PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020

ABSTRACT BOOK

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

Proceedings papers with oral presentations #2-39 are not assigned poster space.

Abstracts are organized first by session, then the last name of the first author. Presenting authors' names are underlined in the Table of Contents and in **bold** text on the abstracts.

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

POSTER PRESENTATIONS

WORKSHOPS WITH POSTER PRESENTATIONS

ATRIFICIAL INTELLIGENCE FOR ENHANCING CLINICAL MEDICINE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

Predicting Longitudinal Outcomes of Alzheimer's Disease via a Tensor-Based Joint Classification and Regression Model

Lodewijk Brand[1], Kai Nichols[1], **Hua Wang**[1], Heng Huang[2], Li Shen[3], for the ADNI

[1]Colorado School of Mines, [2]University of Pittsburgh, [3]University of Pennsylvania

Alzheimer's disease (AD) is a serious neurodegenerative condition that affects millions of people across the world. Recently machine learning models have been used to predict the progression of AD, although they frequently do not take advantage of the longitudinal and structural components associated with multi-modal medical data. To address this, we present a new algorithm that uses the multi-block alternating direction method of multipliers to optimize a novel objective that combines multi-modal longitudinal clinical data of various modalities to simultaneously predict the cognitive scores and diagnoses of the participants in the Alzheimer's Disease Neuroimaging Initiative cohort. Our new model is designed to leverage the structure associated with clinical data that is not incorporated into standard machine learning optimization algorithms. This new approach shows state-of-the-art predictive performance and validates a collection of brain and genetic biomarkers that have been recorded previously in AD literature.

Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph

Irene Y. Chen[1], Monica Agrawal[1], Steven Horng[2], David Sontag[1]

[1]Massachusetts Institute of Technology, [2]Beth Israel Deaconess Medical Center

Increasingly large electronic health records (EHRs) provide an opportunity to algorithmically learn medical knowledge. In one prominent example, a causal health knowledge graph could learn relationships between diseases and symptoms and then serve as a diagnostic tool to be refined with additional clinical input. Prior research has demonstrated the ability to construct such a graph from over 270,000 emergency department patient visits. In this work, we describe methods to evaluate a health knowledge graph for robustness. Moving beyond precision and recall, we analyze for which diseases and for which patients the graph is most accurate. We identify sample size and unmeasured confounders as major sources of error in the health knowledge graph. We introduce a method to leverage non-linear functions in building the causal graph to better understand existing model assumptions. Finally, to assess model generalizability, we extend to a larger set of complete patient visits within a hospital system. We conclude with a discussion on how to robustly extract medical knowledge from EHRs.

Increasing Clinical Trial Accrual via Automated Matching of Biomarker Criteria

Jessica W. Chen, Christian A. Kunder, Nam Bui, James L. Zehnder, Helio A. Costa, Henning Stehr

Stanford University School of Medicine

Successful implementation of precision oncology requires both the deployment of nucleic acid sequencing panels to identify clinically actionable biomarkers, and the efficient screening of patient biomarker eligibility to on-going clinical trials and therapies. This process is typically performed manually by biocurators, geneticists, pathologists, and oncologists; however, this is a time-intensive, and inconsistent process amongst healthcare providers. We present the development of a feature matching algorithmic pipeline that identifies patients who meet eligibility criteria of precision medicine clinical trials via genetic biomarkers and apply it to patients undergoing treatment at the Stanford Cancer Center. This study demonstrates, through our patient eligibility screening algorithm that leverages clinical sequencing derived biomarkers with precision medicine clinical trials, the successful use of an automated algorithmic pipeline as a feasible, accurate and effective alternative to the traditional manual clinical trial curation.

Addressing the Credit Assignment Problem in Treatment Outcome Prediction using Temporal Difference Learning

Sahar Harati[1], Andrea Crowell[2], Helen Mayberg[3], Shamim Nemati[4]

[1]Stanford University, [2]Emory University, [3]Mount Sinai, [4]University of California San Diego

Mental health patients often undergo a variety of treatments before finding an effective one. Improved prediction of treatment response can shorten the duration of trials. A key challenge of applying predictive modeling to this problem is that often the effectiveness of a treatment regimen remains unknown for several weeks, and therefore immediate feedback signals may not be available for supervised learning. Here we propose a Machine Learning approach to extracting audio-visual features from weekly video interview recordings for predicting the likely outcome of Deep Brain Stimulation (DBS) treatment several weeks in advance. In the absence of immediate treatment-response feedback, we utilize a joint state-estimation and temporal difference learning approach to model both the trajectory of a patient's response and the delayed nature of feedbacks. Our results based on longitudinal recordings from 12 patients with depression show that the learned state values are predictive of the long-term success of DBS treatments. We achieve an area under the receiver operating characteristic curve of 0.88, beating all baseline methods.

From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer

Yifeng Tao[1], Chunhui Cai[2], William W. Cohen[1], Xinghua Lu[2]

[1]Carnegie Mellon University, [2]University of Pittsburgh

Cancers are mainly caused by somatic genomic alterations (SGAs) that perturb cellular signaling systems and eventually activate oncogenic processes. Therefore, understanding the functional impact of SGAs is a fundamental task in cancer biology and precision oncology. Here, we present a deep neural network model with encoder-decoder architecture, referred to as genomic impact transformer (GIT), to infer the functional impact of SGAs on cellular signaling systems through modeling the statistical relationships between SGA events and differentially expressed genes (DEGs) in tumors. The model utilizes a multi-head self-attention mechanism to identify SGAs that likely cause DEGs, or in other words, differentiating potential driver SGAs from passenger ones in a tumor. GIT model learns a vector (gene embedding) as an abstract representation of functional impact for each SGA-affected gene. Given SGAs of a tumor, the model can instantiate the states of the hidden layer, providing an abstract representation (tumor embedding) reflecting characteristics of perturbed molecular/cellular processes in the tumor, which in turn can be used to predict multiple phenotypes. We apply the GIT model to 4,468 tumors profiled by The Cancer Genome Atlas (TCGA) project. The attention mechanism enables the model to better capture the statistical relationship between SGAs and DEGs than conventional methods, and distinguishes cancer drivers from passengers. The learned gene embeddings capture the functional similarity of SGAs perturbing common pathways. The tumor embeddings are shown to be useful for tumor status representation, and phenotype prediction including patient survival time and drug response of cancer cell lines.

Automated phenotyping of patients with non-alcoholic fatty liver disease reveals clinically relevant disease subtypes

Maxence Vandromme, Tomi Jun, Ponni Perumalswami, Joel T. Dudley, Andrea Branch, Li Li

Icahn School of Medicine at Mount Sinai, Sema4

Non-alcoholic fatty liver disease (NAFLD) is a complex heterogeneous disease which affects more than 20% of the population worldwide. Some subtypes of NAFLD have been clinically identified using hypothesis-driven methods. In this study, we used data mining techniques to search for subtypes in an unbiased fashion. Using electronic signatures of the disease, we identified a cohort of 13,290 patients with NAFLD from a hospital database. We gathered clinical data from multiple sources and applied unsupervised clustering to identify five subtypes among this cohort. Descriptive statistics and survival analysis showed that the subtypes were clinically distinct and were associated with different rates of death, cirrhosis, hepatocellular carcinoma, chronic kidney disease, cardiovascular disease, and myocardial infarction. Novel disease subtypes identified in this manner could be used to risk-stratify patients and guide management.

