei Fan[1,2], Jianrong Li[1], Qike Li[1], Haiquan Li[1, 2, 4], Yves A. Lussier[1, 2, 3, 5]

[1]Center for Biomedical Informatics and Biostatistics (CB2) of The University of Arizona, [2]Department of Medicine COM-T of The University of Arizona, [3]The Center for Applied Genetics and Genomics in Medicine of The University of Arizona, [4]Department of Biosystems Engineering of The University of Arizona, [5]UA Cancer Center UA Health Science (UAHS) of The University of Arizona

Repurposing existing drugs for new therapeutic indications can improve success rates and streamline development. Use of large-scale biomedical data repositories, including eQTL regulatory relationships and genome-wide disease risk associations, offers opportunities to propose novel indications for drugs targeting common or convergent molecular candidates associated to two or more diseases. This proposed novel computational approach scales across 262 complex diseases, building a multi-partite hierarchical network integrating (i) GWAS-derived SNP-to-disease associations, (ii) eQTL-derived SNP-to-eGene associations incorporating both cis- and trans- relationships from 19 tissues, (iii) protein target-to-drug, and (iv) drug-to-disease indications with (iv) Gene Ontology-based information theoretic semantic (ITS) similarity calculated between protein target functions. Our hypothesis is that if two diseases are associated to a common or functionally similar eGene - and a drug targeting that eGene/protein in one disease exists - the second disease becomes a potential repurposing indication. To explore this, all possible pairs of independently segregating GWAS-derived SNPs were generated, and a statistical network of similarity within each SNP-SNP pair was calculated according to scale-free overrepresentation of convergent biological processes activity in regulated eGenes (ITSeGENE-eGENE) and scale-free overrepresentation of common eGene targets between the two SNPs (ITSSNP-SNP). Significance of ITSSNP-SNP was conservatively estimated using empirical scale-free permutation resampling keeping the node-degree constant for each molecule in each permutation. We identified 26 new drug repurposing indication candidates spanning 89 GWAS diseases, including a potential repurposing of the calcium-channel blocker Verapamil from coronary disease to gout. Predictions from our approach are compared to known drug indications using DrugBank as a gold standard (odds ratio=13.1, p-value=2.49x10-8). Because of specific disease-SNPs associations to candidate drug targets, the proposed method provides evidence for future precision drug repositioning to a patient's specific polymorphisms.

DETECTING POTENTIAL PLEIOTROPY ACROSS CARDIOVASCULAR AND NEUROLOGICAL DISEASES USING UNIVARIATE, BIVARIATE, AND MULTIVARIATE METHODS ON 43,870 INDIVIDUALS FROM THE EMERGE NETWORK

Xinyuan Zhang[1], Yogasudha Veturi[1], Shefali S. Verma[1], William Bone[1], Anurag Verma[1], Anastasia M. Lucas[1], Scott Hebbring[2], Joshua C. Denny[3], Ian Stanaway[4], Gail P. Jarvik[4], David Crosslin[4], Eric B. Larson[5], Laura Rasmussen-Torvik[6], Sarah A. Pendergrass[7], Jordan W. Smoller[8], Hakon Hakonarson[9], Patrick Sleiman[9], Chunhua Weng[10], David Fasel[10], Wei-Qi Wei[3], Iftikhar Kullo[11], Daniel Schaid[11], Wendy K. Chung[10], Marylyn D. Ritchie[1]

[1]University of Pennsylvania, [2]Marshfield Clinic, [3]Vanderbilt University, [4]University of Washington, [5]Kaiser Permanente Washington Health Research Institute, [6]Northwestern University, [7]Geisinger Health System, [8]Massachusetts General Hospital, [9]Children's Hospital of Philadelphia, [10]Columbia University, [11]Mayo Clinic

The link between cardiovascular diseases and neurological disorders has been widely observed in the aging population. Disease prevention and treatment rely on understanding the potential genetic nexus of multiple diseases in these categories. In this study, we were interested in detecting pleiotropy, or the phenomenon in which a genetic variant influences more than one phenotype. Marker-phenotype association approaches can be grouped into univariate, bivariate, and multivariate categories based on the number of phenotypes considered at one time. Here we applied one statistical method per category followed by an eQTL colocalization analysis to identify potential pleiotropic variants that contribute to the link between cardiovascular and neurological diseases. We performed our analyses on ~530,000 common SNPs coupled with 65 electronic health record (EHR)-based phenotypes in 43,870 unrelated European adults from the Electronic Medical Records and Genomics (eMERGE) network. There were 31 variants identified by all three methods that showed significant associations across late onset cardiac- and neurologic- diseases. We further investigated functional implications of gene expression on the detected “lead SNPs” via colocalization analysis, providing a deeper understanding of the discovered associations. In summary, we present the framework and landscape for detecting potential pleiotropy using univariate, bivariate, multivariate, and colocalization methods. Further exploration of these potentially pleiotropic genetic variants will work toward understanding disease causing mechanisms across cardiovascular and neurological diseases and may assist in considering disease prevention as well as drug repositioning in future research.

SINGLE CELL ANALYSIS – WHAT IS THE FUTURE?

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

LISA: ACCURATE RECONSTRUCTION OF CELL TRAJECTORY AND PSEUDO-TIME FOR MASSIVE SINGLE CELL RNA-SEQ DATA

Yang Chen[1], Yuping Zhang[2], **Zhengqing Ouyang**[1]

[1]The Jackson Laboratory for Genomic Medicine, [2]University of Connecticut

Cell trajectory reconstruction based on single cell RNA sequencing is important for obtaining the landscape of different cell types and discovering cell fate transitions. Despite intense effort, analyzing massive single cell RNA-seq datasets is still challenging. We propose a new method named Landmark Isomap for Single-cell Analysis (LISA). LISA is an unsupervised approach to build cell trajectory and compute pseudo-time in the isometric embedding based on geodesic distances. The advantages of LISA include: (1) It utilizes k-nearest-neighbor graph and hierarchical clustering to identify cell clusters, peaks and valleys in low-dimension representation of the data; (2) Based on Landmark Isomap, it constructs the main geometric structure of cell lineages; (3) It projects cells to the edges of the main cell trajectory to generate the global pseudo-time. Assessments on simulated and real datasets demonstrate the advantages of LISA on cell trajectory and pseudo-time reconstruction compared to Monocle2 and TSCAN. LISA is accurate, fast, and requires less memory usage, allowing its applications to massive single cell datasets generated from current experimental platforms.

PARAMETER TUNING IS A KEY PART OF DIMENSIONALITY REDUCTION VIA DEEP VARIATIONAL AUTOENCODERS FOR SINGLE CELL RNA TRANSCRIPTOMICS

Qiwen Hu, Casey S. Greene

University of Pennsylvania

Single-cell RNA sequencing (scRNA-seq) is a powerful tool to profile the transcriptomes of a large number of individual cells at a high resolution. These data usually contain measurements of gene expression for many genes in thousands or tens of thousands of cells, though some datasets now reach the million-cell mark. Projecting high-dimensional scRNA-seq data into a low dimensional space aids downstream analysis and data visualization. Many recent preprints accomplish this using variational autoencoders (VAE), generative models that learn underlying structure of data by compress it into a constrained, low dimensional space. The low dimensional spaces generated by VAEs have revealed complex patterns and novel biological signals from large-scale gene expression data and drug response predictions. Here, we evaluate a simple VAE approach for gene expression data, Tybalt, by training and measuring its performance on sets of simulated scRNA-seq data. We find a number of counter-intuitive performance features: i.e., deeper neural networks can struggle when datasets contain more observations under some parameter configurations. We show that these methods are highly sensitive to parameter tuning: when tuned, the performance of the Tybalt model, which was not optimized for scRNA-seq data, outperforms other popular dimension reduction approaches – PCA, ZIFA, UMAP and t-SNE. On the other hand, without tuning performance can also be remarkably poor on the same data. Our results should discourage authors and reviewers from relying on self-reported performance comparisons to evaluate the relative value of contributions in this area at this time. Instead, we recommend that attempts to compare or benchmark autoencoder methods for scRNA-seq data be performed by disinterested third parties or by methods developers only on unseen benchmark data that are provided to all participants simultaneously because the potential for performance differences due to unequal parameter tuning is so high.

TOPOLOGICAL METHODS FOR VISUALIZATION AND ANALYSIS OF HIGH DIMENSIONAL SINGLE-CELL RNA SEQUENCING DATA

Tongxin Wang[1], Travis Johnson[2], Jie Zhang[3], Kun Huang[4,5]

[1]Department of Computer Science, Indiana University Bloomington; [2]Department of Biomedical Inforamtics, Ohio State University; [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine; [4]Department of Medicine, Indiana University School of Medicine; [5]Regenstrief Institute

Single-cell RNA sequencing (scRNA-seq) techniques have been very powerful in analyzing heterogeneous cell population and identifying cell types. Visualizing scRNA-seq data can help researchers effectively extract meaningful biological information and make new discoveries. While commonly used scRNA-seq visualization methods, such as t-SNE, are useful in detecting cell clusters, they often tear apart the intrinsic continuous structure in gene expression profiles. Topological Data Analysis (TDA) approaches like Mapper capture the shape of data by representing data as topological networks. TDA approaches are robust to noise and different platforms, while preserving the locality and data continuity. Moreover, instead of analyzing the whole dataset, Mapper allows researchers to explore biological meanings of specific pathways and genes by using different filter functions. In this paper, we applied Mapper to visualize scRNA-seq data. Our method can not only capture the clustering structure of cells, but also preserve the continuous gene expression topologies of cells. We demonstrated that by combining with gene co-expression network analysis, our method can reveal differential expression patterns of gene co-expression modules along the Mapper visualization.

WHEN BIOLOGY GETS PERSONAL: HIDDEN CHALLENGES OF PRIVACY AND ETHICS IN BIOLOGICAL BIG DATA

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

LEVERAGING SUMMARY STATISTICS TO MAKE INFERENCES ABOUT COMPLEX PHENOTYPES IN LARGE BIOBANKS

Angela Gasdaska[1], Derek Friend[2], Rachel Chen[3], Jason Westra[4], Matthew Zawistowski[5], William Lindsey[4], **Nathan Tintle**[4]

[1]Emory University, [2]University of Nevada Reno, [3]North Carolina State University, [4]Dordt College, [5]University of Michigan Ann Arbor

As genetic sequencing becomes less expensive and data sets linking genetic data and medical records (e.g., Biobanks) become larger and more common, issues of data privacy and computational challenges become more necessary to address in order to realize the benefits of these datasets. One possibility for alleviating these issues is through the use of already-computed summary statistics (e.g., slopes and standard errors from a regression model of a phenotype on a genotype). If groups share summary statistics from their analyses of biobanks, many of the privacy issues and computational challenges concerning the access of these data could be bypassed. In this paper we explore the possibility of using summary statistics from simple linear models of phenotype on genotype in order to make inferences about more complex phenotypes (those that are derived from two or more simple phenotypes). We provide exact formulas for the slope, intercept, and standard error of the slope for linear regressions when combining phenotypes. Derived equations are validated via simulation and tested on a real data set exploring the genetics of fatty acids.

EVALUATION OF PATIENT RE-IDENTIFICATION USING LABORATORY TEST ORDERS AND MITIGATION VIA LATENT SPACE VARIABLES

Kipp W. Johnson[1], **Jessica K. De Freitas**[1], Benjamin S. Glicksberg[1], Jason R. Bobe[1], Joel T. Dudley[2]

[1]Institute for Next Generation Healthcare - Department of Genetics and Genomics Sciences - Icahn School of Medicine at Mount Sinai, [2]Bakar Computational Health Sciences Institute The University of California San Francisco

A variety of clinical data abstracted and anonymized from electronic health records (EHR) are often used for research purposes. One consistent concern with this type of research is the risk for re-identification of patients from their anonymized data. Here, we use the EHR of 731,850 patients to demonstrate that the average patient is unique from all others 98.4% of the time simply by examining what laboratory tests have been ordered for them. By the time a patient has visited the hospital on two separate days, they are unique in 74.2% of cases. We further present a computational study to identify how accurately the records from a single day of care can be used to re-identify patients from a set of 99 other patients. We show that, given a single visit's laboratory orders for a patient, we can re-identify the patient at least 25% of the time. Furthermore, we can place this patient among the top 10 most similar patients 47% of the time. Finally, we present a proof-of-concept technique using a variational autoencoder to encode laboratory results into a lower-dimensional latent space. We demonstrate that releasing latent- space encoded laboratory orders significantly improves privacy compared to releasing raw laboratory orders (<5% re-identification), while preserving information contained within the laboratory orders (AUC of >0.9 for recreating encoded values). Our findings potentially have consequences for the public release of anonymized laboratory tests to the biomedical research community. We wish to note that our findings do not imply that laboratory tests alone are personally identifiable, but would require a threat actor having an external source of laboratory values which are linked to personal identifiers to begin with.

PROTECTING GENOMIC DATA PRIVACY WITH PROBABILISTIC MODELING

Sean Simmons[1], Bonnie Berger[2], Cenk Sahinalp[3]

[1]Broad Institute, [2]MIT, [3]Indiana University

The proliferation of sequencing technologies in biomedical research has raised many new privacy concerns. These include concerns over the publication of aggregate data at a genomic scale (e.g. minor allele frequencies, regression coefficients). Methods such as differential privacy can overcome these concerns by providing strong privacy guarantees, but come at the cost of greatly perturbing the results of the analysis of interest. Here we investigate an alternative approach for achieving privacy-preserving aggregate genomic data sharing without the high cost to accuracy of differentially private methods. In particular, we demonstrate how other ideas from the statistical disclosure control literature (in particular, the idea of disclosure risk) can be applied to aggregate data to help ensure privacy. This is achieved by combining minimal amounts of perturbation with Bayesian statistics and Markov Chain Monte Carlo techniques.We test our technique on a GWAS dataset to demonstrate its utility in practice. An implementation is available at https://github.com/seanken/PrivMCMC.

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

SNPs2ChIP: LATENT FACTORS OF ChIP-SEQ TO INFER FUNCTIONS OF NON-CODING SNPs

Shankara Anand, Laurynas Kalesinskas, Craig Smail, **Yosuke Tanigawa**

Stanford University

Genetic variations of the human genome are linked to many disease phenotypes. While whole-genome sequencing and genome-wide association studies (GWAS) have uncovered a number of genotype-phenotype associations, their functional interpretation remains challenging given most single nucleotide polymorphisms (SNPs) fall into the non-coding region of the genome. Advances in chromatin immunoprecipitation sequencing (ChIP-seq) have made large-scale repositories of epigenetic data available, allowing investigation of coordinated mechanisms of epigenetic markers and transcriptional regulation and their influence on biological function. To address this, we propose SNPs2ChIP, a method to infer biological functions of non-coding variants through unsupervised statistical learning methods applied to publicly-available epigenetic datasets. We systematically characterized latent factors by applying singular value decomposition to ChIP-seq tracks of lymphoblastoid cell lines, and annotated the biological function of each latent factor using the genomic region enrichment analysis tool. Using these annotated latent factors as reference, we developed SNPs2ChIP, a pipeline that takes genomic region(s) as an input, identifies the relevant latent factors with quantitative scores, and returns them along with their inferred functions. As a case study, we focused on systemic lupus erythematosus and demonstrated our method's ability to infer relevant biological function. We systematically applied SNPs2ChIP on publicly available datasets, including known GWAS associations from the GWAS catalogue and ChIP-seq peaks from a previously published study. Our approach to leverage latent patterns across genome-wide epigenetic datasets to infer the biological function will advance understanding of the genetics of human diseases by accelerating the interpretation of non-coding genomes.

DNA STEGANALYSIS USING DEEP RECURRENT NEURAL NETWORKS

Ho Bae[1], Byunghan Lee[2,3], Sunyoung Kwon[2,4], Sungroh Yoon[1,2,5]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University; [2]Electrical and Computer Engineering, Seoul National University; [3]Electronic and IT Media Engineering, Seoul National University of Science and Technology; [4]Clova AI Research, NAVER Corp; [5]ASRI and INMC, Seoul National University

Recent advances in next-generation sequencing technologies have facilitated the use of deoxyribonucleic acid (DNA) as a novel covert channels in steganography. There are various methods that exist in other domains to detect hidden messages in conventional covert channels. However, they have not been applied to DNA steganography. The current most common detection approaches, namely frequency analysis-based methods, often overlook important signals when directly applied to DNA steganography because those methods depend on the distribution of the number of sequence characters. To address this limitation, we propose a general sequence learning-based DNA steganalysis framework. The proposed approach learns the intrinsic distribution of coding and non-coding sequences and detects hidden messages by exploiting distribution variations after hiding these messages. Using deep recurrent neural networks (RNNs), our framework identifies the distribution variations by using the classification score to predict whether a sequence is to be a coding or non-coding sequence. We compare our proposed method to various existing methods and biological sequence analysis methods implemented on top of our framework. According to our experimental results, our approach delivers a robust detection performance compared to other tools.

LEARNING CONTEXTUAL HIERARCHICAL STRUCTURE OF MEDICAL CONCEPTS WITH POINCAIRÉ EMBEDDINGS TO CLARIFY PHENOTYPES

Brett K. Beaulieu-Jones, Isaac S. Kohane, Andrew L. Beam

Harvard Medical School

Biomedical association studies are increasingly done using clinical concepts, and in particular diagnostic codes from clinical data repositories as phenotypes. Clinical concepts can be represented in a meaningful, vector space using word embedding models. These embeddings allow for comparison between clinical concepts or for straightforward input to machine learning models. Using traditional approaches, good representations require high dimensionality, making downstream tasks such as visualization more difficult. We applied Poincaré embeddings in a 2-dimensional hyperbolic space to a large-scale administrative claims database and show performance comparable to 100-dimensional embeddings in a euclidean space. We then examine disease relationships under different disease contexts to better understand potential phenotypes.

EXPLORING MICRORNA REGULATION OF CANCER WITH CONTEXT-AWARE DEEP CANCER CLASSIFIER

Blake Pyman, Alireza Sedghi, Shekoofeh Azizi, Kathrin Tyryshkin, Neil Renwick, Parvin Mousavi

Queen's University

Background: MicroRNAs (miRNAs) are small, non-coding RNA that regulate gene expression through post-transcriptional silencing. Differential expression observed in miRNAs, combined with advancements in deep learning (DL), have the potential to improve cancer classification by modelling non-linear miRNA-phenotype associations. We propose a novel miRNA-based deep cancer classifier (DCC) incorporating genomic and hierarchical tissue annotation, capable of accurately predicting the presence of cancer in wide range of human tissues. Methods: miRNA expression profiles were analyzed for 1746 neoplastic and 3871 normal samples, across 26 types of cancer involving six organ sub-structures and 68 cell types. miRNAs were ranked and filtered using a specificity score representing their information content in relation to neoplasticity, incorporating 3 levels of hierarchical biological annotation. A DL architecture composed of stacked autoencoders (AE) and a multi-layer perceptron (MLP) was trained to predict neoplasticity using 497 abundant and informative miRNAs. Additional DCCs were trained using expression of miRNA cistrons and sequence families, and combined as a diagnostic ensemble. Important miRNAs were identified using backpropagation, and analyzed in Cytoscape using iCTNet and BiNGO. Results: Nested four-fold cross-validation was used to assess the performance of the DL model. The model achieved an accuracy, AUC/ROC, sensitivity, and specificity of 94.73\%, 98.6\%, 95.1\%, and 94.3\%, respectively. Conclusion: Deep autoencoder networks are a powerful tool for modelling complex miRNA-phenotype associations in cancer. The proposed DCC improves classification accuracy by learning from the biological context of both samples and miRNAs, using anatomical and genomic annotation. Analyzing the deep structure of DCCs with backpropagation can also facilitate biological discovery, by performing gene ontology searches on the most highly significant features.

ESTIMATING CLASSIFICATION ACCURACY IN POSITIVE-UNLABELED LEARNING: CHARACTERIZATION AND CORRECTION STRATEGIES

Rashika Ramola, Shantanu Jain, Predrag Radivojac

Northeastern University

Accurately estimating performance accuracy of machine learning classifiers is of fundamental importance in biomedical research with potentially societal consequences upon the deployment of best-performing tools in everyday life. Although classification has been extensively studied over the past decades, there remain understudied problems when the training data violate the main statistical assumptions relied upon for accurate learning and model characterization. This particularly holds true in the open world setting where observations of a phenomenon generally guarantee its presence but the absence of such evidence cannot be interpreted as the evidence of its absence. Learning from such data is often referred to as positive-unlabeled learning, a form of semi-supervised learning where all labeled data belong to one (say, positive) class. To improve the best practices in the field, we here study the quality of estimated performance in positive-unlabeled learning in the biomedical domain. We provide evidence that such estimates can be wildly inaccurate, depending on the fraction of positive examples in the unlabeled data and the fraction of negative examples mislabeled as positives in the labeled data. We then present correction methods for four such measures and demonstrate that the knowledge or accurate estimates of class priors in the unlabeled data and noise in the labeled data are sufficient for the recovery of true classification performance. We provide theoretical support as well as empirical evidence for the efficacy of the new performance estimation methods.

EXTRACTING ALLELIC READ COUNTS FROM 250,000 HUMAN SEQUENCING RUNS IN SEQUENCE READ ARCHIVE

Brian Tsui, Michelle Dow, Dylan Skola, Hannah Carter

Department of Medicine, University of California San Diego, 9500 Gilman Drive, San Diego, California 92093, USA

The Sequence Read Archive (SRA) contains over one million publicly available sequencing runs from various studies using a variety of sequencing library strategies. These data inherently contain information about underlying genomic sequence variants which we exploit to extract allelic read counts on an unprecedented scale. We reprocessed over 250,000 human sequencing runs (>1000 TB data worth of raw sequence data) into a single unified dataset of allelic read counts for nearly 300,000 variants of biomedical relevance curated by NCBI dbSNP, where germline variants were detected in a median of 912 sequencing runs, and somatic variants were detected in a median of 4,876 sequencing runs, suggesting that this dataset facilitates identification of sequencing runs that harbor variants of interest. Allelic read counts obtained using a targeted alignment were very similar to read counts obtained from whole-genome alignment. Analyzing allelic read count data for matched DNA and RNA samples from tumors, we find that RNA-seq can also recover variants identified by Whole Exome Sequencing (WXS), suggesting that reprocessed allelic read counts can support variant detection across different library strategies in SRA. This study provides a rich database of known human variants across SRA samples that can support future meta-analyses of human sequence variation.

AUTOMATIC HUMAN-LIKE MINING AND CONSTRUCTING RELIABLE GENETIC ASSOCIATION DATABASE WITH DEEP REINFORCEMENT LEARNING

Haohan Wang[1], Xiang Liu[2], Yifeng Tao[1], Wenting Ye[1], Qiao Jin[3], William W. Cohen[4], Eric P. Xing[5]

[1]Carnegie Mellon University, [2]Chinese University of Hong Kong, [3]Tsinghua University, [4]Google AI, [5]Pettum Inc

The increasing amount of scientific literature in biological and biomedical science research has created a challenge in the continuous and reliable curation of the latest knowledge discovered, and automatic biomedical text-mining has been one of the answers to this challenge. In this paper, we aim to further improve the reliability of biomedical text-mining by training the system to directly simulate the human behaviors such as querying the PubMed, selecting articles from queried results, and reading selected articles for knowledge. We take advantage of the efficiency of biomedical text-mining, the flexibility of deep reinforcement learning, and the massive amount of knowledge collected in UMLS into an integrative artificial intelligent reader that can automatically identify the authentic articles and effectively acquire the knowledge conveyed in the articles. We construct a system, whose current primary task is to build the genetic association database between genes and complex traits of the human. Our contributions in this paper are three-fold: 1) We propose to improve the reliability of text-mining by building a system that can directly simulate the behavior of a researcher, and we develop corresponding methods, such as Bi-directional LSTM for text mining and Deep Q-Network for organizing behaviors. 2) We demonstrate the effectiveness of our system with an example in constructing a genetic association database. 3) We release our implementation as a generic framework for researchers in the community to conveniently construct other databases.

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

PROCEEDINGS PAPER WITH POSTER PRESENTATION

INFLUENCE OF TISSUE CONTEXT ON GENE PRIORITIZATION FOR PREDICTED TRANSCRIPTOME-WIDE ASSOCIATION STUDIES

Binglan Li[1], Yogasudha Veturi[1], Yuki Bradford[1], Shefali S. Verma[1], Anurag Verma[1], Anastasia M. Lucas[1], David W. Haas[2], **Marylyn D. Ritchie**[1]

[1]University of Pennsylvania, [2]Vanderbilt University

Transcriptome-wide association studies (TWAS) have recently gained great attention due to their ability to prioritize complex trait-associated genes and promote potential therapeutics development for complex human diseases. TWAS integrates genotypic data with expression quantitative trait loci (eQTLs) to predict genetically regulated gene expression components and associates predictions with a trait of interest. As such, TWAS can prioritize genes whose differential expressions contribute to the trait of interest and provide mechanistic explanation of complex trait(s). Tissue-specific eQTL information grants TWAS the ability to perform association analysis on tissues whose gene expression profiles are otherwise hard to obtain, such as liver and heart. However, as eQTLs are tissue context-dependent, whether and how the tissue-specificity of eQTLs influences TWAS gene prioritization has not been fully investigated. In this study, we addressed this question by adopting two distinct TWAS methods, PrediXcan and UTMOST, which assume single tissue and integrative tissue effects of eQTLs, respectively. Thirty-eight baseline laboratory traits in 4,360 antiretroviral treatment-naïve individuals from the AIDS Clinical Trials Group (ACTG) studies comprised the input dataset for TWAS. We performed TWAS in a tissue-specific manner and obtained a total of 430 significant gene-trait associations (q-value < 0.05) across multiple tissues. Single tissue-based analysis by PrediXcan contributed 116 of the 430 associations including 64 unique gene-trait pairs in 28 tissues. Integrative tissue-based analysis by UTMOST found the other 314 significant associations that include 50 unique gene-trait pairs across all 44 tissues. Both analyses were able to replicate some associations identified in past variant-based genome-wide association studies (GWAS), such as high-density lipoprotein (HDL) and CETP (PrediXcan, q-value = 3.2e-16). Both analyses also identified novel associations. Moreover, single tissue-based and integrative tissue-based analysis shared 11 of 103 unique gene-trait pairs, for example, PSRC1-low-density lipoprotein (PrediXcan's lowest q-value = 8.5e-06; UTMOST's lowest q-value = 1.8e-05). This study suggests that single tissue-based analysis may have performed better at discovering gene-trait associations when combining results from all tissues. Integrative tissue-based analysis was better at prioritizing genes in multiple tissues and in trait-related tissue. Additional exploration is needed to confirm this conclusion. Finally, although single tissue-based and integrative tissue-based analysis shared significant novel discoveries, tissue context-dependency of eQTLs impacted TWAS gene prioritization. This study provides preliminary data to support continued work on tissue context-dependency of eQTL studies and TWAS.

SINGLE CELL ANALYSIS – WHAT IS THE FUTURE?

PROCEEDINGS PAPER WITH POSTER PRESENTATION

SHALLOW SPARSELY-CONNECTED AUTOENCODERS FOR GENE SET PROJECTION

Maxwell P. Gold, Alexander LeNail, Ernest Fraenkel

Massachusetts Institute of Technology

When analyzing biological data, it can be helpful to consider gene sets, or predefined groups of biologically related genes. Methods exist for identifying gene sets that are differential between conditions, but large public datasets from consortium projects and single-cell RNA-Sequencing have opened the door for gene set analysis using more sophisticated machine learning techniques, such as autoencoders and variational autoencoders. We present shallow sparsely-connected autoencoders (SSCAs) and variational autoencoders (SSCVAs) as tools for projecting gene-level data onto gene sets. We tested these approaches on single-cell RNA-Sequencing data from blood cells and on RNA- Sequencing data from breast cancer patients. Both SSCA and SSCVA can recover known biological features from these datasets and the SSCVA method often outperforms SSCA (and six existing gene set scoring algorithms) on classification and prediction tasks.

WHEN BIOLOGY GETS PERSONAL: HIDDEN CHALLENGES OF PRIVACY AND ETHICS IN BIOLOGICAL BIG DATA

PROCEEDINGS PAPER WITH POSTER PRESENTATION

IMPLEMENTING A UNIVERSAL INFORMED CONSENT PROCESS FOR THE ALL OF US RESEARCH PROGRAM

Megan Doerr[1], Shira Grayson[1], Sarah Moore[1], Christine Suver[1], John Wilbanks[1], Jennifer Wagner[2]

[1]Sage Bionetworks, [2]Center for Translational Bioethics & Health Care Policy Geisinger

The United States' All of Us Research Program is a longitudinal research initiative with ambitious national recruitment goals, including of populations traditionally underrepresented in biomedical research, many of whom have high geographic mobility. The program has a distributed infrastructure, with key programmatic resources spread across the US. Given its planned duration and geographic reach both in terms of recruitment and programmatic resources, a diversity of state and territory laws might apply to the program over time as well as to the determination of participants' rights. Here we present a listing and discussion of state and territory guidance and regulation of specific relevance to the program, and our approach to their incorporation within the program's informed consent processes.

GENERAL

POSTER PRESENTATIONS

A CONVOLUTIONAL NEURAL NET PREDICTS BINDING PROPERTIES OF AN ANTIBODY LIBRARY

Rishi Bedi, **Rachel Hovde**, Jacob Glanville

Distributed Bio

Research by Glanville et al. described a method that enabled TCRs of the adaptive immune system to be clustered into specificity groups and allowed de novo design of TCRs with a particular specificity. In this study, we apply deep learning methods to perform characterization and engineering of antibodies.

To generate enough data to address this question with machine learning methods, we created a computationally-optimized antibody library capable of generating thousands of high affinity hits against any antigen. By robotically panning 11 antigens in replicate against the library, we generated, sequenced, and validated a dataset of over 55,000 unique high affinity binders.

To characterize the functional properties of this library, we train a convolutional neural network to predict the binding specificity of each clone. Our model outperforms alternative approaches and successfully predicts binding specificity in held-out, increasingly dissimilar test sets.

Using the trained model to perform optimization on the input sequence, we generate characteristic class examples, as well as "fooling sequences" that represent the boundaries between pairs of binding specificities. We use the real-valued output of the convolutional and linear layers of the network as an embedding and demonstrate physically-meaningful clustering. These techniques let us assess the contribution of particular motifs to the lock-and-key interaction with the target antigen, and enable virtual "epitope binning" to distinguish antibodies in our library that bind similar epitopes. This enables future work in virtual mutagenesis, where we leverage these insights to generate antibodies that exhibit desirable binding properties.

CNVAR: A SOFTWARE TOOL FOR GENOTYPING CYP2D6 USING SHORT READ NEXT GENERATION SEQUENCING TECHNOLOGY

John Logan Black III MD[1], Hugues Sicotte PhD[1], Sandra E. Peterson[1], Kimberley J. Harris[1], Liewei Wang MD PhD[1], Steven Scherer PhD[2], Eric Boerwinkle PhD[2], Richard A. Gibbs PhD[2], Suzette J. Bielinski PhD[1], Richard Weinshilboum MD[1]

[1]Mayo Clinic, [2]Baylor College of Medicine

Introduction: CYP2D6 is an important pharmacogene involved in the metabolism of many medications. CYP2D6 is known to have numerous copy number variations (CNV) including gene duplications/multiplications, gene deletion, and hybrid genes involving the pseudogene, CYP2D7. Software that enables the genotyping of CYP2D6 from short read next generation sequencing (NGS) is urgently needed to cost-effectively and accurately determine clinical CYP2D6 phenotypes.

Methods: Modelling of expected ratios for specific gene regions with and without CNV was done based upon the known configurations of the CYP2D locus. This data was used to generate the CNVAR software which analyzes vcf and bam files to determine variant allelic ratios and read depth for all exons and the promoters of the CYP2D6 and CYP2D7 genes after NGS. The software uses statistical methods to detect the CNVs and employs multiple quality metrics to determine the best fit for possible genotype solutions. It also detects named haplotypes plus any novel variants. CNVAR was previously validated against 500 samples with known genotypes determined by targeted genotyping and Sanger sequencing. Samples sequenced as part of the Mayo Clinic Center for Individualized Medicine's RIGHT 10K Study for Pharmacogenomics are now being analyzed. Sequencing was done at Baylor College of Medicine's Human Genome Sequencing Center using the reagent called PGx-seq and analysis of the CYP2D6 sequence results is being performed in the Personalized Genomics Laboratory at Mayo Clinic.