Monitoring ICU Mortality Risk with A Long Short-Term Memory Recurrent Neural Network

Ke Yu[1], Mingda Zhang[2], Tianyi Cui[2], Milos Hauskrecht[2]

[1]Intelligent Systems Program, University of Pittsburgh; [2]Department of Computer Science, University of Pittsburgh

In intensive care units (ICU), mortality prediction is a critical factor not only for effective medical intervention but also for allocation of clinical resources. Structured electronic health records (EHR) contain valuable information for assessing mortality risk in ICU patients, but current mortality prediction models usually require laborious human-engineered features. Furthermore, substantial missing data in EHR is a common problem for both the construction and implementation of a prediction model. Inspired by language-related models, we design a new framework for dynamic monitoring of patients' mortality risk. Our framework uses the bag-of-words representation for all relevant medical events based on most recent history as inputs. By design, it is robust to missing data in EHR and can be easily implemented as an instant scoring system to monitor the medical development of all ICU patients. Specifically, our model uses latent semantic analysis (LSA) to encode the patients' states into low-dimensional embeddings, which are further fed to long short-term memory networks for mortality risk prediction. Our results show that the deep learning based framework performs better than the existing severity scoring system, SAPS-II. We observe that bidirectional long short-term memory demonstrates superior performance, probably due to the successful capture of both forward and backward temporal dependencies.

INTRINSICALLY DISORDERED PROTEINS (IDPS) AND THEIR FUNCTIONS

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

Disordered Function Conjunction: On the in-silico function annotation of intrinsically disordered regions

Sina Ghadermarzi, Akila Katuwawala, Christopher J. Oldfield, Amita Barik, **Lukasz Kurgan**

Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA

Intrinsically disorder regions (IDRs) lack a stable structure, yet perform biological functions. The functions of IDRs include mediating interactions with other molecules, including proteins, DNA, or RNA and entropic functions, including domain linkers. Computational predictors provide residue-level indications of function for disordered proteins, which contrasts with the need to functionally annotate the thousands of experimentally and computationally discovered IDRs. In this work, we investigate the feasibility of using residue-level prediction methods for region-level function predictions. For an initial examination of the multiple function region-level prediction problem, we constructed a dataset of (likely) single function IDRs in proteins that are dissimilar to the training datasets of the residue-level function predictors. We find that available residue-level prediction methods are only modestly useful in predicting multiple region-level functions. Classification is enhanced by simultaneous use of multiple residue-level function predictions and is further improved by inclusion of amino acids content extracted from the protein sequence. We conclude that multifunction prediction for IDRs is feasible and benefits from the results produced by current residue-level function predictors, however, it has to accommodate inaccuracy in functional annotations.

De novo ensemble modeling suggests that AP2-binding to disordered regions can increase steric volume of Epsin but not Eps15

N. Suhas Jagannathan[1], Christopher W. V. Hogue[2], **Lisa Tucker-Kellogg**[3]

[1]Duke-NUS Medical School; [2]National University of Singapore, 600 Epic Way Unit 345 San Jose CA 95134; [3]Cancer & Stem Cell Biology and Centre for Computational Biology Duke-NUS Medical School

Proteins with intrinsically disordered regions (IDRs) have large hydrodynamic radii, compared with globular proteins of equivalent weight. Recent experiments showed that IDRs with large radii can create steric pressure to drive membrane curvature during Clathrin-mediated endocytosis (CME). Epsin and Eps15 are two CME proteins with IDRs that contain multiple motifs for binding the adaptor protein AP2, but the impact of AP2-binding on these IDRs is unknown. Some IDRs acquire binding-induced function by forming a folded quaternary structure, but we hypothesize that the IDRs of Epsin and/or Eps15 acquire binding-induced function by increasing their steric volume. We explore this hypothesis in silico by generating conformational ensembles of the IDRs of Epsin (4 million structures) or Eps15 (3 million structures), then estimating the impact of AP2-binding on Radius of Gyration (RG). Results show that the ensemble of Epsin IDR conformations that accommodate AP2 binding has a right-shifted distribution of RG (larger radii) than the unbound Epsin ensemble. In contrast, the ensemble of Eps15 IDR conformations has comparable RG distribution between AP2-bound and unbound. We speculate that AP2 triggers the Epsin IDR to function through binding-induced-expansion, which could increase steric pressure and membrane bending during CME.

Modulation of p53 Transactivation Domain Conformations by Ligand Binding and Cancer-Associated Mutations

Xiaorong Liu, **Jianhan Chen**

University of Massachusetts Amherst

Intrinsically disordered proteins (IDPs) are important functional proteins, and their deregulation are linked to numerous human diseases including cancers. Understanding how disease-associated mutations or drug molecules can perturb the sequence-disordered ensemble-function-disease relationship of IDPs remains challenging, because it requires detailed characterization of the heterogeneous structural ensembles of IDPs. In this work, we combine the latest atomistic force field a99SB-disp, enhanced sampling technique replica exchange with solute tempering, and GPU- accelerated molecular dynamics simulations to investigate how four cancer-associated mutations, K24N, N29K/N30D, D49Y, and W53G, and binding of an anti-cancer molecule, epigallocatechin gallate (EGCG), modulate the disordered ensemble of the transactivation domain (TAD) of tumor suppressor p53. Through extensive sampling, in excess of 1.0 μs per replica, well-converged structural ensembles of wild-type and mutant p53-TAD as well as WT p53-TAD in the presence of EGCG were generated. The results reveal that mutants could induce local structural changes and affect secondary structural properties. Interestingly, both EGCG binding and N29K/N30D could also induce long-range structural reorganizations and lead to more compact structures that could shield key binding sites of p53-TAD regulators. Further analysis reveals that the effects of EGCG binding are mainly achieved through nonspecific interactions. These observations are generally consistent with on-going NMR studies and binding assays. Our studies suggest that induced conformational collapse of IDPs may be a general mechanism for shielding functional sites, thus inhibiting recognition of their targets. The current study also demonstrates that atomistic simulations provide a viable approach for studying the sequence-disordered ensemble-function- disease relationships of IDPs and developing new drug design strategies targeting regulatory IDPs.

Exploring Relationships between the Density of Charged Tracts within Disordered Regions and Phase Separation

Ramiz Somjee[1,2], Diana M. Mitrea[1], Richard W. Kriwacki[1,3]

[1]St. Jude Children's Research Hospital; [2]Rhodes College, [3]University of Tennessee Health Sciences Center

Biomolecular condensates form through a process termed phase separation and play diverse roles throughout the cell. Proteins that undergo phase separation often have disordered regions that can engage in weak, multivalent interactions; however, our understanding of the sequence grammar that defines which proteins phase separate is far from complete. Here, we show that proteins that display a high density of charged tracts within intrinsically disordered regions are likely to be constituents of electrostatically organized biomolecular condensates. We scored the human proteome using an algorithm termed ABTdensity that quantifies the density of charged tracts and observed that proteins with more charged tracts are enriched in particular Gene Ontology annotations and, based upon analysis of interaction networks, cluster into distinct biomolecular condensates. These results suggest that electrostatically-driven, multivalent interactions involving charged tracts within disordered regions serve to organize certain biomolecular condensates through phase separation.

MUTATIONAL SIGNATURES

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

PhySigs: Phylogenetic Inference of Mutational Signature Dynamics

Sarah Christensen[1], Mark D.M. Leiserson[2], Mohammed El-Kebir[1]

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

Distinct mutational processes shape the genomes of the clones comprising a tumor. These processes result in distinct mutational patterns, summarized by a small number of mutational signatures. Current analyses of clone-specific exposures to mutational signatures do not fully incorporate a tumor's evolutionary context, either inferring identical exposures for all tumor clones, or inferring exposures for each clone independently. Here, we introduce the Tree-constrained Exposure problem to infer a small number of exposure shifts along the edges of a given tumor phylogeny. Our algorithm, PhySigs, solves this problem and includes model selection to identify the number of exposure shifts that best explain the data. We validate our approach on simulated data and identify exposure shifts in lung cancer data, including at least one shift with a matching subclonal driver mutation in the mismatch repair pathway. Moreover, we show that our approach enables the prioritization of alternative phylogenies inferred from the same sequencing data. PhySigs is publicly available at https://github.com/elkebir-group/PhySigs

TrackSigFreq: subclonal reconstructions based on mutation signatures and allele frequencies

Caitlin F. Harrigan[1,2,4], Yulia Rubanova[1,2,4], Quaid Morris[1,2,3,4,5,6], Alina Selega[2,4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada; [2]Donnelly Centre for Cellular and Biomolecular Research , University of Toronto, Toronto, Canada; [3]Department of Molecular Genetics, University of Toronto, Toronto, Canada; [4]Vector Institute, Toronto, Canada; [5]Ontario Institute for Cancer Research, Toronto, Canada; [6]Memorial Sloan Kettering Cancer Centre, New York, USA (pending)

Mutational signatures are patterns of mutation types, many of which are linked to known mutagenic processes. Signature activity represents the proportion of mutations a signature generates. In cancer, cells may gain advantageous phenotypes through mutation accumulation, causing rapid growth of that subpopulation within the tumour. The presence of many subclones can make cancers harder to treat and have other clinical implications. Recon- structing changes in signature activities can give insight into the evolution of cells within a tumour. Recently, we introduced a new method, TrackSig, to detect changes in signature activities across time from single bulk tumour sample. By design, TrackSig is unable to identify mutation populations with different frequencies but little to no difference in signature activity. Here we present an extension of this method, TrackSigFreq, which enables trajectory reconstruction based on both observed density of mutation frequencies and changes in mutational signature activities. TrackSigFreq preserves the advantages of TrackSig, namely optimal and rapid mutation clustering through segmentation, while extending it so that it can identify distinct mutation populations that share similar signature activities.

DNA Repair Footprint Uncovers Contribution of DNA Repair Mechanism to Mutational Signatures

Damian Wojtowicz[1], Mark D.M. Leiserson[2], Roded Sharan[3], **Teresa M. Przytycka**[1]

[1]NIH, [2]University of Maryland, [3]Tel Aviv University

Cancer genomes accumulate a large number of somatic mutations resulting from imperfection of DNA processing during normal cell cycle as well as from carcinogenic exposures or cancer related aberrations of DNA maintenance machinery. These processes often lead to distinctive patterns of mutations, called mutational signatures. Several computational methods have been developed to uncover such signatures from catalogs of somatic mutations. However, cancer mutational signatures are the end-effect of several interplaying factors including carcinogenic exposures and potential deficiencies of the DNA repair mechanism. To fully understand the nature of each signature, it is important to disambiguate the atomic components that contribute to the final signature. Here, we introduce a new descriptor of mutational signatures, DNA Repair FootPrint (RePrint), and show that it can capture common properties of deficiencies in repair mechanisms contributing to diverse signatures. We validate the method with published mutational signatures from cell lines targeted with CRISPR-Cas9-based knockouts of DNA repair genes.

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

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data

Andrew L. Beam[1], **Benjamin Kompa**[2], Allen Schmaltz[1], Inbar Fried[3], Griffin Weber[2], Nathan Palmer[2], Xu Shi[1], Tianxi Cai[1], Isaac S. Kohane[3]

[1]Harvard T.H. Chan School of Public Health, [2]Harvard Medical School, [3]University of North Carolina School of Medicine

Word embeddings are a popular approach to unsupervised learning of word relationships that are widely used in natural language processing. In this article, we present a new set of embeddings for medical concepts learned using an extremely large collection of multimodal medical data. Leaning on recent theoretical insights, we demonstrate how an insurance claims database of 60 million members, a collection of 20 million clinical notes, and 1.7 million full text biomedical journal articles can be combined to embed concepts into a common space, resulting in the largest ever set of embeddings for 108,477 medical concepts. To evaluate our approach, we present a new benchmark methodology based on statistical power specifically designed to test embeddings of medical concepts. Our approach, called cui2vec, attains state-of-the-art performance relative to previous methods in most instances. Finally, we provide a downloadable set of pre-trained embeddings for other researchers to use, as well as an online tool for interactive exploration of the cui2vec embeddings.

Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients

Carly A. Bobak, Lauren McDonnell, Matthew D. Nemesure, Justin Lin, Jane E. Hill

Dartmouth College

The growth of publicly available repositories, such as the Gene Expression Omnibus, has allowed researchers to conduct meta-analysis of gene expression data across distinct cohorts. In this work, we assess eight imputation methods for their ability to impute gene expression data when values are missing across an entire cohort of Tuberculosis (TB) patients. We investigate how varying proportions of missing data (across 10%, 20%, and 30% of patient samples) influence the imputation results, and test for significantly differentially expressed genes and enriched pathways in patients with active TB. Our results indicate that truncating to common genes observed across cohorts, which is the current method used by researchers, results in the exclusion of important biology and suggest that LASSO and LLS imputation methodologies can reasonably impute genes across cohorts when total missingness rates are below 20%.

Towards identifying drug side effects from social media using active learning and crowd sourcing

Sophie Burkhardt, Julia Siekiera, Josua Glodde, Miguel A. Andrade-Navarro, Stefan Kramer

University of Mainz

Motivation: Social media is a largely untapped source of information on side effects of drugs. Twitter in particular is widely used to report on everyday events and personal ailments. However, labeling this noisy data is a difficult problem because labeled training data is sparse and automatic labeling is error-prone. Crowd sourcing can help in such a scenario to obtain more reliable labels, but is expensive in comparison because workers have to be paid. To remedy this, semi-supervised active learning may reduce the number of labeled data needed and focus the manual labeling process on important information. Results: We extracted data from Twitter using the public API. We subsequently use Amazon Mechanical Turk in combination with a state-of-the-art semi-supervised active learning method to label tweets with their associated drugs and side effects in two stages. Our results show that our method is an effective way of discovering side effects in tweets with an improvement from 53% F-measure to 67% F-measure as compared to a one stage work flow. Additionally, we show the effectiveness of the active learning scheme in reducing the labeling cost in comparison to a non-active baseline.

Microvascular Dynamics from 4D Microscopy Using Temporal Segmentation

Shir Gur, **Lior Wolf**, Lior Golgher, Pablo Blinder

Tel Aviv University

Recently developed methods for rapid continuous volumetric two-photon microscopy facilitate the observation of neuronal activity in hundreds of individual neurons and changes in blood flow in adjacent blood vessels across a large volume of living brain at unprecedented spatio-temporal resolution. However, the high imaging rate necessitates fully automated image analysis, whereas tissue turbidity and photo-toxicity limitations lead to extremely sparse and noisy imagery. In this work, we extend a recently proposed deep learning volumetric blood vessel segmentation network, such that it supports temporal analysis. With this technology, we are able to track changes in cerebral blood volume over time and identify spontaneous arterial dilations that propagate towards the pial surface. This new capability is a promising step towards characterizing the hemodynamic response function upon which functional magnetic resonance imaging (fMRI) is based.