Results: 6921 samples have been analyzed using the CNVAR software to derive CYP2D6 diplotypes. 968 (14%) samples had quality flags indicating either unexpected allele frequencies, CNV ratios, a novel variant was detected, or several diplotype solutions fit the findings equally well. 102 (1.5%) samples were determined to have novel variants or novel hybrid genes. All of the remaining samples, except 55 (0.79%), could be resolved by visual inspection of CNVAR outputs. These 55 remaining samples were referred for additional Sanger sequencing to determine the actual diplotype and quantitative rtPCR to determine actual copy number.

Conclusions: CNVAR is a software tool which can detect CYP2D6 diplotypes, CNVs and hybrid genes from NGS short read technology. The software identifies samples that cannot be genotyped with certainty so that additional evaluation can be performed to derive the actual genotype. Novel variants and hybrid alleles were also identified so that variant curation and classification could be done.

This work was supported by Mayo Clinic Center for Individualized Medicine and the Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, National Institutes of Health grants U19 GM61388 (The Pharmacogenomics Research Network), R01 GM28157, U01 HG005137, R01 GM125633, R01 AG034676 (The Rochester Epidemiology Project), and U01 HG06379 and U01 HG06379 Supplement (The Electronic Medical Record and Genomics (eMERGE) Network).

NETWORK ANALYSIS OF DISTINCT COHORTS ALLOWS FOR THE COMPARISON OF KEY BIOLOGICAL FUNCTIONS RELATED TO TB PATHOGENESIS

Carly Bobak, Meghan E. Muse, Alexander J. Titus, Brock C. Christensen, A. James O'Malley, Jane E. Hill

Dartmouth College

Challenges with reproducibility of microarray data sets can limit the ability to analyze and interpret integrated gene expression data sets. One approach to tackle reproducibility across microarray datasets builds a multi-cohort framework using publicly available data to better mirror diverse populations seen in clinics. An alternative way of increasing the reproducibility of results is emphasizing underlying pathway or network level analyses. While differential expression of genes may vary between datasets and data analysis techniques, the biological processes underlying gene expression are more robust. The results from these analyses can drive hypotheses regarding the biological mechanisms behind diseases.

We propose using a multi-cohort design and a pathway-level gene expression analysis to identify key biological processes in active Tuberculosis (TB) disease. A multi-cohort approach is particularly important when analyzing TB because phenotypic presentation of the disease differs among patients, especially those who are co-infected with human immunodeficiency virus (HIV), or among children. As such, these subgroups are often excluded from studies examining human gene expression array data. However, in 2016, 10% of incident TB cases were people living with HIV, and 10% were children, and despite the difficulty of studying these populations alongside adults, they make up a substantial proportion of both currently TB infected and the overall TB susceptible population.

To visualize differences across cohorts containing these TB subgroups, we use an approach called an EnrichmentMap which allows us to represent each distinct dataset in one network. We selected three representative publicly available datasets (n=1148) and used Differential Expression and Gene Set Enrichment Analysis. Gene sets which were significantly enriched became the nodes of the network, with edges representative of the overlap between these gene sets. The results of these combined analyses were used as an input to EnrichmentMap, to cluster and annotate important biological functions. The EnrichmentMap network identified many processes expected based on current TB knowledge, such as interferon-gamma activity (6 gene sets). As well, some other processes which represent potentially novel insights to the disease are identified. We examine one cluster of nodes related to DNA methylation (6 gene sets) in depth. The DNA methylation gene set within this cluster was strongly enriched in the dataset with no HIV+ patients (FDR=0.004) and appears to be enriched, although insignificantly so, in the two datasets including HIV+ patients (FDR=0.518, 0.879). Further unsupervised analysis of DNA methylation genes within these sets reveals clear clustering of active TB patients from those with latent TB infection, irrespective of HIV+. Thus, we theorize that while conventional methods would not implicate DNA methylation as playing a role in active TB infection, by comparing enrichments across datasets at the network level we can observe patterns in gene expression with a finer degree of granularity.

VARIATION IN OPIOID PRESCRIBING PATTERNS IN SURGICAL POPULATIONS

Soline M. Boussard[1], Marylyn D. Ritchie[2], Michelle Whirl-Carrillo[3], Tina Hernandez-Boussard[3], Teri E. Klein[3]

[1]Castilleja School, [2]University of Pennsylvania, [3]Stanford University

Introduction

In clinical settings, patients' response to opioids can vary by as much as 40-fold. Common opioids require metabolism by liver enzyme CYP-2D6 and considerable variation exists in the amount of CYP-2D6 produced by individuals. Therefore, pharmacogenomics may shed light as to how to address different responses by finding the most effective medication and dosage for each patient. As a first step at identifying opportunities for personalized pain management, we analyzed postoperative pain and opioid prescribing patterns across four common surgeries known for high postoperative pain.

Methods

We used EHRs to identify patients undergoing 4 surgeries (total knee replacement (TKA), thoracotomy, distal radius fracture, and mastectomy). The main outcomes were discharge pain medications and postoperative pain scores. This research was possible through the use of structured EHR data and the mapping of medications to ontologies. Patients were identified using procedural codes; pain scores (pain scores range from 0 to 10 with 10 being the most severe) were identified from flow sheets within the EHR, and discharge medications were mapped to RXNorm. Data were aggregated to the patient level. Pain scores were averaged across different time points. RStudio was used for statistical computing and graphics. Chi-square, t-tests and analysis of variance were used for statistical testing.

Results

A total of 63,500 patients were included. The mean age was 61.31(SD: 14.3), 65.3% were female, 62.1% were white and 13.3% were Hispanic/Latino ethnicity. On average, pain scores were lower at 30 days follow-up compared to pre-operative and patients received 4.1 different types of opioids during their inpatient stay, with a majority of patients switching between hydrocodone and oxycodone. Total knee replacement represented 61.6% followed by 20.0% thoracotomy, 16.4% mastectomy and 2.0% distal radius fracture. At discharge, the majority of patients received oxycodone (69.15%) and hydrocodone (15.29%). In mastectomy, 47.89% received hydrocodone and 44.05% received oxycodone. For TKA, 78.20% received oxycodone, followed by 8.90% receiving tramadol. Follow-up pain was similar across the 4 surgeries, however the follow-up pain differed by opioids received with patients on oxymorphone having the highest follow-up pain (6.24) and patients on propoxyphene having the lowest pain (1.29, $p<.0001$).

Discussion

In this study that examines post-operative outcomes and prescriptions in a real-world setting, opioid prescribing patterns varied significantly across surgery type. Our data suggest codeine was associated with lower follow-up pain in TKA compared to other opioids. This data from real-world evidence suggests that we can use such methodology to identify a cohort of patients that may be targeted for genotyping for personalized medicine. Targeting patients with poor pain relief from opioids that require CYP-2D6 for activation could identify patients with gene variations that affect opioid metabolism. Future studies could look at what variants that could affect patients' metabolism for codeine.

REGIONAL HETEROGENEITY IN GENE EXPRESSION, REGULATION AND COHERENCE IN HIPPOCAMPUS AND DORSOLATERAL PREFRONTAL CORTEX ACROSS DEVELOPMENT AND SCHIZOPHRENIA

Leonardo Collado-Torres[1], Emily E. Burke[1], Amy Peterson[1], Joo Heon Shin[1], RIchard E. Straub[1], Anandita Rajpurohit[1], Stephen A. Semick[1], William S. Ulrich[1], BrainSeq Consortium, Cristian Valencia[1], Ran Tao[1], Amy Deep-Soboslay[1], Thomas M. Hyde[1], Joel E. Kleinman[1], Daniel R Weinberger[1,+,] Andrew E. Jaffe[1,+]

[1]*Lieber Institute for Brain Development, Baltimore, MD, USA*

Background: We previously identified widespread genetic, developmental, and schizophrenia-associated (SCZD) changes in polyadenylated RNAs in the dorsolateral prefrontal cortex (DLPFC), but the landscape of hippocampal (HIPPO) expression using RNA sequencing is less well-explored.

Methods: We performed RNA-seq using RiboZero on 900 RNA-seq samples across 551 individuals (SCZD N=286) in DLPFC (N=453) and HIPPO (N=447). We quantified expression of multiple feature summarizations of the Gencode v25 reference transcriptome, including genes, exons and splice junctions. Within and across brain regions, we modeled age-related changes in controls using linear splines, integrated genetic data to perform expression quantitative trait loci (eQTL) analyses, and performed differential expression analyses controlling for observed and latent confounders.

Results: We identified widespread developmental regulation between the DLPFC and HIPPO over aging with 10,839 genes differentially expressed (Bonferroni < 0.01) and replicating in BrainSpan (n = 79 tissue samples, DLPFC=40, HIPPO=39). Of these genes, 5,982 (55%) contain differentially expressed exons and splice junctions that replicated in BrainSpan. By extending quality surrogate variable analysis (qSVA) to multiple brain regions, we identified 48 and 245 differentially expressed genes (DEG) by SCZD diagnosis (FDR<5%) in HIPPO and DLPFC, respectively, with surprisingly minimal overlap in DEG between the two brain regions. We further identified 205,618 brain region-dependent eQTLs (FDR<1%) and found that 124 GWAS risk loci contain eQTLs in at least one of the regions. We also identify potential molecular correlates of in vivo evidence of altered prefrontal-hippocampal functional coherence in schizophrenia. Through our eQTL browser resource http://eqtl.brainseq.org/ we have made all eQTLs sets available for further exploration.

Discussion: We show extensive regional specificity of developmental and genetic regulation, and SCZD-associated expression differences between HIPPO and DLPFC. These results underscore the complexity and regional heterogeneity of the transcriptional correlates of schizophrenia, and suggest future schizophrenia therapeutics may need to target molecular pathologies localized to specific brain regions.

FULL-LENGTH SEQUENCE ASSEMBLY AND CHARACTERIZATION OF HIGHLY PURIFIED CIRCRNA ISOFORMS

Supriyo De, Amaresh C. Panda, Myriam Gorospe

Laboratory of Genetics and Genomics, National Institute on Aging IRP, NIH

Circular RNAs are a large heterogenous class of highly stable noncoding RNAs but they are poorly characterized. Many software tools exist for identifying circular RNAs by finding their circularizing junctions, but very little is known about the sequence of their full length or their isoforms/alternately spliced forms. The assembly and characterization of isoforms is also limited by the lack of methodologies to extract highly pure circRNAs. While exoribonuclease (RNase R) treatment is widely used to degrade linear RNAs and enrich circRNAs from total RNA, it does not efficiently eliminate all linear RNAs. This limitation complicates the assembly process to get full-length circRNAs. Here we describe a novel method for isolating highly pure circRNA populations involving RNase R treatment followed by Polyadenylation and poly(A)+ RNA Depletion (RPAD), which removes linear RNA to near completion. Once the RNA population is highly enriched, sequence assembly algorithms such as Cufflinks can be used to identify the body of the circRNA, while the circularizing/back-spliced junctions can be found using many different software tools such as Circexplorer, CIRI etc. High-throughput sequencing of RNA prepared using RPAD from human cervical carcinoma HeLa cells and mouse C2C12 myoblasts followed by this novel analysis pipeline led to identification of many circRNA isoforms with an identical back-splice sequence (circularizing junction) but with different body sizes and sequences. As one of the main functions of circRNAs is sponging regulatory RNAs and proteins, full-length characterization of circRNA isoforms will be critical for enabling the functional characterization of circRNAs.

Acknowledgement: This research was supported by Intramural Research Program of the National Institute on Aging, NIH.

A COMPREHENSIVE REVIEW AND ASSESSMENT OF EXISTING PATHWAY ANALYSIS APPROACHES

Tuan-Minh Nguyen[1], Adib Shafi[1], Tin Nguyen[2], **Sorin Draghici**[1]

[1]Dept of Computer Science, Wayne State University; [2]Dept of Computer Science, University of Nevada

In many high-throughput experiments, it is crucial to understand the biological mechanisms of genes and their products from expression data. Pathway analysis is a crucial step in any phenotype comparison because it allows us to gain insights into the underlying biological phenomena. Because of the importance of this type of analysis, more than 35 pathway analysis methods have been proposed so far. These can be categorized into two main categories: non-pathway topology based (non-TB) and topology-based (TB) approaches. Non-TB methods consider pathways as simple gene sets and ignore the position and role of the genes, as well as the direction and type of signals described by the pathway while TB methods include this additional information in the analysis. Although there are some review papers discussing this topic, there has been no study that systematically assesses the performances of the methods using an unbiased and large number of data sets available. Furthermore, the majority of the pathway analysis approaches rely on the assumption of uniformity of p-values under the null hypothesis, which is not always true. None of these existing reviews take the performances of the studied methods under the null into account in their comparisons. In order to provide an accurate and objective assessment so that researchers and biologists can choose a method suitable for their purpose, we provide an extensive analysis of 11 widely used pathway analysis methods from both non-TB and TB groups using 2601 samples from 75 human disease data sets and 8 methods using 121 samples from 11 knock-out mouse data sets. In addition, we investigate the extent to which each method is biased under the null hypothesis. Overall, the result shows TB methods perform better than non-TB methods since they take into consideration the topology information and signal propagation. Via permutation and bootstrap, we discover another critical conclusion that most if not all listed approaches are biased and produce very skewed results under the null.

A NEW PHYLOGENETIC SAMPLING METHOD USING GENERALIZED-ENSEMBLE ALGORITHM

Tetsu Furukawa, Hiroyuki Toh

Department of Biomedical Chemistry, School of Science and Technology, Kwansei-Gakuin University, Sanda, Hyogo, Japan 669-1337

Bayesian inference has been widely utilized for the evolutionary analysis including phylogenetic tree reconstruction, where Monte Carlo sampling such as Markov chain Monte Carlo (MCMC) or Metropolis-coupled MCMC (MC3) generates a posterior distribution. Monte Carlo sampling is also utilized for molecular simulation of biopolymers like proteins and DNA. One of the representative methods is the replica exchange algorithm, which is equivalent to MC3 in the molecular phylogeny. Besides the replica exchange algorithm, several different sampling methods have been developed for molecular simulation, which are collectively termed as the generalized ensemble algorithm. In this study, we examined the possibility to apply the other algorithms belonging to the generalized ensemble algorithm to the tree reconstruction, in order to develop more efficient sampling method for the molecular phylogeny. The program implemented with the other generalized ensemble algorithm was developed based on the source code of BEAST version 2.5.1. To evaluate the performance, artificial alignments were generated, so that the posterior distributions of the corresponding trees are difficult to be regenerated by sampling, i.e. the distribution with multiple peaks. We applied our program and existing tools to the artificial data. Then, we compared the results such as the times required for the convergence and the degree of regeneration of the posterior distributions. The benefit and pitfalls of our program will be discussed based on the comparison.

CONVERGENT MECHANISMS PERTURBED BY SCATTERED SNPs SUSCEPTIBLE TO ALZHEIMER'S DISEASE

Jiali Han[1,2], Edwin Baldwin[1], Jin Zhou[3], Fei Yin[4,5], Haiquan Li[1,6,]

[1]University of Arizona, Department of Biosystems Engineering; [2]University of Arizona, Department of Systems and Industrial Engineering; [3]University of Arizona, Department of Public Health; [4]University of Arizona, Department of Pharmacology; [5]University of Arizona Center for Innovation in Brain Science; [6]University of Arizona Center for Biomedical Informatics and Biostatistics

Alzheimer's Disease (AD) is the most prevalent neurodegenerative disorder affecting approximately 50 million people worldwide. Genome-wide association studies (GWAS) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with AD, while the effect size of each individual SNP is largely modest. The molecular mechanisms underlying these associations are yet to be understood. Our recent genomic analysis focused on unveiling common downstream biological effectors of intergenic SNPs associated with AD, aiming to understand the interactive- and synergetic effects that the genetic variants across non-coding and intergenic regions are playing in the pathogenesis of AD. In this study, data from GWAS and expression quantitative trait locus (eQTL) studies by GTEx project are integrated, and downstream functional similarity between two SNPs is imputed using an enhanced multiscale information theoretic distance model [1]. The significance levels are determined through extensive permutations of the eQTL-derived multiscale network for mRNA overlap, functional similarity and shared biological processes [2]. Convergent molecular mechanisms based on gene ontology are prioritized at FDR<0.05.