Using Transcriptional Signatures to Find Cancer Drivers with LURE

David Haan, Ruikang Tao, Verena Friedl, Ioannis N. Anastopoulos, Christopher K. Wong, Alana S. Weinstein, Joshua M. Stuart

Dept. of Biomolecular Engineering and UC Santa Cruz Genomics Institute, University Of California Santa Cruz, Santa Cruz, CA 95064 USA

Cancer genome projects have produced multidimensional datasets on thousands of samples.Yet, depending on the tumor type, 5-50% of samples have no known driving event. Weintroduce a semi-supervised method called Learning UnRealized Events (LURE) that usesa progressive label learning framework and minimum spanning analysis to predict cancerdrivers based on their altered samples sharing a gene expression signature with the samplesof a known event. We demonstrate the utility of the method on the TCGA Pan-Cancer At-las dataset for which it produced a high-confidence result relating 59 new connections to 18known mutation events including alterations in the same gene, family, and pathway. We giveexamples of predicted drivers involved in TP53, telomere maintenance, and MAPK/RTKsignaling pathways. LURE identifies connections between genes with no known prior rela-tionship, some of which may offer clues for targeting specific forms of cancer. Code and Sup-plemental Material are available on the LURE website: https://sysbiowiki.soe.ucsc.edu/lure.

PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data

Jie Hao[1], Sai Chandra Kosaraju[2], Nelson Zange Tsaku[3], Dae Hyun Song[4], Mingon Kang[2]

[1]University of Pennsylvania, [2]University of Nevada Las Vegas, [3]Kennesaw State University, [4]Gyeongsang National University Changwon Hospital

The integration of multi-modal data, such as histopathological images and genomic data, is essential for understanding cancer heterogeneity and complexity for personalized treatments, as well as for enhancing survival predictions in cancer study. Histopathology, as a clinical gold-standard tool for diagnosis and prognosis in cancers, allows clinicians to make precise decisions on therapies, whereas high-throughput genomic data have been investigated to dissect the genetic mechanisms of cancers. We propose a biologically interpretable deep learning model (PAGE-Net) that integrates histopathological images and genomic data, not only to improve survival prediction, but also to identify genetic and histopathological patterns that cause different survival rates in patients. PAGE-Net consists of pathology/genome/demography-specific layers, each of which provides comprehensive biological interpretation. In particular, we propose a novel patch-wise texture-based convolutional neural network, with a patch aggregation strategy, to extract global survival-discriminative features, without manual annotation for the pathology-specific layers. We adapted the pathway-based sparse deep neural network, named Cox-PASNet, for the genome-specific layers. The proposed deep learning model was assessed with the histopathological images and the gene expression data of Glioblastoma Multiforme (GBM) at The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA). PAGE-Net achieved a C-index of 0.702, which is higher than the results achieved with only histopathological images (0.509) and Cox-PASNet (0.640). More importantly, PAGE-Net can simultaneously identify histopathological and genomic prognostic factors associated with patients' survivals. The source code of PAGE-Net is publicly available at https://github.com/DataX-JieHao/PAGE-Net

Machine learning algorithms for simultaneous supervised detection of peaks in multiple samples and cell types

Toby Dylan Hocking[1], Guillaume Bourque[2]

[1]Northern Arizona University, [2]McGill University

Joint peak detection is a central problem when comparing samples in epigenomic data analysis, but current algorithms for this task are unsupervised and limited to at most two sample types. We propose PeakSegPipeline, a new genome-wide multi-sample peak calling pipeline for epigenomic data sets. It performs peak detection using a constrained maximum likelihood segmentation model with essentially only one free parameter that needs to be tuned: the number of peaks. To select the number of peaks, we propose to learn a penalty function based on user-provided labels that indicate genomic regions with or without peaks in specific samples. In comparisons with state-of-the-art peak detection algorithms, PeakSegPipeline achieves similar or better accuracy, and a more interpretable model with overlapping peaks that occur in exactly the same positions across all samples. Our novel approach is able to learn that predicted peak sizes vary by experiment type.

Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks

Minh Pham, Olivier Lichtarge

Baylor College of Medicine

Shortest path length methods are routinely used to validate whether genes of interest are functionally related to each other based on biological network information. However, the methods are computationally intensive, impeding extensive utilization of network information. In addition, non- weighted shortest path length approach, which is more frequently used, often treat all network connections equally without taking into account of confidence levels of the associations. On the other hand, graph-based information diffusion method, which employs both the presence and confidence weights of network edges, can efficiently explore large networks and has previously detected meaningful biological patterns. Therefore, in this study, we hypothesized that the graph-based information diffusion method could prioritize genes with relevant functions more efficiently and accurately than the shortest path length approaches. We demonstrated that the graph-based information diffusion method substantially differentiated not only genes participating in same biological pathways ($p << 0.0001$) but also genes associated with specific human drug-induced clinical symptoms ($p << 0.0001$) from random. Furthermore, the diffusion method prioritized these functionally related genes faster and more accurately than the shortest path length approaches (pathways: $p = 2.7e-28$, clinical symptoms: $p = 0.032$). These data show the graph-based information diffusion method can be routinely used for robust prioritization of functionally related genes, facilitating efficient network validation and hypothesis generation, especially for human phenotype- specific genes.

A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases

Daniel N. Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, Russ B. Altman

Stanford University

Millions of Americans are affected by rare diseases, many of which have poor survival rates. However, the small market size of individual rare diseases, combined with the time and capital requirements of pharmaceutical R&D, have hindered the development of new drugs for these cases. A promising alternative is drug repurposing, whereby existing FDA-approved drugs might be used to treat diseases different from their original indications. In order to generate drug repurposing hypotheses in a systematic and comprehensive fashion, it is essential to integrate information from across the literature of pharmacology, genetics, and pathology. To this end, we leverage a newly developed knowledge graph, the Global Network of Biomedical Relationships (GNBR). GNBR is a large, heterogeneous knowledge graph comprising drug, disease, and gene (or protein) entities linked by a small set of semantic "themes" derived from the abstracts of biomedical literature. We apply a knowledge graph embedding method that explicitly models the uncertainty associated with literature-derived relationships and uses link prediction to generate drug repurposing hypotheses. This approach achieves high performance on a gold-standard test set of known drug indications (AUROC = 0.89) and is capable of generating novel repurposing hypotheses, which we independently validate using external literature sources and protein interaction networks. Finally, we demonstrate the ability of our model to produce explanations of its predictions.

Two-stage ML Classifier for Identifying Host Protein Targets of the Dengue Protease

Jacob T. Stanley, Alison R. Gilchrist, Alex C. Stabell, Mary A. Allen, Sara L. Sawyer, Robin D. Dowell

Department of Molecular, Cellular and Developmental Biology; BioFrontiers Institute; University of Colorado Boulder (all authors have the same affiliation)

Flaviviruses such as dengue encode a protease that is essential for viral replication. The protease functions by cleaving well-conserved positions in the viral polyprotein. In addition to the viral polyprotein, the dengue protease cleaves at least one host protein involved in immune response. This raises the question, what other host proteins are targeted and cleaved? Here we present a new computational method for identifying putative host protein targets of the dengue virus protease. Our method relies on biochemical and secondary structure features at the known cleavage sites in the viral polyprotein in a two-stage classification process to identify putative cleavage targets. The accuracy of our predictions scaled inversely with evolutionary distance when we applied it to the known cleavage sites of several other flaviviruses---a good indication of the validity of our predictions. Ultimately, our classifier identified 257 human protein sites possessing both a similar target motif and accessible local structure. These proteins are promising candidates for further investigation. As the number of viral sequences expands, our method could be adopted to predict host targets of other flaviviruses.

Enhancing Model Interpretability and Accuracy for Disease Progression Prediction via Phenotype-Based Patient Similarity Learning

Yue Wang[1], Tong Wu[1,2], Yunlong Wang[1], Gao Wang[3]

[1]IQVIA Inc., [2]University of Minnesota, [3] University of Chicago

Models have been proposed to extract temporal patterns from longitudinal electronic health records (EHR) for clinical predictive models. However, the common relations among patients (e.g., receiving the same medical treatments) were rarely considered. In this paper, we propose to learn patient similarity features as phenotypes from the aggregated patient-medical service matrix using non-negative matrix factorization. On real-world medical claim data, we show that the learned phenotypes are coherent within each group, and also explanatory and indicative of targeted diseases. We conducted experiments to predict the diagnoses for Chronic Lymphocytic Leukemia (CLL) patients. Results show that the phenotype-based similarity features can improve prediction over multiple baselines, including logistic regression, random forest, convolutional neural network, and more.

PRECISION MEDICINE: ADDRESSING THE CHALLENGES OF SHARING, ANALYSIS, AND PRIVACY AT SCALE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

Integrated Cancer Subtyping using Heterogeneous Genome-Scale Molecular Datasets

Suzan Arslanturk[1], **Sorin Draghici**[1], Tin Nguyen[2]

[1]Wayne State University, [2]University of Nevada

Vast repositories of heterogeneous data from existing sources present unique opportunities. Taken individually, each of the datasets offers solutions to important domain and source-specific questions. Collectively, they represent complementary views of related data entities with an aggregate information value often well exceeding the sum of its parts. Integration of heterogeneous data is therefore paramount to i) obtain a more unified picture and comprehensive view of the relations, ii) achieve more robust results, iii) improve the accuracy and integrity, and iv) illuminate the complex interactions among data features. In this paper, we have proposed a data integration methodology to identify subtypes of cancer using multiple data types (mRNA, methylation, microRNA and somatic variants) and different data scales that come from different platforms (microarray, sequencing, etc.). The Cancer Genome Atlas (TCGA) dataset is used to build the data integration and cancer subtyping framework. The proposed data integration and disease subtyping approach accurately identifies novel subgroups of patients with significantly different survival profiles. With current availability of vast genomics, and variant data for cancer, the proposed data integration system will better differentiate cancer and patient subtypes for risk and outcome prediction and targeted treatment planning without additional cost and precious lost time.

Assessment of coverage for endogenous metabolites and exogenous chemical compounds using an untargeted metabolomics platform

Sek Won Kong[1], Carles Hernandez-Ferrer[2]

[1]Computational Health Informatics Program, Boston Children's Hospital, 300 Longwood Avenue Boston, MA 02115, USA; [2]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

Physiological status and pathological changes in an individual can be captured by metabolic state that reflects the influence of both genetic variants and environmental factors such as diet, lifestyle and gut microbiome. The totality of environmental exposure throughout lifetime – i.e., exposome – is difficult to measure with current technologies. However, targeted measurement of exogenous chemicals and untargeted profiling of endogenous metabolites have been widely used to discover biomarkers of pathophysiologic changes and to understand functional impacts of genetic variants. To investigate the coverage of chemical space and interindividual variation related to demographic and pathological conditions, we profiled 169 plasma samples using an untargeted metabolomics platform. On average, 1,009 metabolites were quantified in each individual (range 906 – 1,038) out of 1,244 total chemical compounds detected in our cohort. Of note, age was positively correlated with the total number of detected metabolites in both males and females. Using the robust Qn estimator, we found metabolite outliers in each sample (mean 22, range from 7 to 86). A total of 50 metabolites were outliers in a patient with phenylketonuria including the ones known for phenylalanine pathway suggesting multiple metabolic pathways perturbed in this patient. The largest number of outliers (N=86) was found in a 5-year-old boy with alpha-1-antitrypsin deficiency who were waiting for liver transplantation due to cirrhosis. Xenobiotics including drugs, diets and environmental chemicals were significantly correlated with diverse endogenous metabolites and the use of antibiotics significantly changed gut microbial products detected in host circulation. Several challenges such as annotation of features, reference range and variance for each feature per age group and gender, and population scale reference datasets need to be addressed; however, untargeted metabolomics could be immediately deployed as a biomarker discovery platform and to evaluate the impact of genomic variants and exposures on metabolic pathways for some diseases.

Coverage profile correction of shallow-depth circulating cell-free DNA sequencing via multi-distance learning

Nicholas B. Larson, Melissa C. Larson, Jie Na, Carlos P. Sosa, Chen Wang, Jean-Pierre Kocher, Ross Rowsey

Mayo Clinic College of Medicine and Sciences

Shallow-depth whole-genome sequencing (WGS) of circulating cell-free DNA (ccfDNA) is a popular approach for non-invasive genomic screening assays, including liquid biopsy for early detection of invasive tumors as well as non-invasive prenatal screening (NIPS) for common fetal trisomies. In contrast to nuclear DNA WGS, ccfDNA WGS exhibits extensive inter- and intra-sample coverage variability that is not fully explained by typical sources of variation in WGS, such as GC content. This variability may inflate false positive and false negative screening rates of copy-number alterations and aneuploidy, particularly if these features are present at a relatively low proportion of total sequenced content. Herein, we propose an empirically-driven coverage correction strategy that leverages prior annotation information in a multi-distance learning context to improve within-sample coverage profile correction. Specifically, we train a weighted k-nearest neighbors-style method on non-pregnant female donor ccfDNA WGS samples, and apply it to NIPS samples to evaluate coverage profile variability reduction. We additionally characterize improvement in the discrimination of positive fetal trisomy cases relative to normal controls, and compare our results against a more traditional regression-based approach to profile coverage correction based on GC content and mappability. Under cross-validation, performance measures indicated benefit to combining the two feature sets relative to either in isolation. We also observed substantial improvement in coverage profile variability reduction in leave-out clinical NIPS samples, with variability reduced by 26.5-53.5% relative to the standard regression-based method as quantified by median absolute deviation. Finally, we observed improvement discrimination for screening positive trisomy cases reducing ccfDNA WGS coverage variability while additionally improving NIPS trisomy screening assay performance. Overall, our results indicate that machine learning approaches can substantially improve ccfDNA WGS coverage profile correction and downstream analyses.

PGxMine: Text mining for curation of PharmGKB

Jake Lever[1], Julia M. Barbarino[2], Li Gong[2], Rachel Huddart[2], Katrin Sangkuhl[2], Ryan Whaley[2], Michelle Whirl-Carrillo[2], Mark Woon[2], Teri E. Klein[2,3], Russ B. Altman[1,2,3]

[1]Department of Bioengineering, Stanford University, Stanford, CA, 94305; [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, 94305; [3]Department of Medicine, Stanford University, Stanford, CA, 94305

Precision medicine tailors treatment to individuals personal data including differences in their genome. The Pharmacogenomics Knowledgebase (PharmGKB) provides highly curated information on the effect of genetic variation on drug response and side effects for a wide range of drugs. PharmGKB's scientific curators triage, review and annotate a large number of papers each year but the task is challenging. We present the PGxMine resource, a text-mined resource of pharmacogenomic associations from all accessible published literature to assist in the curation of PharmGKB. We developed a supervised machine learning pipeline to extract associations between a variant (DNA and protein changes, star alleles and dbSNP identifiers) and a chemical. PGxMine covers 452 chemicals and 2,426 variants and contains 19,930 mentions of pharmacogenomic associations across 7,170 papers. An evaluation by PharmGKB curators found that 57 of the top 100 associations not found in PharmGKB led to 83 curatable papers and a further 24 associations would likely lead to curatable papers through citations. The results can be viewed at https://pgxmine.pharmgkb.org/ and code can be downloaded at https://github.com/jakelever/pgxmine.

The power of dynamic social networks to predict individuals' mental health

Shikang Liu[1], David Hachen[1], Omar Lizardo[2], Christian Poellabauer[1], Aaron Striegel[1], Tijana Milenkovic[1]

[1]University of Notre Dame, [2]University of California Los Angeles

Precision medicine has received attention both in and outside the clinic. We focus on the latter, by exploiting the relationship between individuals' social interactions and their mental health to predict one's likelihood of being depressed or anxious from rich dynamic social network data. Existing studies differ from our work in at least one aspect: they do not model social interaction data as a network; they do so but analyze static network data; they examine ''correlation'' between social networks and health but without making any predictions; or they study other individual traits but not mental health. In a comprehensive evaluation, we show that our predictive model that uses dynamic social network data is superior to its static network as well as non-network equivalents when run on the same data.