The prioritized mechanistic network for AD renders several functional modules perturbed by either cis-eQTL or trans-eQTL elements, corresponding to multiple common mechanisms downstream of distinct eQTLs with some of them being cross-chromosome. For instance, SNPs on chromosomes six and one are both associated with antigen processing and presentation via regulating multiple human leukocyte antigen genes (e.g., HLA-DRB1 and HLA-DQA1) and cytokine genes, suggesting the genetic involvement of the immune systems and neuroinflammation in the pathogenesis of AD. SNPs on chromosome 17 and chromosome 19 co-regulate genes involved in synaptic transmission, which is essential for neurons communication and its dysfunction is known in AD leading to memory loss. Other than cross-chromosome SNPs, independent intergenic SNPs on the same chromosome also provide insights to AD genetic risks. A pair of SNPs on chromosome 17 is prioritized by our method through their convergent association with the MAPT gene, which encodes tau protein, regulates axon extension, and is known as a risk factor of a variety of neurodegenerative disorders including not only AD but also Frontotemporal dementia and Parkinson disease. Another pair of SNPs on chromosome 19 is prioritized by their common association with the ABCA7 gene, which regulates lipid metabolism across cellular membranes and is suggested to be susceptible loci for the late-onset AD.

This study suggests a new strategy connecting scattered AD-susceptible genetic variants with risk genes and convergent downstream mechanisms implicated in AD pathogenesis. The results will help to understand how genetic variants and underlying functional modules work interactively and systematically toward AD onset and could thus identify genetics-specific molecular targets and inspire new personalized therapeutic strategies.

[1] Li, H., et al. npj Genomic Medicine 1:16006, 2016.
[2] Han, J., et al. PSB, 2018, pp. 524-535.

IDENTIFICATION AND EVALUATION OF CO-EXPRESSION GENE NETWORKS FOR PACLITAXEL-INDUCED PERIPHERAL NEUROPATHY IN BREAST CANCER SURVIVORS

Kord M. Kober[1], Jon D. Levine[2], Judy Mastick[1], Bruce Cooper[1], Steven Paul[1], Christine Miaskowski[1]

[1]UCSF School of Nursing, [2]UCSF School of Medicine

Chronic chemotherapy-induced peripheral neuropathy (CIPN) is the most common and severe adverse drug reaction associated with neurotoxic chemotherapy (CTX) with prevalence rates that range from 30% to 70% in cancer survivors. No pharmacologic interventions are available to prevent CIPN. Lack of knowledge of the fundamental mechanisms that underlie CIPN thwart our efforts to develop interventions to prevent or treat it. Increased knowledge of CIPN's molecular mechanisms could identify therapeutic targets for this condition. Findings from animal studies suggest that a number of diverse mechanisms are involved in the development of chronic PIPN including damage to DRG cell bodies; microtubule associated toxicity; inflammation; distal axonal injury; damage to the peripheral vasculature; modulation of ion channels; and mitochondrial dysfunction. Taxol is a common CTX drug that is associated with the development of CIPN. Paclitaxel-induced peripheral neuropathy (PIPN) is the dose limiting toxicity of this CTX drug. The purpose of this pilot study was to evaluate for coordinated expression variations of genes in RNA extracted from peripheral blood from breast cancer survivors, and from these modules identify co-expressed genes that are associated with chronic PIPN. Gene expression in peripheral blood was assayed using RNA-seq in a sample of breast cancer (BC) survivors who did (n=25) and did not (n=25) develop PIPN. BC survivors with PIPN were significantly older; more likely to be unemployed; reported lower alcohol use; had a higher BMI and a poorer functional status; and had a higher number of lower extremity sites with loss of light touch, cold, and pain sensations, and higher vibration thresholds. No between group differences were found in the cumulative dose of paclitaxel received or in the percentage of patients who had a dose reduction or delay due to PIPN. Co-expression network analysis was performed to identify modules of genes with highly correlated expression using the top 5000 most variant genes. Thirteen color-coded modules were detected ranging in size from 36 to 1653 genes. The eigengenes of the "black" module (n=1653 genes) were significantly correlated with the CIPN phenotype (Pearson R2=0.224, p=0.02). GO enrichment was found in inflammation-related terms (e.g., C-C chemokine receptor activity, Chemokine-mediated signaling pathway, T cell co-stimulation). Functional protein association network analysis identified an enrichment of protein-protein interactions ($p<0.0002$) including highly connected genes that have previously been identified to be related to CIPN (i.e., G protein-coupled receptor 55, GPR55, and C-X-C Motif Chemokine Receptor 5, CXCR5). To our knowledge, this is the first study to apply systems biology approaches using circulating blood RNA-seq data in a sample of breast cancer survivors with and without chronic PIPN. We revealed networks and candidate genes associated with chronic PIPN related to inflammation, and suggest genes for validation and as potential therapeutic targets.

VARIFI - WEB-BASED AUTOMATIC VARIANT IDENTIFICATION, FILTERING AND ANNOTATION OF AMPLICON SEQUENCING DATA

Milica Krunic[1], Peter Venhuizen[2], Leonhard Müllauer[3], Bettina Kaserer[3], Arndt von Haeseler[1,4]

[1]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Dr. Bohrgasse 9, 1030 Vienna, Austria; [2]Department of Applied Genetics und Cell Biology, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria; [3]Institute of Pathology, Medical University Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria; [4]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

Fast and affordable benchtop sequencers are becoming more important in improving personalized medical treatment. Still, distinguishing genetic variants between healthy and diseased individuals from sequencing errors remains a challenge. Here we present VARIFI, a pipeline for finding reliable genetic variants (SNPs and INDELs). We optimized parameters in VARIFI by analyzing more than 170 amplicon sequenced cancer samples produced on the Personal Genome Machine (PGM). In contrast to existing pipelines, VARIFI combines different analysis methods and, based on their concordance, assigns a confidence score to every identified variant. Furthermore, VARIFI applies variant filters for biases associated with the sequencing technologies (e.g. incorrectly called homopolymer-associated indels with Ion Torrent). VARIFI automatically extracts variant information from publicly available databases and incorporates methods for variant effect prediction. VARIFI requires only little computational experience and no in-house compute power since the analyses are done on our server. VARIFI is a web-based tool available at varifi.cibiv.univie.ac.at.

STATISTICAL INFERENCE RELIEF (STIR) FEATURE SELECTION

Trang T. Le[1], Ryan J. Urbanowicz[1], Jason H. Moore[1], Brett A. McKinney[2]

[1]Institute of Biomedical Informatics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA; [2]Tandy School of Computer Science, Department of Mathematics, University of Tulsa, Tulsa, OK

Motivation: Identifying relevant features in high-dimensional data can be challenging when their effect on an outcome may be obscured by a complex interaction architecture. Using nearest neighbors, Relief-based algorithms account for statistical interactions when selecting features. However, Relief-based estimators are non-parametric in the statistical sense that they do not have a parameterized model with an underlying probability distribution for the estimator, making it difficult to determine the statistical significance of Relief-based attribute estimates. Thus, a statistical inferential formalism is needed to avoid imposing arbitrary thresholds to select the most important features.

Method: We reconceptualize the Relief-based feature selection algorithm to create a new family of STatistical Inference Relief (STIR) estimators that retains the ability to identify interactions while incorporating sample variance of the nearest neighbor distances into the attribute importance estimation. This variance permits the calculation of statistical significance of features and adjustment for multiple testing of Relief-based scores. Specifically, we develop a pseudo t-test version of Relief-based algorithms for case-control data.

Results: We demonstrate the statistical power and control of type I error of the STIR family of feature selection methods on a panel of simulated data that exhibits properties reflected in real gene expression data, including main effects and network interaction effects. We showed that the statistical performance using STIR p-values is the same as using permutation p-values but much more computationally efficient. We compare the performance of STIR when the adaptive radius method is used as the nearest neighbor constructor with STIR when the fixed-k nearest neighbor constructor is used. Applying STIR to real RNA-Seq data from a study of major depressive disorder, we found that 32 significant STIR genes include all 8 significant genes from standard t-test. STIR genes outside of the intersection with t-test may be good candidates for interaction effects.

Conclusion: STIR is the first method to use a theoretical distribution to calculate the statistical significance of Relief attribute scores without the computational expense of permutation. This validates the STIR pseudo t-test and means one can use it instead of costly permutation testing. STIR formalism generalizes to all Relief-based neighbor finding algorithms, including MultiSURF. k=m/6 offers a better default than the pervasive use of k=10, which is an arbitrary choice in the early literature. Extensions of STIR will involve multi-class data, quantitative trait data (regression) and correction for covariates. Similarly, we envision regression-STIR to follow a linear model formalism. Future studies will apply STIR to GWAS as well as eQTL and other high dimensional data to identify interaction effects.

DEEP LEARNING-BASED LONGITUDINAL HETEROGENEOUS DATA INTEGRATION FRAMEWORK FOR AD-RELEVANT FEATURE EXTRACTION

Garam Lee[1], Kwangsik Nho[2], Byungkon Kang[1], Kyung-Ah Sohn[1], **Dokyoon Kim**[3]

[1]Ajou University, [2]Indiana University School of Medicine, [3]Geisinger

Alzheimer's disease (AD) is a progressive neurodegenerative condition marked by a decline in cognitive functions with no validated disease modifying treatment. It is critical for timely treatment to detect AD in its earlier stage before clinical manifestation. Mild cognitive impairment (MCI) is an intermediate stage between cognitively normal older adults and AD. To predict conversion from MCI to probable AD, we applied a deep learning approach, multimodal recurrent neural network. We developed an integrative framework that combines not only cross-sectional neuroimaging biomarkers at baseline but also longitudinal cerebrospinal fluid (CSF) and cognitive performance biomarkers obtained from the Alzheimer's Disease Neuroimaging Initiative cohort (ADNI). The proposed framework integrated longitudinal multi-domain data with missing values. The python package LIFAD (Deep learning-based Longitudinal heterogeneous data Integration Framework for AD-relevant feature extraction) provides pre-constructed deep learning architecture for a classification task. Our results showed that 1) our prediction model for MCI conversion to AD yielded up to 75% accuracy (area under the curve (AUC)= 0.83) when using only a single modality of data separately; and 2) our prediction model achieved the best performance with 80% accuracy (AUC= 0.86) when incorporating longitudinal multi-domain data. A multi-modal deep learning approach has potential to identify persons at risk of developing AD who might benefit most from a clinical trial or as a stratification approach within clinical trials.

MICROBIOME ANALYSIS OF UNEXPLAINED CASES OF PNEUMONIA IN SOUTH KOREA

Sooyeon Lim, Jae Kyung Lee, Ji Yun Noh, Woo Joo Kim

Department of Internal Medicine, Guro Hospital, Korea University

Nasal swab samples were obtained from patients with symptoms of pneumonia through the tertiary hospital-based influenza surveillance system in South Korea during 2011-2017. Although the symptoms were suspected to be of viral cause pneumonia, collected samples were confirmed negative, using the respiratory virus panel, for 16 common respiratory pathogens, in addition to the following five viruses: Enterovirus D68, WU polyomavirus, KI polyomavirus, Parechovirus type 1, 3, 6, and Pteropine orthoreovirus. Therefore, 16S rRNA screening was performed to study the microbiome community of the patients. V3 and V4 sequences of 16S rRNA were obtained using Nextera XT DNA library preparation kit and MiSeq Reagent Kit v3(Illumina). Microbiome profiles of 92 patient samples were obtained through Illumina MiSeq. The total taxonomic composition of the samples consisted of 99 bacterial genus, whose sequences were detected in more than 1% of the samples. Common bacterial pathogens were present as either single pathogen or in combination with other organisms in the patient samples. Although samples collected were different in conditions, such as age, gender, location, and season, common dominant genus of bacteria commonly known as pathogens were revealed. The most dominant genera of bacteria were the following: Streptococcus, Corynebactierum, Haemophilus, Rhizobium. Based on comparative analysis of genus compositions are similar but demonstrated the difference in microbial composition between age groups. We tried to isolation dominant colonies through the media culture for whole genome sequencing and isolated single colony and 8 species are identified using sanger sequencing. After more isolation of single colonies, we will focused on whole genome sequencing to find out reason of pneumonia symptoms in detail.

POTRA: PATHWAY ANALYSIS OF CANCER GENOMICS DATA IN THE CLOUD

Margaret Linan[1,2], Junwen Wang[1,2], **Valentin Dinu**[1,2]

[1]Department of Biomedical Informatics, Arizona State University, Scottsdale, Arizona, USA; [2]Department of Health Sciences, Mayo Clinic, Scottsdale, Arizona, USA

We have recently developed PoTRA (Pathways of Topological Rank Analysis), a novel algorithm that uses the Google Search PageRank algorithm to identify biological pathways involved in cancer. The analytical approach is motivated by the observation that loss of connectivity is a common topological trait of gene regulatory networks in cancer. We leveraged the Cancer Genomics Cloud environment and applied PoTRA to analyze The Cancer Genome Atlas (TCGA) genomic data, a high-quality publicly available data set of tumor and matched normal samples. The top most influential pathways and most dysregulated pathways in 17 TCGA projects were found, using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database. Overall, pathways in cancer is the most common dysregulated pathway, and the MAPK signaling pathway is the most influential, while the purine metabolism pathway is the most significantly dysregulated metabolic pathway. Additionally, genomic analysis workflows were created using docker and rabix for the detection of mRNA mediated dysregulated pathways in the open access TCGA repository with the PoTRA tool in the CGC platform. Our approach illustrates the advantages of employing powerful computational methods to analyze large genomic data sets with the aim of improving our understanding of cancer and identifying better diagnoses and treatments.

EVALUATING CELL LINES AS MODELS FOR METASTATIC CANCER THROUGH INTEGRATIVE ANALYSIS OF OPEN GENOMIC DATA

Ke Liu[1], Patrick A. Newbury[1], Benjamin S. Glicksberg[2], William Zeng[2], Eran R. Andrechek[3], **Bin Chen**[1]

[1]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA; [2]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA; [3]Department of Physiology, Michigan State University, East Lansing, MI, USA

Metastasis is the most common cause of cancer-related death and, as such, there is an urgent need to discover new therapies to treat metastasized cancers. Cancer cell lines are widely-used models to study cancer biology and test drug candidates. However, it is still unknown to what extent they adequately resemble the disease in patients. The recent accumulation of large-scale genomic data in cell lines, mouse models, and patient tissue samples provides an unprecedented opportunity to evaluate the suitability of cell lines for metastatic cancer research. In this work, we used breast cancer as a case study. The comprehensive comparison of the genetic profiles of 57 breast cancer cell lines with those of metastatic breast cancer samples revealed substantial genetic differences. In addition, we identified cell lines that more closely resemble different subtypes of metastatic breast cancer. Surprisingly, a combined analysis of mutation, copy number variation and gene expression data suggested that MDA-MB-231, the most commonly used triple negative cell line for metastatic breast cancer research, had little genomic similarity with Basal-like metastatic breast cancer samples. We further compared cell lines with organoids, a new type of preclinical model which are becoming more popular in recent years. We found that organoids outperformed cell lines in resembling the transcriptome of metastatic breast cancer samples. However, additional differential expression analysis suggested that both types of models could not mimic the effects of tumor microenvironment and meanwhile had their own bias towards modeling specific biological processes. Our work provides a guide of cell line selection in metastasis-related study and sheds light on the potential of organoids in translational research.