Implementing a Cloud Based Method for Protected Clinical Trial Data Sharing

Gaurav Luthria, Qingbo Wang

Harvard University

Clinical trials generate a large amount of data that have been underutilized due to obstacles that prevent data sharing including risking patient privacy, data misrepresentation, and invalid secondary analyses. In order to address these obstacles, we developed a novel data sharing method which ensures patient privacy while also protecting the interests of clinical trial investigators. Our flexible and robust approach involves two components: (1) an advanced cloud-based querying language that allows users to test hypotheses without direct access to the real clinical trial data and (2) corresponding synthetic data for the query of interest that allows for exploratory research and model development. Both components can be modified by the clinical trial investigator depending on factors such as the type of trial or number of patients enrolled. To test the effectiveness of our system, we first implement a simple and robust permutation based synthetic data generator. We then use the synthetic data generator coupled with our querying language to identify significant relationships among variables in a realistic clinical trial dataset.

Pathway and network embedding methods for prioritizing psychiatric drugs

Yash Pershad[1], Margaret Guo[2], Russ B. Altman[3]

[1]Stanford University Department of Bioengineering, [2]Stanford University Biomedical Informatics Program, [3]Stanford University Departments of Bioengineering, Genetics, & Medicine

One in five Americans experience mental illness, and roughly 75% of psychiatric prescriptions do not successfully treat the patient's condition. Extensive evidence implicates genetic factors and signaling disruption in the pathophysiology of these diseases. Changes in transcription often underlie this molecular pathway dysregulation; individual patient transcriptional data can improve the efficacy of diagnosis and treatment. Recent large-scale genomic studies have uncovered shared genetic modules across multiple psychiatric disorders — providing an opportunity for an integrated multi-disease approach for diagnosis. Moreover, network-based models informed by gene expression can represent pathological biological mechanisms and suggest new genes for diagnosis and treatment. Here, we use patient gene expression data from multiple studies to classify psychiatric diseases, integrate knowledge from expert-curated databases and publicly available experimental data to create augmented disease-specific gene sets, and use these to recommend disease-relevant drugs. From Gene Expression Omnibus, we extract expression data from 145 cases of schizophrenia, 82 cases of bipolar disorder, 190 cases of major depressive disorder, and 307 shared controls. We use pathway-based approaches to predict psychiatric disease diagnosis with a random forest model (78% accuracy) and derive important features to augment available drug and disease signatures. Using protein-protein-interaction networks and embedding-based methods, we build a pipeline to prioritize treatments for psychiatric diseases that achieves a 3.4-fold improvement over a background model. Thus, we demonstrate that gene-expression-derived pathway features can diagnose psychiatric diseases and that molecular insights derived from this classification task can inform treatment prioritization for psychiatric diseases.

Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data

Jiayi Tong[1], Rui Duan[1], Ruowang Li[1], Martijn J. Scheuemie[2], Jason H. Moore[1], Yong Chen[1]

[1]University of Pennsylvania, [2]Janssen Research and Development LLC

Electronic Health Records (EHR) contain extensive patient data on various health outcomes and risk predictors, providing an efficient and wide-reaching source for health research. Integrated EHR data can provide a larger sample size of the population to improve estimation and prediction accuracy. To overcome the obstacle of sharing patient-level data, distributed algorithms were developed to conduct statistical analyses across multiple clinical sites through sharing only aggregated information. However, the heterogeneity of data across sites is often ignored by existing distributed algorithms, which leads to substantial bias when studying the association between the outcomes and exposures. In this study, we propose a privacy-preserving and communication-efficient distributed algorithm which accounts for the heterogeneity caused by a small number of the clinical sites. We evaluated our algorithm through a systematic simulation study motivated by real-world scenarios and applied our algorithm to multiple claims datasets from the Observational Health Data Sciences and Informatics (OHDSI) network. The results showed that the proposed method performed better than the existing distributed algorithm ODAL and a meta-analysis method.

Computationally efficient, exact, covariate-adjusted genetic principal component analysis by leveraging individual marker summary statistics from large biobanks

Jack Wolf[1], Martha Barnard[1], Xueting Xia[2], Nathan Ryder[3], Jason Westra[4], **Nathan Tintle**[4]

[1]St. Olaf College, [2]Texas Tech University, [3]Colorado State University, [4]Dordt University

The popularization of biobanks provides an unprecedented amount of genetic and phenotypic information that can be used to research the relationship between genetics and human health. Despite the opportunities these datasets provide, they also pose many problems associated with computational time and costs, data size and transfer, and privacy and security. The publishing of summary statistics from these biobanks, and the use of them in a variety of downstream statistical analyses, alleviates many of these logistical problems. However, major questions remain about how to use summary statistics in all but the simplest downstream applications. Here, we present a novel approach to utilize basic summary statistics (estimates from single marker regressions on single phenotypes) to evaluate more complex phenotypes using multivariate methods. In particular, we present a covariate-adjusted method for conducting principal component analysis (PCA) utilizing only biobank summary statistics. We validate exact formulas for this method, as well as provide a framework of estimation when specific summary statistics are not available, through simulation. We apply our method to a real data set of fatty acid and genomic data.

ARTIFICIAL INTELLIGENCE FOR ENHANCING CLINICAL MEDICINE

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

Multiclass Disease Classification from Microbial Whole-Community Metagenomes

Saad Khan, Libusha Kelly

Albert Einstein College of Medicine

The microbiome, the community of microorganisms living within an individual, is a promising avenue for developing non-invasive methods for disease screening and diagnosis. Here, we utilize 5643 aggregated, annotated whole-community metagenomes to implement the first multiclass microbiome disease classifier of this scale, able to discriminate between 18 different diseases and healthy. We compared three different machine learning models: random forests, deep neural nets, and a novel graph convolutional architecture which exploits the graph structure of phylogenetic trees as its input. We show that the graph convolutional model outperforms deep neural nets in terms of accuracy (achieving 75% average test-set accuracy), receiver-operator-characteristics (92.1% average area-under-ROC (AUC)), and precision-recall (50% average area-under-precision-recall (AUPR)). Additionally, the convolutional net's performance complements that of the random forest, showing a lower propensity for Type-I errors (false-positives) while the random forest makes less Type-II errors (false-negatives). Lastly, we are able to achieve over 90% average top-3 accuracy across all of our models. Together, these results indicate that there are predictive, disease-specific signatures across microbiomes that can be used for diagnostic purposes.

LitGen: Genetic Literature Recommendation Guided by Human Explanations

Allen Nie[1], Arturo L. Pineda[1], Matt W. Wright[1], Hannah Wand[1], Bryan Wulf[1], Helio A. Costa[1], Ronak Y. Patel[2], Carlos D. Bustamante[1], James Zou[1]

[1]Stanford University, [2]Baylor College of Medicine

As genetic sequencing costs decrease, the lack of clinical interpretation of variants has become the bottleneck in using genetics data. A major rate limiting step in clinical interpretation is the manual curation of evidence in the genetic literature by highly trained biocurators. What makes curation particularly time-consuming is that the curator needs to identify papers that study variant pathogenicity using different types of approaches and evidences---e.g. biochemical assays or case control analysis. In collaboration with the Clinical Genomic Resource (ClinGen)---the flagship NIH program for clinical curation---we propose the first machine learning system, LitGen, that can retrieve papers for a particular variant and filter them by specific evidence types used by curators to assess for pathogenicity. LitGen uses semi-supervised deep learning to predict the type of evi+dence provided by each paper. It is trained on papers annotated by ClinGen curators and systematically evaluated on new test data collected by ClinGen. LitGen further leverages rich human explanations and unlabeled data to gain 7.9%-12.6% relative performance improvement over models learned only on the annotated papers. It is a useful framework to improve clinical variant curation.