PATHWAY ANALYSIS OF EHR AND NON-EHR-BASED GWAS CONNECTS LIPID METABOLISM TO THE IMMUNE RESPONSE

Jason E. Miller[1], Thomas J. Hoffmann[2,3], Elizabeth Theusch[4], Carlos Iribarren[5], Marisa W. Medina[4], Neil Risch[2,3,5], Ronald M. Krauss[4], Marylyn D. Ritchie[1]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA; [2]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA; [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA; [4]Children's Hospital Oakland Research Institute, Oakland, CA, USA; [5]Division of Research, Kaiser Permanente, Northern California, Oakland, CA, USA

Pathway-analysis is a commonly used method to interpret genome-wide association study (GWAS) results. Recently it has been illustrated that electronic-health-record (EHR) data from a single-cohort can be used to perform GWAS. However, it is unclear how this new study design might affect replication of pathway-level results when compared to a non-EHR-based GWAS. It is also unclear how an EHR-based study will affect downstream analyses such as the identification of genes that are associated with said pathways. We propose evaluating the pathway-level similarities from analyses of two separate GWAS studies that used different methodologies to investigate the same traits. Here, we employ the software PARIS (Pathway Analysis by Randomization Incorporating Structure) to compare summary-level results across studies, thus making it more generalizable. PARIS generates randomized collections of features which mimic pathways to calculate empirical p-values. This process reduces type I error and the multiple testing burden. We compared EHR to non-EHR-based GWAS results using four different lipid traits: low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides (TG), and total cholesterol (TC). The data came from two GWAS, the Genetic Epidemiology Resource on Adult Health and Aging (GERA), a single-cohort EHR-based GWAS, and the Global Lipids Genetics Consortium (GLGC), which used a meta-analysis study design. KEGG pathways expected to explain variation in lipid values such as "cholesterol metabolism" and "PPAR signaling pathway" were identified from both studies. Moreover, there was a significant overlap between the pathways identified between studies for the same traits ($p < 1x10^{-14}$). Thus, specific pathways can be replicated across distinct cohorts and study designs. Several pathways made up of genes whose proteins are important for an immune response were identified in both datasets and across multiple lipid traits. To see if lipid modifying therapy affects the same pathways of interest, we performed pathway analysis of CAP RNA-seq expression from Theusch, E., et al., 2016, which measured expression in immortalized cells pre and post statin exposure. Among the pathways represented in both the PARIS results ($p < 0.01$) and LCL RNA-seq gene set enrichment results (FDR < 25%) are "cholesterol metabolism" (CM) and "Hepatitis C" (HC) pathways. Hepatitis C virus (HCV) infection can cause chronic liver disease and is associated with a host of lipid and lipoprotein metabolic disorders. PARIS can also identify genes that are statistically significant within each pathway. Interestingly, in both the LDL and TC GWAS, the gene that was significantly associated with both the CM and HC pathways was low density lipoprotein receptor or LDLR, a gene that affects both lipid metabolism and HC viral activity. Statins, in combination with other therapies, can increase efficacy of antiviral therapy by blocking viral replication. Our results highlight the need for further investigation into how genetic variation affects outcomes from the treatment of HCV with statins, particularly with respect to loci associated with lipid traits. In conclusion, pathway-level analysis of GWAS summary-level results can be used to characterize similarities across EHR and non-EHR-based studies and improve biological interpretation.

META-ANALYSIS OF HETEROGENEITY AND BATCH EFFECTS IN THE A549 CELL LINE

Abigail Moore, John Castorino

School of Natural Sciences, Hampshire College, Amherst, MA

Meta-analysis of RNA-seq data offers the opportunity to increase reproducibility by integrating data from multiple studies. Such analyses are challenged by heterogenous cell culture and RNA-seq techniques, which may confound or hide true biological findings. Thus, we sought to identify batch characteristics that most significantly affect gene expression in a cell line common to lung cancer and viral studies. We queried the NCBI GEO for RNA-seq data from the A549 cell line and filtered the results for paired-end data obtained via total RNA extraction. Across eight studies, we downloaded raw RNA-seq data for 23 untreated samples and collected corresponding metadata. Differential expression analysis with Salmon, TXimport and edgeR identified 3,802 differentially expressed genes (at least twofold-change, FDR < 0.05). Principal variant component analysis revealed that media choice alone explains 54% of expression variation within 139 differentially expressed lung cancer prognostic genes. Our findings highlight the impact of specific batch effects on biologically significant genes. In future work, hope to extend this analysis to consider single nucleotide variants.

HYPERPARAMETER TUNING FOR CHIP-SEQ PEAK CALLING SOFTWARE TOOLS USING PARALLELIZED BAYESIAN OPTIMIZATION.

Dongpin Oh, **Jinhee Lee**, Seonghyeon Kim, Dohyeon Lee, Dongwon Choo, Giltae Song

School of Computer Science and Engineering, Pusan National University

ChIP-Seq is widely used to understand protein-DNA interaction and gene regulation. In ChiP-seq data analysis, identifying peak signals is one of core computational steps, but most existing software tools still suffer from large portion of false positive calls owing to sequencing errors and bias, in part caused by copy number variations. ChiP-seq analysis tools require hyperparameters set by users depending on sequencing quality and copy number variation rate. However, it is hard for users to know the valid values of the hyperparameters before running the software tools. In addition, we would have more false positive peak calls for given ChiP-seq data if the hyperparameters of peak calling tools are less than optimal. In this study, we develop a software pipeline for identifying the optimal values of the hyperparameters in major ChiP-seq peak calling tools. First we collect ChiP-seq data whose peak signals are labeled manually by experts. These data are used as training data in our hyperparameter tuning. Second we define an objective function to measure the accuracy of peak calling results. Then we learn optimal hyperparameters using these training data and objective function based on Bayesian optimization. We use Matern5/2 kernel function for the optimization and Monte Carlo Markov Chain for parallel processing. We validate our approach using our collection of ChiP-seq data labeled for around 2,000 genomic segments including peaks or no peaks. We apply our software pipeline for major ChiP-seq peak calling tools such as MACS, SICER, HOMER, and PeakSeq.

CROSS-STUDY META-ANALYSIS IDENTIFIES ALTERED BACTERIAL STRAINS SEPARATING RESPONDER AND NON-RESPONDER POPULATIONS ACROSS MULTIPLE CHECKPOINT-INHIBITOR THERAPY DATASETS

Jayamary Divya Ravichandar, Erica Rutherford, Yonggan Wu, Thomas Weinmaier, Cheryl-Emiliane Chow, Shoko Iwai, Helena Kiefel, Kareem Graham, Karim Dabbagh, Todd DeSantis

Second Genome

The gut microbiota has emerged as an important modulator in cancer progression and a growing body of evidence supports the influence of gut microbiota on response to cancer therapy, especially in the context of checkpoint inhibitor therapy. While several studies present insight into the landscape of microbial shifts modulating response to checkpoint inhibitors, they may be unduly influenced by cohort, sequencing-technology, and data analysis methods. Further, individual studies are often under-powered to detect microbes differentially abundant in responder and non-responder populations, which can limit therapeutic development. Key to microbiome-based drug discovery is the identification of proteins with therapeutic potential that are efficacious across cohorts. Herein, existing published datasets in the checkpoint-inhibitor space were mined and integrated via a cross-study meta-analysis to identify bacterial strains separating responder and non-responder populations.

We compared the baseline gut microbiota associated with stool samples collected from five discrete cancer patient cohorts undergoing checkpoint-inhibitor therapy. Samples were sequenced on one or more technologies (Illumina 16S NGS, 454 16S NGS, and Illumina shotgun metagenomics) and a total of seven publicly-available datasets were analyzed herein. Leveraging our multi-faceted bioinformatics platform, which enables appropriate method-specific quality filtering and statistical testing to identify differentially abundant bacteria at the strain-level, we were able to successfully integrate analysis results across multiple microbiome-profiling technologies. We performed a random effects model based meta-analysis and identified strains that were concordantly enriched in responder populations across datasets. In a separate analysis we also applied natural language processing to the text of cancer checkpoint inhibitor studies (available in Pubmed) in order to obtain additional insights about the microbiome and strains of interest from publications with no raw data available. The strains identified herein present opportunities for mining proteins with potential to improve response to checkpoint inhibitors.

This cross study meta-analysis demonstrates the power of Second Genome's bioinformatics pipeline to leverage publicly available datasets and systematically integrate microbial shifts not only across samples from multiple cohorts but also across samples sequenced on different technologies. Our in-house strain database that enables taxonomic annotation down to the strain-level allowed for comparison of fine-grained bacterial identities across datasets, resolving a key challenge with microbiome meta-analysis. This systematic and statistically-driven integration of datasets enabled identification of strains associated with response across multiple responder populations that were not previously reported in the independent analysis of these datasets.

A HYPOTHESIS OF THE STABILIZING ROLE OF ALU EXPANSION VIA HOMOLOGY DIRECTED REPAIR OF SPONTANEOUS DNA DOUBLE STRANDED BREAKS

Tanmoy Roychowdhury, **Alexej Abyzov**

Mayo Clinic

Structural variations (SVs) in the human genome originate from different mechanisms related to DNA repair, replication errors, and retrotransposition. Our analyses of 26,927 SVs from the 1000 Genomes Project revealed differential distributions and consequences of SVs of different origin, e.g., deletions from non-allelic homologous recombination (NAHR) are more prone to disrupt chromatin organization while processed pseudogenes can create accessible chromatin. Spontaneous double stranded breaks (DSBs) are the best predictor of enrichment of NAHR deletions in open chromatin. This evidence, along with strong physical interaction of NAHR breakpoints belonging to the same deletion suggests that majority of NAHR deletions are non-meiotic i.e., originate from errors during homology directed repair (HDR) of spontaneous DSBs. In turn, the origin of the spontaneous DSBs is associated with transcription factor binding in accessible chromatin revealing the vulnerability of functional, open chromatin. The chromatin itself is enriched with repeats, particularly Alu elements that provide the homology required to maintain stability via HDR. Additionally, we observed a striking difference between distributions of fixed and variable Alus across genome compartments. Through co-localization of fixed Alus and NAHR deletions in open chromatin we hypothesize that old Alu expansion in hominid lineage had a stabilizing role on the human genome.

STATISTICAL LEARNING WITH HIGH-DIMENSIONAL MASS CYTOMETRY DATA

Pratyaydipta Rudra[1], Elena Hsieh[2], Debashis Ghosh[2]

[1]Oklahoma State University, [2]University of Colorado Denver

Recent developments in single-cell based technologies, such as mass cytometry (CyTOF), has led to the need for computational and analytic approaches that can accommodate the high dimensionality and single-cell granularity. The analysis of CyTOF data can elucidate novel disease biomarkers and mechanisms of the underlying immunopathology, leading to improved treatments and prognostic measures. The use of single-cell technologies allows for consideration of expression from both a spatial and temporal framework. In spite of the promising nature of these platforms, much work remains in order to be able to meaningfully interpret the data in the context of biological questions. While end-to-end reproducible methods exist for fluorescence flow cytometry data analysis, they do not scale well for CyTOF data which have much higher dimensionality.

The data are often clustered into cell sub-populations first, which can then be used to answer scientific questions regarding the abundance of cell types and expressions of specific parameters (e.g. surface markers, signaling proteins, cytokines) across groups, such as disease and control groups, or stimulation regimes. The statistical questions about the tree-structured cell population data can be visualized in two layers. First, it is clinically interesting to know if the abundance of the cell subpopulations is different across two or more groups and/or conditions. Given the proportion of cell types for each sample, the next question is whether there is any differential expression of signaling proteins or cytokines (functional measurements of the cell populations studied).

Modeling data with multiple layers of correlation using a classical parametric model often becomes a challenging task. The classical parametric models also have limiting distributional assumptions such as normality, which may not be true for cytometry data. In order to tackle this, we developed a new statistical learning methodology based on the kernel distance covariance framework to compare the cell type composition different disease groups and stimulation conditions. High-dimensional statistical learning using a kernel machine regression is also developed to test the difference in cytokine expression levels across different cell-types and different conditions.

The methods are applied to high-dimensional dataset we collected containing different subgroups of populations including Systemic Lupus Erythematosus patients and healthy control subjects. The samples from the peripheral blood of the subjects were treated using three different stimulation methods. Preliminary analysis of the data revealed clinically relevant patterns such as differential cell type abundance between the disease and the control group, and also differential expression of several cytokines. For example, the expression of the cytokines MCP1, Mip1b and IL-1RA were found to be different among CD14 high monocytes across the two groups. An extensive simulation study to compare our statistical method with the existing approaches is currently being conducted.

HARDWARE ACCELERATION OF APPROXIMATE STRING MATCHING FOR BOTH SHORT AND LONG READ MAPPING

Damla Senol Cali[1], Lavanya Subramanian[2], Zülal Bingöl[3], **Jeremie S. Kim**[1,4], Rachata Ausavarungnirun[1], Anant V. Nori[2], Gurpreet S. Kalsi[2], Sreenivas Subramoney[2], Saugata Ghose[1], Can Alkan[3], Onur Mutlu[1,4]

[1]Carnegie Mellon University, [2]Intel Labs, [3]Bilkent University, [4]ETH Zurich

High throughput sequencing (HTS) technology enables fast and inexpensive generation of billions of DNA sequences (i.e., reads) from a genome [1, 2]. To quickly and accurately process the plethora of reads, we need new computational techniques. Analyzing HTS data requires finding the original locations of each read via an approximate string matching process against a long reference genome. Approximate string matching is typically performed with an expensive dynamic programming algorithm, which consumes over 90% of the first step's execution time. Many prior studies [3, 4] have identified this bottleneck in mapping and have proposed numerous methods for accelerating this expensive step on a wide-array of computational platforms.

Our goal in this work is to provide a fast and efficient implementation of approximate string matching towards enabling faster read mapping. We choose to accelerate Bitap [6, 7] due to its ability to perform approximate string matching with fast and simple bitwise operations, that can be highly parallelized for high throughput. We modified the algorithm to enable searching longer patterns and to remove the data dependency between the iterations and provide parallelism for the large amount of iterations. Unfortunately, in our study of Bitap on existing systems, we find that CPUs and GPUs alone are both limited by their respective architectures and thus cannot fully utilize the available hardware for maximal efficiency. Specifically, we find that the CPU implementation of Bitap is bottlenecked by computation since the working set fits within the L1 cache and the limited number of cores prevents the further parallel speedup. The GPU implementation of Bitap is bottlenecked by limited amount of private memory and destructive interference of threads while accessing the shared memory.

In order to overcome the imbalance in each of the above systems, we propose a custom accelerator for Bitap with characteristics that falls between the CPU and GPU. This achieves a finer balance in compute resources and memory for higher performance in approximate string matching. We also explore the design space of various accelerators, including processing in memory.

REFERENCES

[1] Alkan, Can, et al. "Limitations of Next-generation Genome Sequence Assembly," Nature Methods, 2011.
[2] Van Dijk, Erwin L., et al. "Ten Years of Next-generation Sequencing Technology," Trends in Genetics, 2014.
[3] Alser, Mohammed, et al. "GateKeeper: A New Hardware Architecture for Accelerating Pre-alignment in DNA Short Read Mapping," Bioinformatics, 2017.
[4] Kim, Jeremie S., et al. "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping using Processing-in-Memory Technologies," BMC Genomics, 2018.
[5] Baeza-Yates, Ricardo, et al. "A New Approach to Text Searching," Communications of the ACM, 1992.
[6] Wu, Sun, et al. "Fast Text Search Allowing Errors." Communications of the ACM, 1992.

TRANSITION OF REGULATORY FORCE TOWARD THE GENE EXPRESSIONS DURING OSTEOBLAST CELL DIFFERENTIATION

Yoichi Takenaka

Kansai University

Understanding the dynamics of cell differentiation system is one of the big issue in biology and medicine. It helps to acquire cells of diseased organs from pluripotent stem cells such as ES cell of iPS cell. To analyze the dynamics, the time-series gene expression profiles of cell lines from various organisms have been measured. It has been made clear that the movement of the gene expression. However, the dynamics of the system such as gene regulation mechanism are not well revealed yet.

In the poster, the author shows the dynamics of gene regulations during the osteoblast cell differentiation process from mesenchymal stem cell. There are many genes and gene regulations that are known to be active during the process. However, it has been not reported that the activity time of each regulation and the strength of the activated regulations. The author proposed a method to elucidate the transitions between the activation and inactivation of gene regulations at the temporal resolution of single time points. The method measures the strength of the gene regulations of each time point by leave one-time-point out way. Then it decomposes the time series of the gene expression data into partial series using information criterion. Finally, it determines whether each gene regulation of each partial time series is activated or in activated.

The gene expression profile of the osteoblast cell differentiation process includes 65 time points ranged from minus 6 hour to 192 hour where 0 hour is the time the cell differentiation process starts. The profile was downloaded from Genome Network Platform of National Institute of Genetics, Japan. The gene regulatory network that is activated at least one time point during the differentiation was composed from three reviewed papers. It includes 19 genes and 22 regulations where Runx2, the key transcription factor associated with osteoblast differentiation, is located at the center of the network. Osx, transcription factor Sp7, which serves as a maker for osteoblast differentiation, is at the downstream of Runx2.

The result shows there are four distinct periods during the osteoblast cell differentiation. And each period indicates when the expressions of genes are strongly controlled.