Multilevel Self-Attention Model and its Use on Medical Risk Prediction

Xianlong Zeng[1,2], Yunyi Feng[1,2], Soheil Moosavinasab[2], Deborah Lin[2], Simon Lin[2], Chang Liu[1]

[1]School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA; [2]The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA

Various deep learning models have been developed for different healthcare predictive tasks using Electronic Health Records and have shown promising performance. In these models, medical codes are often aggregated into visit representation without considering their heterogeneity, e.g., the same diagnosis might imply different healthcare concerns with different procedures or medications. Then the visits are often fed into deep learning models, such as recurrent neural networks, sequentially without considering the irregular temporal information and dependencies among visits. To address these limitations, we developed a Multilevel Self-Attention Model (MSAM) that can capture the underlying relationships between medical codes and between medical visits. We compared MSAM with various baseline models on two predictive tasks, i.e., future disease prediction and future medical cost prediction, with two large datasets, i.e., MIMIC-3 and PFK. In the experiments, MSAM consistently outperformed baseline models. Additionally, for future medical cost prediction, we used disease prediction as an auxiliary task, which not only guides the model to achieve a stronger and more stable financial prediction, but also allows managed care organizations to provide a better care coordination.

Identifying Transitional High Cost Users from Unstructured Patient Profiles Written by Primary Care Physicians

Haoran Zhang[1,2,3], Elisa Candido[3], Andrew S. Wilton[3], Raquel Duchen[3], Liisa Jaakkimainen[3], Walter Wodchis[3,4,5], Quaid Morris[1,2,6,7]

[1]Department of Computer Science, University of Toronto; [2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada; [3]ICES, Toronto, Ontario, Canada; [4]Institute of Health Policy, Management, and Evaluation, University of Toronto; [5]Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada; [6]Terrence Donnelly Center for Cellular and Biomolecular Research, University of Toronto; [7]Department of Molecular Genetics, University of Toronto

Identification and subsequent intervention of patients at risk of becoming High Cost Users (HCUs) presents the opportunity to improve outcomes while also providing significant savings for the healthcare system. In this paper, the 2016 HCU status of patients was predicted using free-form text data from the 2015 cumulative patient profiles within the electronic medical records of family care practices in Ontario. These unstructured notes make substantial use of domain-specific spellings and abbreviations; we show that word embeddings derived from the same context provide more informative features than pre-trained ones based on Wikipedia, MIMIC, and Pubmed. We further demonstrate that a model using features derived from aggregated word embeddings (EmbEncode) provides a significant performance improvement over the bag-of-words representation (82.48±0.35% versus 81.85±0.36% held-out AUROC, p=3.2E-4), using far fewer input features (5,492 versus 214,750) and fewer non-zero coefficients (1,177 versus 4,284). The future HCUs of greatest interest are the transitional ones who are not already HCUs, because they provide the greatest scope for interventions. Predicting these new HCU is challenging because most HCUs recur. We show that removing recurrent HCUs from the training set improves the ability of EmbEncode to predict new HCUs, while only slightly decreasing its ability to predict recurrent ones.

Obtaining dual-energy computed tomography (CT) information from a single-energy CT image for quantitative imaging analysis of living subjects by using deep learning

Wei Zhao[1], Tianling Lv[2], Rena Lee[3], Yang Chen[2], **Lei Xing**[1]

[1]Stanford University, [2]Southeast University, [3]Ehwa Womens University

Computed tomographic (CT) is a fundamental imaging modality to generate cross-sectional views of internal anatomy in a living subject or interrogate material composition of an object, and it has been routinely used in clinical applications and nondestructive testing. In a standard CT image, pixels having the same Hounsfield Units (HU) can correspond to different materials, and it is therefore challenging to differentiate and quantify materials. Dual-energy CT (DECT) is desirable to differentiate multiple materials, but the costly DECT scanners are not widely available as single-energy CT (SECT) scanners. Recent advancement in deep learning provides an enabling tool to map images between different modalities with incorporated prior knowledge. Here we develop a deep learning approach to perform DECT imaging by using the standard SECT data. The end point of the approach is a model capable of providing the high-energy CT image for a given input low-energy CT image. The feasibility of the deep learning-based DECT imaging method using a SECT data is demonstrated using contrast-enhanced DECT images and evaluated using clinical relevant indexes. This work opens new opportunities for numerous DECT clinical applications with a standard SECT data and may enable significantly simplified hardware design, scanning dose, and image cost reduction for future DECT systems.

INTRINSICALLY DISORDERED PROTEINS (IDPS) AND THEIR FUNCTIONS

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

Many-to-one binding by intrinsically disordered protein regions

Wei-Lun Alterovitz[1*], Eshel Faraggi[1,2,3*], Christopher J. Oldfield[1], Jingwei Meng[1], Bin Xue[1], Fei Huang[1], Pedro Romero[1], Andrzej Kloczkowski[2], Vladimir N. Uversky[1], **A. Keith Dunker**[1]

[1]Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 410 W. 10th St, HS5000, Indianapolis, IN 46202, USA (kedunker@iupui.edu); [2]Battelle Center for Mathematical Medicine, and the Nationwide Children's Hospital, Department of Pediatrics, The Ohio State University, Columbus, OH 43210, USA; [3]Research and Information Systems, LLC, 1620 E. 72nd St. Indianapolis, IN 46240 USA

**Contributed equally (weilun.hsu@gmail.com, efaaggi@gmail.com)*

Disordered binding regions (DBRs), which are embedded within intrinsically disordered proteins or regions (IDPs or IDRs), enable IDPs or IDRs to mediate multiple protein-protein interactions. DBR-protein complexes were collected from the Protein Data Bank for which two or more DBRs having different amino acid sequences bind to the same (100% sequence identical) globular protein partner, a type of interaction herein called many-to-one binding. Two distinct binding profiles were identified: independent and overlapping. For the overlapping binding profiles, the distinct DBRs interact by means of almost identical binding sites (herein called "similar"), or the binding sites contain both common and divergent interaction residues (herein called "intersecting"). Further analysis of the sequence and structural differences among these three groups indicate how IDP flexibility allows different segments to adjust to similar, intersecting, and independent binding pockets.

MUTATIONAL SIGNATURES

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

Impact of mutational signatures on microRNA and their response elements

Eirini Stamoulakatou[1], Pietro Pinoli[1], Stefano Ceri[1], Rosario Piro[2]

[1]Politecnico di Milano, [2]Freie Universitat Berlin

MicroRNAs are a class of small non-coding RNA molecules with great importance for regulating a large number of diverse biological processes in health and disease, mostly by binding to complementary microRNA response elements (MREs) on protein-coding messenger RNAs and other non-coding RNAs and subsequently inducing their degradation. A growing body of evidence indicates that the dysregulation of certain microRNAs may either drive or suppress oncogenesis. The seed region of a microRNA is of crucial importance for its target recognition. Mutations in these seed regions may disrupt the binding of microRNAs to their target genes. In this study, we investigate the theoretical impact of cancer-associated mutagenic processes and their mutational signatures on microRNA seeds and their MREs. To our knowledge, this is the first study which provides a probabilistic framework for microRNA and MRE sequence alteration analysis based on mutational signatures and computationally assessing the disruptive impact of mutational signatures on human microRNA–target interactions.