Before the cell differentiation process starts, Osx, BMP2, DLX5 and HDAC3 are the most strongly controlled among all the 65 time points. Next, EP300 is controlled strongly at the first period. Then, Creb, HDAC3, HDAC4, HDAC5 and Osx are. And at the final period, Runx2, Bglap, DLX5, DHAC7 and SMAD6 are. The analysis gives the hint to control the cell differentiation process.

METHYLATION PROFILES OF MELANOMA TO PREDICT TILs

Yihsuan Tsai[1], Nana Nikolaishvili Feinberg[1], Kathleen Conway[2], Sharon N. Edmiston[1], Nancy E. Thomas[3], Joel S. Parker[4]

[1]Lineberger Comprehensive Cancer Center (LCCC), University of North Carolina at Chapel Hill; [2]Department of Epidemiology, School of Public Health, Department of Dermatology, School of Medicine, Lineberger Comprehensive Cancer Center (LCCC), University of North Carolina at Chapel Hill; [3]Department of Dermatology, School of Medicine, Lineberger Comprehensive Cancer Center (LCCC), University of North Carolina at Chapel Hill; [4]Lineberger Comprehensive Cancer Center (LCCC), Department of Genetics, School of Medicine

Correlations between tumor infiltrating lymphocytes (TILs) and prolonged survival have been reported in many cancers including melanoma. However, current TIL assessment by pathologists reviewing the slide sections is not always ideal. Inter-observer agreement between pathologists may be low if the assessment was quantitative. To achieve a higher agreement, the estimates may be translated to categories. Here we proposed to train an epigenomics model to estimate the T-cell populations in melanoma samples using immunofluorescence (IF) image of CD3 and CD8 T-cells, which provides a more objective estimation of TILs.

In previous work, we generated methylation profiles for 89 melanoma and 78 nevi samples. To have a gold standard of TIL estimate, 80 out of the 89 melanoma samples were stained with IF to image CD3, CD8, S100 (melanoma marker) and a nuclear counterstain. We defined the fraction of CD3 and/or CD8 positive cells as T-cell fraction and found its estimate from the IF image has the most significant association with patient survival. Therefore, an elastic net model was built using features from the methylation dataset with T-cell fraction estimates from the IF image as response. Monte-Carlo cross validation was performed on 2/3 of the samples to tune the parameters. We identified 121 CpGs in the final model to estimate T-cell fraction which gave us the highest correlation with pearson r=0.87 in validation and r=0.91 in all samples. We also compared this method with two other methods. In a naÃ¯ve method, we identified CpGs with high methylation level in external lymphocyte samples and low in our nevi samples. These probes represent a lymphocyte methylation signature on an unmethylated nevi background. Therefore, we calculated the mode of kernel-smoothed DNA methylation distribution at these sites for each sample as a surrogate for lymphocyte fraction for that sample. This method gave a correlation relative to the gold standard of R= 0.64. Another method uses reference-based cell deconvolution algorithms, where a pre-built methylation reference was used to compute the fractions of each cell types via three different algorithms. While all three algorithms gave similar results, Robust Partial Correlations (RPC) provides the highest correlation with the gold standard (R= 0.58).

We then applied our final model (121 methylation markers) to an external dataset, TCGA-SKCM, to estimate the T-cell fractions. Since there is no gold standard for TCGA-SKCM dataset, we used survival as a surrogate. We found the T-cell fraction estimate from our model had a strong survival association (cox p-value = 3.85e-05). We will look at the correlation of our estimation with expression of T-cell gene modules next.

In summary, the predicted T-cell fraction from our methylation markers has very high correlation with the estimates from IF images and it's also highly correlated with patient survival.

HIGH-THROUGHPUT GENE TO KNOWLEDGE MAPPING THROUGH MASSIVE INTEGRATION OF PUBLIC SEQUENCING DATA

Brian Tsui, Hannah Carter

Department of Medicine, University of California, San Diego

Sequencing Read Archive contains more than one million runs of publicly available sequencing data. However, the lack of consistently preprocessed summary and molecular quantification data (for example, gene expression quantification for RNAseq) for each sequencing run hinders efficient Big Data interpolation. Here, we introduce Skymap, a standalone database that offers a single, multi-species data matrix incorporating all public sequencing studies. The data matrix contains several omic layers, including expression quantification, allelic read counts, microbes read counts, chip-seq. We reprocessed petabytes of sequencing data to generate the data matrix for each data type. We also offer a reprocessed biological metadata file that describes the relationships between the sequencing runs and the associated keywords, extracted from over 3 million freetext annotations using natural language processing. The processed data can fit into a single hard drive (<500GB). In https://github.com/brianyiktaktsui/Skymap, we showcase how one can (1) retrieve and analyze the SNPs and expression of a genetic variant across >250k runs in less than a minute and (2) increase the temporal resolution for tracking gene expression in mouse developmental hierarchy.

MANTA-RAE, PREDICTING THE IMPACT OF GENOME VARIANTS ON THE TRANSCRIPTION FACTOR BINDING POTENTIAL OF REGULATORY ELEMENTS

Robin van der Lee, Phillip A. Richmond, Oriol Fornes, Wyeth W. Wasserman

Centre for Molecular Medicine and Therapeutics - Department of Medical Genetics - BC Children's Hospital Research Institute - University of British Columbia - Vancouver, Canada

Interpreting the functional impact and pathogenicity of noncoding variants remains challenging. Increasing evidence suggests an important role for alterations that impact cis-regulatory elements and transcription factor (TF) binding sites (TFBSs). We are developing MANTA-RAE, a tool for Mutational ANalysis of Tfbs Alterations by Reconstruction of Altered regulatory Elements. MANTA-RAE will predict the effects of variants on TFBSs in regulatory elements in a three-step approach: (i) reconstructing reference and alternative genotypes based on user-supplied sets of genomic variants and regulatory elements, (ii) predicting TFBS through sequence scanning with curated TF binding models from JASPAR, and (iii) delta regulatory capacity analysis by comparing the TFBS potential of the reference and alternative sequences. MANTA-RAE will have the capacity to evaluate (i) both losses and gains of TFBSs and (ii) changes beyond single nucleotide variants, including small insertions, deletions, and larger copy number changes. Envisioned applications include prioritization of variants from rare disease and cancer genomes. These features should contribute to richer detection of regulation-altering noncoding variants that may contribute to disease.

USING QUANTITATIVE PHOSPHOPROTEOMICS TO UNDERSTAND FUNCTIONAL SELECTIVITY OF RECEPTOR TYROSINE KINASES

J. Watson, C. Francavilla, J.M. Schwartz

Faculty of Biology, Medicine and Health, University of Manchester

Cell signalling is the process of translating extracellular messages, or signals, to the inside of the cell in order to coordinate cellular activity. Cells receive signals from the external environment in a myriad of ways, including by the binding of extracellular proteins, called ligands, to receptors on the surface of the cell. Upon ligand binding, the signal is transmitted across the cell surface by the receptor and the signal propagates though the cell, primarily by the post-translational modification of proteins. For the receptor tyrosine kinase (RTK) family, this process is mediated by phosphorylation, a modification which is added to serine, threonine or tyrosine residues of proteins by the activity of kinases and removed by phosphatases. The addition of phosphoryl-groups is associated with activation of protein function. Ligand binding induces RTK dimerization and activation of kinase activity, allowing full activation of the receptor. This initiates a sequential cascade of protein phosphorylation, ultimately regulating transcription factor activity to modulate cellular behavior. An unanswered question in the field is how different ligands binding to the same receptor induce distinct signaling cascades, defined by changes in phosphorylation dynamics and consequent cellular behavior, a concept known as functional selectivity. This is demonstrated by fibroblast growth factor (FGF)-receptor 2b; when stimulated by either FGF7 or FGF10 an increase in proliferation or migration respectively is observed. Quantitative phosphoproteomics is a powerful method for comparing on a global scale the signaling cascades inducing these different behaviors. This comparison will allow us to define patterns of phosphorylation associated with signalling by different ligands, and use this to identify key phosphorylation sites associated with particular cell behaviors. We have developed a workflow to interrogate temporal phosphoproteomics datasets to directly compare the phosphorylation dynamics of cells stimulated by different ligands. As proteins may have multiple phosphorylation sites which can have independent effects and regulation, our approach considers data on the level of both the phosphorylation site and associated protein. Initial clustering of phosphorylated sites with similar dynamics over time is followed by protein-level analysis of functional similarity, using connectivity in graph databases, enrichment for ontological terms, and roles in well-studied signalling pathways (extracted from KEGG). Subsequent steps in the workflow aim to move the analysis from the protein to the phosphorylated sites. By integrating network-based analyses with phosphoproteomics data, we will develop novel methods for understanding and visualizing the role of phosphorylation in functional selectivity.

ANERIS APPLIED: SPARK-ENABLED ANALYTICS FOR FULL-SCALE AND REPRODUCIBLE ANNOTATION-BASED GENOMIC STUDIES

Nicholas Wheeler, Jeremy Fondran, Penny Benchek, Jonathan Haines, William S. Bush

Case Western Reserve University

Modern genomic studies are rapidly growing in scale, and the analytical approaches used to analyze genomic data are increasing in complexity. Genomic data management poses logistic and computational challenges, and analyses are increasingly reliant on genomic annotation resources that create their own data management and versioning issues. As a result, genomic datasets are increasingly handled in ways that limit the rigor and reproducibility of many analyses. In this work, we describe an analysis framework based on Spark infrastructure that provides management, rapid access, and flexible analysis of genomic data. By storing large-scale genomic and variant annotation resources alongside genomic data in a distributed system, we provide efficient methods for testing a variety of biologically-driven hypotheses for rare variants. Using the well-established Spark framework and analyses designed using Jupyter notebooks, we provide tools that improve processing speed, reduce user-driven data partitioning, and enhance the reproducibility of large-scale genomic studies.

PUTTING RELICANTHUS IN ITS PLACE: IMPACT OF MIXTURE MODEL CHOICE ON PHYLOGENETIC RECONSTRUCTION

Madelyne Xiao[1], Mercer R. Brugler[2], Estefania Rodriguez[1]

[1]Department of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024; [2]Biological Sciences Department NYC College of Technology (CUNY), 285 Jay Street, Brooklyn, NY 11201

First described in 2006, Relicanthus daphneae is a deep-sea anthozoan that lives on the ocean floor near hydrothermal vents in the East Pacific. It was originally classified as an anemone until a phylogenetic analysis in 2014 called this classification into question. The tree resulting from a maximum likelihood analysis for the Order Actiniaria (anemones) placed Relicanthus outside of Actiniaria; a recent analysis of Relicanthus' mitochondrial gene order, however, suggests its membership among the anemones. An ongoing study seeks to relate the choice of mixture model (e.g., maximum likelihood, maximum parsimony, Bayesian inference) to the resulting phylogenetic tree, taking into account the robustness of the data set in question (number of genes, specimens, etc). In particular, we are interested in the impact of mixture model choice on the placement of Relicanthus with respect to the actiniarians.

RATIONAL DESIGN OF NOVEL SKP2 INHIBITORS USING DEEP NEURAL NETWORKS

Shuxing Zhang, Beibei Huang, Lon W. Fong

Intelligent Molecular Discovery Laboratory, Department of Experimental Therapeutics, MD Anderson Cancer Center, Houston, TX 77054

Recently it has gained more and more attention with deep learning techniques, which show significant promise in generating predictive models for pharmaceutical research. In the present study, we attempt to develop deep neural networks method to design novel therapeutic agents for triple-negative breast cancer (TNBC) by targeting a crucial E3 ligase Skp2. TNBC represents about 20% of breast-cancer cases. It is highly aggressive with poor clinical outcome, and no targeted agents have been shown to be clinically effective in treating TNBC. Skp2 is an F-box protein, constituting one of the four subunits of the Skp1-Cullin-1 (Cul-1)-F-Box (SCF) ubiquitin E3 ligase complex. Earlier studies showed that Skp2 regulates cell cycle progression and proliferation by targeting ubiquitination and degradation of its substrates such as cell cycle inhibitor p27. Our in-house data also revealed that Skp2 was overexpressed in TNBC and correlated with poor prognosis. In addition, we revealed that genetic Skp2 inactivation also triggered a massive cellular senescence and/or apoptosis response in a p19Arf/p53-independent, but p27-dependent manner. Taken together, our results suggest that targeting Skp2 may represent a general "pro-senescence/apoptosis" and "anti-glycolysis" approach and is a promising therapeutic strategy for TNBC development and metastasis.

Herein we developed a novel deep neural network (DNN) method to predict TNBC cell responses to drugs based solely on their chemical features. In particular a cost function was employed to suppress overfitting. We also adopted an "early stopping" strategy to further reduce overfit and improve the accuracy of our models. Currently the software has been integrated with a genetic algorithm-based variable selection approach and implemented as part of our DL4DR package. We observed that DL4DR could handle big data set efficiently, significantly outperforming other methods in model-building and prediction and obtaining better results in big data analysis. When employed to predict drug responses of several highly aggressive TNBC cell lines, DL4DR produced robust and accurate predictions. Therefore, we applied these TNBC models to rationally design new small molecule inhibitors by targeting Skp2. After screening of millions of chemical compounds and designing novel structures based on our lead compound ZL25, we conducted a series of biochemical and cellular studies. These experimental examinations demonstrate that the top ranked molecules indeed inhibit Skp2 E3 ubiquitination functions significantly and kill TNBC cells effectively. Hence it has been used for our lead optimization of Skp2 inhibitors, and we anticipate that DL4DR can be employed as a general tool for hit identification and lead rational design for cancer therapeutics development.

PATTERN RECOGNITION IN BIOMEDICAL DATA: CHALLENGES IN PUTTING BIG DATA TO WORK

POSTER PRESENTATIONS

ODAL: A ONE-SHOT DISTRIBUTED ALGORITHM TO PERFORM LOGISTIC REGRESSIONS ON ELECTRONIC HEALTH RECORDS DATA FROM MULTIPLE CLINICAL SITES

Rui Duan, Mary Regina Boland, Jason H. Moore, **Yong Chen**

Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania

Electronic Health Records (EHR) contain extensive information on various health outcomes and risk factors, and therefore have been broadly used in healthcare research. Integrating EHR data from multiple clinical sites can accelerate knowledge discovery and risk prediction by providing a larger sample size in a more general population which potentially reduces clinical bias and improves estimation and prediction accuracy. To overcome the barrier of patient-level data sharing, distributed algorithms are developed to conduct statistical analyses across multiple sites through sharing only aggregated information. The current distributed algorithm often requires iterative information evaluation and transferring across sites, which can potentially lead to a high communication cost in practical settings. In this study, we propose a privacy-preserving and communication-efficient distributed algorithm for logistic regression without requiring iterative communications across sites. Our simulation study showed our algorithm reached comparative accuracy comparing to the oracle estimator where data are pooled together. We applied our algorithm to an EHR data from the University of Pennsylvania health system to evaluate the risks of fetal loss due to various medication exposures.

PLATYPUS: A MULTIPLE-VIEW LEARNING PREDICTIVE FRAMEWORK FOR CANCER DRUG SENSITIVITY PREDICTION

Kiley Graim[1], **Verena Friedl**[2], Kathleen E. Houlahan[3], Joshua M. Stuart[3]

[1]Flatiron Institute & Princeton University, [2]University of California Santa Cruz, [3]Ontario Institute of Cancer Research

Cancer is a complex collection of diseases that are to some degree unique to each patient. Precision oncology aims to identify the best drug treatment regime using molecular data on tumor samples. While omics-level data is becoming more widely available for tumor specimens, the datasets upon which computational learning methods can be trained vary in coverage from sample to sample and from data type to data type. Methods that can "connect the dots" to leverage more of the information provided by these studies could offer major advantages for maximizing predictive potential. We introduce a multi-view machine-learning strategy called PLATYPUS that builds "views" from multiple data sources that are all used as features for predicting patient outcomes. We show that a learning strategy that finds agreement across the views on unlabeled data increases the performance of the learning methods over any single view. We illustrate the power of the approach by deriving signatures for drug sensitivity in a large cancer cell line database. Code and additional information are available from the PLATYPUS website https://sysbiowiki.soe.ucsc.edu/platypus.

A SOFTWARE PIPELINE FOR DETERMINING FINE-SCALE TEMPORAL GENOME VARIATION PATTERNS IN EVOLVING POPULATIONS USING A NON-PARAMETRIC STATISTICAL TEST

Minjung Kwak[1], Seokwoo Kang[2], Dongwon Choo[2], Dohyeon Lee[2], Jinhee Lee[2], Seonghyeon Kim[2], **Giltae Song**[2]

[1]Yeungnam University, [2]Pusan National Univeristy

Abnormal variations are frequent in clonal genome evolution of cancers. Such aberrational variations often function as a driver in cancer cell growth. Understanding fundamental evolutionary dynamics underlying these variations in tumor metastasis still is understudied owing to their genetic complexity.

Recently, whole genome sequencing empowers to determine genome variations in short-term evolution of cell populations. This approach has been applied to evolving populations of model organisms including yeast. It is substantial progress in evolutionary genomics to examine sequence changes at such fine-scale resolution. However, existing statistical tests for analyzing variation temporal changes in multiple time-points are limited to identify the full spectrum of intermediate changes.

We designed a new statistical approach based on Kolmogorov-Smirnov test and integrated it into a software tool for determining the variation patterns in fine-scale temporal resolution in experimental evolution studies. We validated our method using simulation data that mimic the evolution of fruit fly populations. We compared the results of ours and other existing methods such as the Cochran-Mantel-Haenszel (CMH) test and the beta-binomial Gaussian process (BBGP) method. We analyzed yeast (Saccharomyces cerevisiae) W303 strain genomes from 40 populations at 12 time-points using our software pipeline. Our toolset can be also applied for identifying abnormal variation changes in other evolving populations.

A DEEP LEARNING APPROACH TO IDENTIFYING THE CELLULAR COMPOSITION OF SOLID TISSUE WITH DNA METHYLATION DATA

Meghan E. Muse[1], Curtis L. Petersen[1], Carmen J. Marsit[2], Diane Gilbert-Diamond[1], Brock C. Christensen[1]

[1]Dartmouth College, [2]Emory University

DNA methylation is involved in the establishment of cellular identity and measured profiles of DNA methylation can be leveraged to deconvolute the underlying cellular composition of a tissue sample. Currently, both reference-based and reference-free methods exist to estimate the relative proportion of inferred cell types in solid tissue using DNA methylation data. However, establishing DNA methylation libraries for reference-based deconvolution in solid tissues is challenging and use of reference-free approaches to estimate putative cell type proportions are computationally intensive, particularly as sample size increases. As observed patterns in DNA methylation can be most strongly explained by the relative proportion of cell types in a tissue sample, we investigated the utility of implementing an unsupervised variational autoencoder (VAE) approach to learn a defined number of latent dimensions in DNA methylation data and tested their relationship with inferred cell type proportions from a reference-free approach. We implement the Tybalt model developed by Way et al. to learn latent representations of DNA methylation data measured on the Illumina 450K array in 334 placental samples. We compare the results of this method to those from a well-established reference free method for inferring the relative proportions of putative cell types. We considered models that learned 10 to 100 latent dimensions and selected the model in which the greatest number of putative cell types identified by the reference free method had moderate correlation ($r2 > 0.5$) with at least one latent dimension. This resulted in the selection of a model learning 10 latent dimensions. In this model, learned latent dimensions had moderate correlation with 5 of the 9 putative placental cell types identified by the reference free method and strong correlation ($r2 > 0.7$) with 2 putative placental cell types. To better understand the underlying biology represented by these latent dimensions, we assess the CpG loci most strongly correlated with the activations of these 5 latent dimensions as a means of identifying genes that are representative of cellular identity.

DIRECTLY MEASURING THE RATE AND DYNAMICS HUMAN MUTATION BY SEQUENCING LARGE, MULTI-GENERATIONAL PEDIGREES

Thomas A. Sasani, Brent S. Pedersen, Mark Leppert, Ray White, Lisa Baird, **Aaron R. Quinlan**, Lynn B. Jorde

Department of Human Genetics, University of Utah

Developing an accurate estimate of the human germline mutation rate is critical to our understanding of evolution, demography, and genetic disease. Early phylogenetic analyses inferred mutation rates from the observed sequence divergence between humans and related primate species at particular genes and pseudogenes. However, as whole genome sequencing has become ubiquitous, these estimates have been refined using pedigree-based approaches. By identifying mutations present in offspring that are absent from their parents (de novo mutations), it is possible to more accurately approximate the human germline mutation rate.

To obtain a precise, unbiased estimate of the mutation rate in humans, we performed deep whole-genome sequencing on blood-derived DNA from 34 of the original three-generation CEPH families from Utah, comprising a total of 604 individuals. These families, which each contain grandparents (P0 generation), parents (F1), and their children (F2), are considerably larger than any used in prior estimates of the human mutation rate, and offer unique power to detect and validate de novo mutation. With a median of 8 F2 individuals per pedigree, we were able to biologically validate putative de novo mutations in the F1 generation by assessing their transmission to a third generation. Using this dataset, we have generated a high-confidence estimate of the human mutation rate (1.31 x 10-8 / bp / generation), observe a significant parental age effect on the rate of de novo mutation, and identify wide variability in family-specific age effects across CEPH pedigrees. To our knowledge, this study represents the first example of a longitudinal analysis of the effect of parental age within individual families. Additionally, we have identified recurrent de novo variants present in multiple F2 offspring, which are likely the result of mosaicism in the parental germline.

Finally, we have trained a classification model on the high-quality, transmitted de novo variants in our dataset, and used this model to identify de novo mutations in a large cohort of children from the Simons Foundation for Autism Research Initiative (SFARI). Combining the de novo mutations observed in 34 Utah families with the SFARI callset, we have generated a dense genomic map of spontaneous human mutation. We observe regional enrichment of de novo variation in the human genome, and explore the role of sequence context, as well as molecular processes like recombination and gene conversion, on the rate of human mutation.

AVAILABLE PROTEIN 3D STRUCTURES DO NOT REFLECT HUMAN GENETIC AND FUNCTIONAL DIVERSITY

Gregory Sliwoski, Neel Patel, R. Michael Sivley, Charles R. Sanders, Jens Meiler, **William S. Bush**, John A. Capra

Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA, Center for Structural Biology, Vanderbilt University, Nashville, TN, USA; Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA; Department of Biochemistry, Vanderbilt University, Nashville, TN, USA; Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; Department of Chemistry, Vanderbilt University, Nashville, TN, USA; Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA; Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA; Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Genomic databases and clinical trials are substantially biased towards European ancestry populations, and this bias significantly contributes to health disparities. Structural biology has an essential role in investigating protein function and clinical variant interpretation, providing powerful tools for investigating the impact of genetic variants on protein structure and function. However, studies that analyze the 3D structure of proteins typically consider a single canonical amino acid sequence as representative of the protein.

Here, we evaluate the potential for this simplification to bias results toward different populations by evaluating how well 66,971 experimentally characterized human protein 3D structures represent the sequence diversity of the proteins they model. Thousands of protein structures have unrepresented alternative sequences commonly found in human populations, and African ancestry individuals' sequences are the least likely to be represented by available structures. Because sequence variability is often limited to a few positions within a protein, we evaluate the likelihood of these small changes to influence protein function. Combining existing annotations and computational modeling, we identify thousands of proteins for which use of a single structure as representative of "wild type" may bias results against certain populations or individuals. Variants segregating in human populations, but unrepresented in structures, are observed across functional sites involved in stability (134 disulfide bond cysteines), regulation (94 phosphorylation sites), DNA binding (322 residues), small molecule binding (1,463 residues with 362 within drug binding sites), and protein-protein interfaces (6,144 residues).

We computationally model more than 700 unrepresented variants' effects on protein stability and protein-protein interaction. Changes in predicted protein stability are found for 28% (156) of the 556 variants, with stabilizing (41) and destabilizing (115) effects predicted. Of 161 protein-interface variants modeled, 25% (41) are predicted to impact protein-protein binding. These variants in human populations have potential to impact the study of their protein's structure and function.

With the widespread use of protein structures in basic science and clinical variant interpretation, human protein sequence and structural diversity must be considered to enable accurate and reproducible conclusions from structural analyses.

SEMANTIC WORKFLOWS FOR BENCHMARK CHALLENGES: ENHANCING COMPARABILITY, REUSABILITY AND REPRODUCIBILITY

Arunima Srivastava[1], Ravali Adusumilli[2], Hunter Boyce[2], Daniel Garijo[3], Varun Ratnakar[3], Rajiv Mayani[3], Thomas Yu[4], Raghu Machiraju[1], Yolanda Gil[3], Parag Mallick[2]

[1]The Ohio State University, [2]Stanford University, [3]University of Southern California, [4]Sage Bionetworks

Benchmark challenges, such as the Critical Assessment of Structure Prediction (CASP) and Dialogue for Reverse Engineering Assessments and Methods (DREAM) have been instrumental in driving the development of bioinformatics methods. Typically, challenges are posted, and then competitors perform a prediction based upon blinded test data. Challengers then submit their answers to a central server where they are scored. Recent efforts to automate these challenges have been enabled by systems in which challengers submit Docker containers, a unit of software that packages up code and all of its dependencies, to be run on the cloud. Despite their incredible value for providing an unbiased test-bed for the bioinformatics community, there remain opportunities to further enhance the potential impact of benchmark challenges. Specifically, current approaches only evaluate end-to-end performance; it is nearly impossible to directly compare methodologies or parameters. Furthermore, the scientific community cannot easily reuse challengers' approaches, due to lack of specifics, ambiguity in tools and parameters as well as problems in sharing and maintenance. Lastly, the intuition behind why particular steps are used is not captured, as the proposed workflows are not explicitly defined, making it cumbersome to understand the flow and utilization of data. Here we introduce an approach to overcome these limitations based upon the WINGS semantic workflow system. Specifically, WINGS enables researchers to submit complete semantic workflows as challenge submissions. By submitting entries as workflows, it then becomes possible to compare not just the results and performance of a challenger, but also the methodology employed. This is particularly important when dozens of challenge entries may use nearly identical tools, but with only subtle changes in parameters (and radical differences in results). WINGS uses a component driven workflow design and offers intelligent parameter and data selection by reasoning about data characteristics. This proves to be especially critical in bioinformatics workflows where using default or incorrect parameter values is prone to drastically altering results. Different challenge entries may be readily compared through the use of abstract workflows, which also facilitate reuse. WINGS is housed on a cloud based setup, which stores data, dependencies and workflows for easy sharing and utility. It also has the ability to scale workflow executions using distributed computing through the Pegasus workflow execution system. We demonstrate the application of this architecture to the DREAM proteogenomic challenge.

PRECISION MEDICINE: IMPROVING HEALTH THROUGH HIGH-RESOLUTION ANALYSIS OF PERSONAL DATA

POSTER PRESENTATIONS

CLASS PRIOR ESTIMATION AND QUANTIFICATION OF THE LOSS AND GAIN OF RESIDUE FUNCTION UPON MUTATION

Shantanu Jain[1], Jose Lugo-Martinez[2], Martha White[3], Michael W. Trosset[4], Predrag Radivojac[1]

[1]Northeastern University, [2]Carnegie-Mellon University, [3]University of Alberta, [4]Indiana University

Standard algorithms for binary classification assume access to labeled data from both the positive and the negative class. However, in many biological problems, labeled examples from one of the classes (say, negatives) is not available. In this scenario, a positive-unlabeled learner, that relies on positive and unlabeled examples only, is used. Surprisingly, this strategy leads to an optimal score function. However, picking an optimal threshold to construct the final classifier requires the knowledge of the class priors, the proportion of positives and negatives in the unlabeled data. I will 1) present an nonparametric algorithm for estimation of the class priors based on a mixture model formulation, 2) elucidate the assumptions necessary for the algorithm, and 3) derive a class prior preserving univariate transform for dimensionality reduction and thereby obtain a practical algorithm for multivariate data. Moreover, I will also demonstrate how the posterior can be estimated using the estimate of the class priors. I will further extend these results to a more general setting where some of the examples labeled as positive are in fact negative. I will present experimental results demonstrating the efficacy of our algorithm, comparing it with the state of the art methods and other baseline methods on many real and synthetic datasets. Lastly, I will present a biological application of this work to establish the loss and gain of residue function as a common mechanism for inherited diseases.

PREDICTION OF TIME TO INSULIN USING CLINICAL AND GENETIC BIOMARKERS IN TYPE 2 DIABETES PATIENTS

Rikke Linnemann Nielsen[1], Louise Donnelly[2], Agnes Martine Nielsen[3], Konstantinos Tsirigos[1], Kaixin Zhou[2], Bjarne Ersboell[3], Line Clemmensen[3], Ewan Pearson[2], Ramneek Gupta[1]

[1]Department of Bio and Health Informatics, Technical University of Denmark; [2]Medical Research Institute, University of Dundee, United Kingdom; [3]Department of Applied Mathematics and Computer Science, Technical University of Denmark

Type II diabetes (T2D) is a complex metabolic disorder where the risk of a fast or slow disease progression is highly dependent of each individual. Therefore, it is useful to identify predictive biomarkers for diabetes progression and relevant patient subgroups characteristics that may assist clinical decisions in T2D treatment management.

In this study, we obtained electronic medical records from a cohort-based population in Tayside, UK registered from December 1994 to September 2015. Using life-style data, anthropometry, biochemical data, drug-prescription data and genetic features from electronic medical records on 6871 T2D patients, artificial neural network models (ANN) were trained with two-layer cross-validation to classify T2D patients' progression given as patients' time to insulin (TTI). TTI was defined as the first day of insulin treatment or as the clinical need for insulin (HbA1c >8.5% treated with two or more non-insulin diabetes therapies). Prediction targets were TTI within year 1, 3 or 5 from the time of diagnosis. Genetic variants were selected by prior knowledge on T2D and glycemic trait predisposition SNPs from ~80M imputed SNPs. Prediction models were investigated for understanding which biomarkers were most predictive of progression.

ANNs with all data except genetic variants predicted TTI for year 1 (0.92Â±0.02, 0.83Â±0.04, 0.86Â±0.04 for AUC, sensitivity and specificity, respectively), year 3 (0.82Â±0.03, 0.71Â±0.05, 0.78Â±0.04) and year 5 (0.78Â±0.02, 0.66Â±0.02, 0.76Â±0.02). Most important features included HbA1c, GAD antibody concentration and the type of diabetes therapy patients were receiving at the time of confirmed diagnosis. Integration of genetic variants, using a forward selection strategy, resulted in a slightly improved performance in all three models; year 1 (0.94 Â± 0.01, 0.83 Â± 0.03, 0.90 Â± 0.01), year 3 (0.85 Â± 0.02, 0.72 Â± 0.05, 0.80 Â± 0.02), and year 5 (0.80Â±0.03, 0.68Â±0.04, 0.78Â±0.02).

We are currently examining the robustness of the selected SNPs by building an ensemble of multiple models with different features and investigating if the genetic features are relevant to specific patient subgroups, as well as carrying out further longitudinal work with the phenotype to include more information about a given patient using longitudinal patient information across irregular sampled time points.

PATHOGENICITY AND FUNCTIONAL IMPACT OF INSERTION/DELETION AND STOP GAIN VARIATION IN THE HUMAN GENOME

Kymberleigh A. Pagel[1], Danny Antaki[2], Matthew Mort[3], David N. Cooper[3], Jonathan Sebat[2], Lilia M. Iakoucheva[2], Sean D. Mooney[4], **Predrag Radivojac**[5]

[1]Indiana University, [2]University of California San Diego, [3]Cardiff University, [4]University of Washington, [5]Northeastern University

An individual human exome may contain hundreds of protein-coding insertion/deletions (indels) and dozens of protein truncating variants. Accurate differentiation between phenotypically neutral and disease-causing genetic variation remains an open problem, particularly among the excess of indel variants brought about by recent developments in sequencing technologies. Indel and protein truncating variants exhibit diverse impact on protein sequence, from a single residue to the deletion of entire functional domains. We present machine learning methods to predict the pathogenicity and the types of functional residues impacted by loss-of-function and indel variation. The models show good predictive performance and the potential to identify effect upon resides predicted to effect structural and functional features, including secondary structure, intrinsic disorder, metal and macromolecular binding, post-translational modifications, and catalytic residues. We identify structural and functional mechanisms that are impacted preferentially by germline variation from the Human Gene Mutation Database, recurrent somatic variation in COSMIC, and de novo variation from individuals with neurodevelopmental disorders. Collectively, the pathogenicity prediction and predicted functional effects provide a framework to facilitate the interrogation of indel and protein truncating variants.

DETECTING POTENTIAL PLEIOTROPY ACROSS CARDIOVASCULAR AND NEUROLOGICAL DISEASES USING UNIVARIATE, BIVARIATE, AND MULTIVARIATE METHODS ON 43,870 INDIVIDUALS FROM THE EMERGE NETWORK

Xinyuan Zhang[1], Yogasudha Veturi[1], Shefali S. Verma[1], William Bone[1], Anurag Verma[1], Anastasia M. Lucas[1], Scott Hebbring[2], Joshua C. Denny[3], Ian Stanaway[4], Gail P. Jarvik[4], David Crosslin[4], Eric B. Larson[5], Laura Rasmussen-Torvik[6], Sarah A. Pendergrass[7], Jordan W. Smoller[8], Hakon Hakonarson[9], Patrick Sleiman[9], Chunhua Weng[10], David Fasel[10], Wei-Qi Wei[3], Iftikhar Kullo[11], Daniel Schaid[11], Wendy K. Chung[10], Marylyn D. Ritchie[1]

[1]University of Pennsylvania, [2]Marshfield Clinic, [3]Vanderbilt University, [4]University of Washington, [5]Kaiser Permanente Washington Health Research Institute, [6]Northwestern University, [7]Geisinger Health System, [8]Massachusetts General Hospital, [9]Children's Hospital of Philadelphia, [10]Columbia University, [11]Mayo Clinic

The link between cardiovascular diseases and neurological disorders has been widely observed in the aging population. Disease prevention and treatment rely on understanding the potential genetic nexus of multiple diseases in these categories. In this study, we were interested in detecting pleiotropy, or the phenomenon in which a genetic variant influences more than one phenotype. Marker-phenotype association approaches can be grouped into univariate, bivariate, and multivariate categories based on the number of phenotypes considered at one time. Here we applied one statistical method per category followed by an eQTL colocalization analysis to identify potential pleiotropic variants that contribute to the link between cardiovascular and neurological diseases. We performed our analyses on ~530,000 common SNPs coupled with 65 electronic health record (EHR)-based phenotypes in 43,870 unrelated European adults from the Electronic Medical Records and Genomics (eMERGE) network. There were 31 variants identified by all three methods that showed significant associations across late onset cardiac- and neurologic- diseases. We further investigated functional implications of gene expression on the detected "lead SNPs" via colocalization analysis, providing a deeper understanding of the discovered associations. In summary, we present the framework and landscape for detecting potential pleiotropy using univariate, bivariate, multivariate, and colocalization methods. Further exploration of these potentially pleiotropic genetic variants will work toward understanding disease causing mechanisms across cardiovascular and neurological diseases and may assist in considering disease prevention as well as drug repositioning in future research.

PHARMGKB: THE API AND INFOBUTTONS

Michelle Whirl-Carrillo[1], Ryan M. Whaley[1], Mark Woon[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. PharmGKB knowledge is defined by a data model, stored in a database, and accessed through the Application Programming Interface (API). The API supplies data to the www.pharmgkb.org website which is the most common way for people to query and view the knowledge content of PharmGKB.

Additionally, the PharmGKB API supports the InfoButton specification which is used on the ClinGen website as well as by others in their EHR systems. The Infobutton Implementation Guide provides a standard mechanism for EHR systems to submit knowledge requests to knowledge resources over the HTTP protocol for point-of-care decision support. PharmGKB provides this as part of its standard API, using RXCUIs (RxNorm concept unique identifiers) and normalization of drug names, and returns HTML, with plans to support JSON and XML (https://api.pharmgkb.org/infobutton.html).

For InfoButtons, the EHR displays a button for the user to click that will query the PharmGKB and display information directly in the EHR application. Inside the application, a list of drug identifiers (RXCUIs) are created and then submitted to the InfoButton service's URL. The URL then returns a report in HTML that is displayed to the EHR user directly in the interface. The PharmGKB InfoButton implementation displays dosing guideline annotations, drug label annotations, and top-level clinical annotations that are relevant to the drug identifiers provided by the user. We monitor the API request logs to assess usage.

SINGLE CELL ANALYSIS – WHAT IS IN THE FUTURE?

POSTER PRESENTATIONS

INTRA TUMOR HETEROGENEITY (ITH) METRIC OF CIRCULATING TUMOR CELL (CTC)-DERIVED XENOGRAFT MODELS IN SMALL CELL LUNG CANCER.

Yuanxin Xi[1], C. Allison Stewart[2], Carl M. Gay[2], Hai Tran[2], Bonnie Glisson[2], John V. Heymach[2], Paul Robson[3], Lauren A. Byers[2], Jing Wang[1]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [2]Department of Thoracic/Head & Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [3]*The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA*

Small cell lung cancer (SCLC) is an aggressive malignancy characterized by rapid onset of platinum-resistance. Once considered a homogeneous disease, recent analyses of SCLC have shown intra-tumoral heterogeneity (ITH) associated with treatment-resistance. To further investigate the contribution of intra-tumoral heterogeneity (ITH) to clinical outcomes in SCLC, we profiled single-cell RNAseq expression of circulating tumor cell (CTC)-derived xenograft (CDX) models from SCLC patients that recapitulate patient tumor genomics and response to platinum chemotherapy.

Characterizing the heterogeneity of tumor cell subpopulations remains a bioinformatics challenge in analyzing single-cell RNAseq data for CTC-derived CDX models, mostly due to lack of an accurate method to quantify the complexity of tumor cell expression patterns at single cell resolution and discover the correlations with different tumor development or treatment response mechanisms. In this study, we developed a variance-based metric to measure the overall heterogeneity of tumor cell populations based on single cell RNAseq expression profiles We applied this metric to the Chromium 10x single cell RNAseq data of 4 SCLC CDX models that has different platinum treatment responses, and identified a global increase of intra-tumor heterogeneity in platinum-resistant models compared with platinum-sensitive models, and defined variable gene expression as a reliable hallmark of increasing therapeutic resistance in SCLC. Further gene set enrichment analysis (GSEA) of the treatment naÃ¯ve and relapsed samples revealed that the increased ITH metric were associated with multiple concurrent resistance mechanisms , suggesting that resistance to molecularly targeted therapies does not follow a predictable, reproducible pathway within the same CDX model. These results showed that the variance-based ITH metric successfully characterized the resistance associated heterogeneity increases in SCLC tumor cells, and more broadly, it provides a general purpose quantitative measurement of the tumor cell subpopulation heterogeneity in single cell analysis.

WHEN BIOLOGY GETS PERSONAL: HIDDEN CHALLENGES OF PRIVACY AND ETHICS IN BIOLOGICAL BIG DATA

POSTER PRESENTATIONS

QUANTIFYING THE IDENTIFIABILITY OF INDIVIDUALS USING A SPARSE SET OF SNPs

Prashant S. Emani, Gamze Gursoy, Mark B. Gerstein

Department of Molecular Biophysics and Biochemistry, Yale University

The recent revolution in high-throughput genomics has led to the proliferation of publicly available datasets and databases enabling queries on individual genotypes, whether in the form of reference genotypes, single nucleotide polymorphism (SNP) "beacons" or functional genomics data with significant identifying-information leakage. It is therefore of interest to quantify the power of a sparse set of SNPs to reveal the identity of an individual, as this would help determine the privacy risks of making particular datasets accessible to the research community. Such an evaluation would enable a principled cost-benefit analysis to determine the right balance of public and private data accessibility. We present a tool for such quantification based on well-established Hidden Markov Models (HMMs) of chromosomal recombination (Li and Stephens, 2003): the central idea is to explore the state space of reference haplotypes from a database, and find the trajectory through this space that best describes observed genotypes. The tool enables simple SNP-based kinship analysis by the identification of queried individuals as a "mosaic", or piecewise combination, of the input reference haplotypes, while allowing for genotyping error and de novo mutation. The output includes the best-fit reference haplotype trajectories, which for a small set of input SNPs, could result in several equal-probability possibilities. However, even in this case, inferences could be made on the membership of an individual in certain haplotype communities based on their enrichment within the best-fit trajectories. This approach parallels linkage disequilibrium- (LD-) based methods, but avoids any assumptions of population homogeneity as it does not require explicit calculation of allele frequencies or SNP correlations. It is, of course, dependent on the availability of a sufficiently rich database to ensure that the queried individual is at least related. This limitation is fast becoming a non-issue, however, with the constant expansion of population-level genotype databases. The results of representative simulations using the 1000 Genomes reference dataset with randomly chosen common SNPs (allele frequency > 0.05) from a single chromosome are: searching for a genotyped individual among 100 phased genotypes (= 200 reference haplotypes) yielded accurate discovery with as few as 12 SNPs; including a mutation rate of 0.1 â€“ 0.2 increased the number of SNPs required for reliable identification to ~25; simulations of mosaic samples composed of two reference individuals, each contributing half of the SNPs, suggested that ~30 SNPs could be sufficient to identify the two constituent individuals. These numbers would likely be improved upon when all chromosomes are combined. In summary, we provide a tool that can serve to identify observed genotypes either known to be members of a database, or related to individuals within the database, under varying conditions of mutation and recombination rates with no assumptions about the population-specific allele frequencies of SNPs.

TRANSCRIPTOMIC SUMMARY SPLICING DATA MAY LEAK PERSONAL PRIVATE INFORMATION BY COMPUTATIONAL LINKAGE TO THE GENOMIC VARIANTS

Zhiqiang Hu[1], Mark B. Gerstein[2], **Steven E. Brenner**[1]

[1]University of California, Berkeley, [2]Yale University

Sharing genomes without personal identifiers has been common practice in biological and medical research. However, recent studies revealed the risk of re-identifying people from their genomes, or attached quasi-identifiers, such as sex, birthdate, and zip code. Moreover, consumer databases now contain genetic data for millions of individuals; a recent study suggested that most Americans have detectable family relationships in these databases, allowing their identification using demographic identifiers. The additional availability of an individual's RNA-seq data has implications for privacy, as it may be linked to the genome, potentially allowing the person's privacy to be breached. For example, sex and ethnicity information may be inferred directly from a genome, and the study may provide a zip code. This genome could be linked to RNA-seq data from a diabetes study with attached birthdates and income. These combined quasi-identifiers may uniquely identify the person, and the study reveals the person's diabetes disease status. NEW PARAGRAPH RNA-seq reads contain genetic variants, and thus can be directly linked to the genome. To avoid this risk, some researchers now release gene expression, isoform expression and exon read count data instead of the raw sequencing reads. NEW PARAGRAPH However, gene expression can also be linked to the genome based on expression QTLs (eQTLs). Using a Bayesian framework, we found that it is feasible to predict genomic variants from summarized splicing data. Based on GTEx splicing QTLs (sQTL) data, using relative isoform expression from 15 genes, we could identify the target genome within a pool containing hundreds of individuals with >90% accuracy. We could also link RNA-seq data from a certain tissue or cell type to the genome using parameters trained from a similar tissue, indicating parameters trained on major tissues may enable the linkage of RNA-seq from all types of human samples to the genome. By quantitatively measuring the information leakage from each sQTL, we found that it is possible to identify the target genome of an RNA-seq dataset from millions of individuals using more sQTLs. Researchers have proposed to eliminate the risk of eQTL-based linking attacks by adding noise to the gene expressions, based on the observation that only a few genes enable linkage. However, our framework suggested that there are now many more such genes than previously reported. We find that expression data enables the re-identification of target genome from a pool containing billions of genomes. Our result implies that mitigation of the linking risk by adding noise would severely abrogate biological entity of the data, since the data will no longer be biologically meaningful when over half of gene expressions are modified. Our study also implies that other kinds of "omic" data, including DNA modification and protein metabolite levels, may also leak genome privacy.

WORKSHOP: MERGING HETEROGENEOUS DATA TO ENABLE KNOWLEDGE DISCOVERY

POSTER PRESENTATION

TO SEARCH A HETNET... HOW ARE TWO NODES CONNECTED?

Daniel Himmelstein[1], Michael Zietz[1], Kyle Kloster[2], Michael Nagle[3], Blair Sullivan[2], Casey S. Greene[1]

[1]University of Pennsylvania, [2]North Carolina State University, [3]Pfizer Inc.

Networks with multiple node and relationship types, called hetnets, provide an ideal data structure to integrate biomedical knowledge. One example, Hetionet, has 47 thousand nodes of 11 types and 2.25 million relationships of 24 types covering diseases, small molecule drugs, and the entities in between, which range from molecular (e.g. genes & pathways) to organismal (e.g. side effects & symptoms).

We are building a search engine for hetnet connectivity on the Hetionet network. We want to provide users with an immediate answer to the question, "how are these two nodes connected?" We approach this problem by identifying types of paths where a source and target node are connected more than expected by chance (i.e. based on their degrees alone).

While still a work in progress on GitHub (https://github.com/greenelab/hetmech), the project is nearing a prototype web application. Reaching this stage required several methodological advances. First, we implemented efficient path counting algorithms in Python based on matrix multiplication. A new HetMat data structure provides efficient on-disk storage of hetnets, optimized for matrix operations and caching. We designed a novel gamma-hurdle method for assessing the null distribution of a degree-weighted path count (DWPC) for a given pair of source-target node degrees.

Using these techniques, we computed measures of connectivity between all node-pairs for the 2,205 types of paths (metapaths) with length ≤ 3 in Hetionet v1.0 (available at https://doi.org/cww7). Now, we aim to expose the hidden information these measures capture: namely, how are two nodes related in terms of metapaths, individual paths, and intermediate nodes. Stop by our poster to learn more and discuss how this search engine can help you peruse biomedical knowledge or interpret your computational predictions.

WORKSHOP: TEXT MINING AND MACHINE LEARNING FOR PRECISION MEDICINE

POSTER PRESENTATION

LITVAR: MINING GENOMIC VARIANTS FROM BIOMEDICAL LITERATURE FOR DATABASE CURATION AND PRECISION MEDICINE

Alexis Allot, Yifan Peng, Chih-Hsuan Wei, Kyubum Lee, Lon Phan, **Zhiyong Lu**

National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894

The identification and interpretation of genomic variants play a key role in the diagnosis of genetic diseases and related research in the era of precision medicine. To stay up to date, researchers must process an ever-increasing amount of new publications. This task is complicated by two factors. First, authors use multiple abbreviations to refer to the same variant. For example, "A146T", "c.436G>A", and Ala146Thr all refer to the same variant rs121913527. Second, the same abbreviation (e.g., p.Ala94Thr) can refer to different variants in different genes. A simple search on PubMed would thus return only a subset of all relevant articles for the variant of interest, while returning many articles that are irrelevant.

To help scientists, healthcare professionals, and database curators find the most up-to-date published variant research, we have developed LitVar, a novel webserver for linking genomic variant data in the literature with intuitive UI (1). Specifically, it employs a suite of state-of-the-art entity recognition tools as its backend processing method. LitVar combines robust and advanced text mining with data integrations from PubMed (>28 million abstracts) and PubMed Central Subset (>2.7 million full-length articles) to improve both sensitivity and specificity. As of May 2018, there are more than 2 million unique variants in our system, associated with hundreds of thousands of publications from PubMed and PMC Open Access Subset. While comparing with PubMed, LitVar achieved an increase in sensitivity and specificity. For example, with a search of "rs113488022", no results can be found in PubMed, but over 6,000 articles are returned by LitVar. On the other hand, a search for "H199R" on PubMed will return articles where this variant presents both on the gene LIN28B (PMID:22964795) and CFTR (PMID:15084222), while the disambiguation process of LitVar will allow the user to select precisely the variant (and gene) of interest.

To further assist users, LitVar allows matching publications to be filtered by journal, type, date or part of publication. Moreover, publications' popularity in time can be visualised as a zoomable histogram. In addition to the website, LitVar provides REST APIs to allow users to disambiguate a textual query into a list of top matching variants, or perform large-scale analysis, by retrieving publications linked to hundreds of supplied dbSNP identifiers in one query.

LitVar is now integrated in dbSNP. The newly added link allows users not only to view more publications than with the link to PubMed, but also to assess the context (sentence and related diseases, chemicals and other variants) in which the variant appears in each publication.

LitVar is publicly available at https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar.

[1] Allot, A., Peng, Y., Wei, C.H., Lee, K., Phan, L. and Lu, Z. (2018) LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res.

AUTHOR INDEX