Genome Gerrymandering: optimal divison of the genome into regions with cancer type specific differences in mutation rates

Adamo Young, Jacob Chmura, Yoonsik Park, Quaid Morris, Gurnit Atwal

University of Toronto

The activity of mutational processes differs across the genome, and is influenced by chromatin state and spatial genome organization. At the scale of one megabase-pair (Mb), regional mutation density correlate strongly with chromatin features and mutation density at this scale can be used to accurately identify cancer type. Here, we explore the relationship between genomic region and mutation rate by developing an information theory driven, dynamic programming algorithm for dividing the genome into regions with differing relative mutation rates between cancer types. Our algorithm improves mutual information when compared to the naive approach, effectively reducing the average number of mutations required to identify cancer type. Our approach provides an efficient method for associating regional mutation density with mutation labels, and has future applications in exploring the role of somatic mutations in a number of diseases.

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

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

Learning a Latent Space of Highly Multidimensional Cancer Data

Benjamin Kompa[1], Beau Coker[2]

[1]Harvard Medical School, [2]Harvard School of Public Health

We introduce a Unified Disentanglement Network (UFDN) trained on The Cancer Genome Atlas (TCGA), which we refer to as UFDN-TCGA. We demonstrate that UFDN-TCGA learns a biologically relevant, low-dimensional latent space of high-dimensional gene expression data by applying our network to two classification tasks of cancer status and cancer type. UFDN-TCGA performs comparably to random forest methods. The UFDN allows for continuous, partial interpolation between distinct cancer types. Furthermore, we perform an analysis of differentially expressed genes between skin cutaneous melanoma (SKCM) samples and the same samples interpolated into glioblastoma (GBM). We demonstrate that our interpolations consist of relevant metagenes that recapitulate known glioblastoma mechanisms.

Scaling structural learning with NO-BEARS to infer causal transcriptome networks

Hao-Chih Lee[1,3], Matteo Danieletto[1,2,3], Riccardo Miotto[1,2,3], Sarah T. Cherng[1,3], Joel T. Dudley[1,2,3]

[1]Institute for Next Generation Healthcare, [2]Hasso Plattner Institute for Digital Health, [3]Department of Genetics and Genomic Sciences Icahn School of Medicine at Mount Sinai New York, NY 10065, USA

Constructing gene regulatory networks is a critical step in revealing disease mechanisms from transcriptomic data. In this work, we present NO-BEARS, a novel algorithm for estimating gene regulatory networks. The NO-BEARS algorithm is built on the basis of the NO- TEARS algorithm with two improvements. First, we propose a new constraint and its fast approximation to reduce the computational cost of the NO-TEARS algorithm. Next, we introduce a polynomial regression loss to handle non-linearity in gene expressions. Our implementation utilizes modern GPU computation that can decrease the time of hours-long CPU computation to seconds. Using synthetic data, we demonstrate improved performance, both in processing time and accuracy, on inferring gene regulatory networks from gene expression data.

PathFlowAI: A High-Throughput Workflow for Preprocessing, Deep Learning and Interpretation in Digital Pathology

Joshua J. Levy[1], Lucas A. Salas[1], Brock C. Christensen[1], Aravindhan Sriharan[2], Louis J. Vaickus[2]

[1]Geisel School of Medicine at Dartmouth, [2]Dartmouth Hitchcock Medical Center

The diagnosis of disease often requires analysis of a biopsy. Many diagnoses depend not only on the presence of certain features but on their location within the tissue. Recently, a number of deep learning diagnostic aids have been developed to classify digitized biopsy slides. Clinical workflows often involve processing of more than 500 slides per day. But, clinical use of deep learning diagnostic aids would require a preprocessing workflow that is cost-effective, flexible, scalable, rapid, interpretable, and transparent. Here, we present such a workflow, optimized using Dask and mixed precision training via APEX, capable of handling any patch-level or slide level classification and prediction problem. The workflow uses a flexible and fast preprocessing and deep learning analytics pipeline, incorporates model interpretation and has a highly storage-efficient audit trail. We demonstrate the utility of this package on the analysis of a prototypical anatomic pathology specimen, liver biopsies for evaluation of hepatitis from a prospective cohort. The preliminary data indicate that PathFlowAI may become a cost-effective and time-efficient tool for clinical use of Artificial Intelligence (AI) algorithms.

Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data*

Lisa Neums, Richard Meier, Devin C. Koestler, Jeffrey A. Thompson

Department of Biostatistics and Data Science, University of Kansas Medical Center, and University of Kansas Cancer Center

The accurate prediction of a cancer patient's risk of progression or death can guide clinicians in the selection of treatment and help patients in planning personal affairs. Predictive models based on patient-level data represent a tool for determining risk. Ideally, predictive models will use multiple sources of data (e.g., clinical, demographic, molecular, etc.). However, there are many challenges associated with data integration, such as overfitting and redundant features. In this paper we aim to address those challenges through the development of a novel feature selection and feature reduction framework that can handle correlated data. Our method begins by computing a survival distance score for gene expression, which in combination with a score for clinical independence, results in the selection of highly predictive genes that are non-redundant with clinical features. The survival distance score is a measure of variation of gene expression over time, weighted by the variance of the gene expression over all patients. Selected genes, in combination with clinical data, are used to build a predictive model for survival. We benchmark our approach against commonly used methods, namely lasso- as well as ridge-penalized Cox proportional hazards models, using three publicly available cancer data sets: kidney cancer (521 samples), lung cancer (454 samples) and bladder cancer (335 samples). Across all data sets, our approach built on the training set outperformed the clinical data alone in the test set in terms of predictive power with a c.Index of 0.773 vs 0.755 for kidney cancer, 0.695 vs 0.664 for lung cancer and 0.648 vs 0.636 for bladder cancer. Further, we were able to show increased predictive performance of our method compared to lasso-penalized models fit to both gene expression and clinical data, which had a c.Index of 0.767, 0.677, and 0.645, as well as increased or comparable predictive power compared to ridge models, which had a c.Index of 0.773, 0.668 and 0.650 for the kidney, lung, and bladder cancer data sets, respectively. Therefore, our score for clinical independence improves prognostic performance as compared to modeling approaches that do not consider combining non-redundant data. Future work will concentrate on optimizing the survival distance score in order to achieve improved results for all types of cancer.

Bayesian semi-nonnegative matrix tri-factorization to identify pathways associated with cancer phenotypes

Sunho Park[1], Nabhonil Kar[1], Jae-Ho Cheong[2], Tae Hyun Hwang[1]

[1]Cleveland Clinic, [2]Yonsei University College of Medicine

Accurate identification of pathways associated with cancer phenotypes (e.g., cancer subtypes and treatment outcome) could lead to discovering reliable prognostic and/or predictive biomarkers for better patients stratification and treatment guidance. In our previous work, we have shown that non-negative matrix tri-factorization (NMTF) can be successfully applied to identify pathways associated with specific cancer types or disease classes as a prognostic and predictive biomarker. However, one key limitation of non-negative factorization methods, including various non-negative bi-factorization methods, is their limited ability to handle negative input data. For example, many molecular data that consist of real-values containing both positive and negative values (e.g., normalized/log transformed gene expression data where negative value represents down-regulated expression of genes) are not suitable input for these algorithms. In addition, most previous methods provide just a single point estimate and hence cannot deal with uncertainty effectively. To address these limitations, we propose a Bayesian semi-nonnegative matrix tri-factorization method to identify pathways associated with cancer phenotypes from a real-valued input matrix, e.g., gene expression values. Motivated by semi-nonnegative factorization, we allow one of the factor matrices, the centroid matrix, to be real-valued so that each centroid can express either the up- or down-regulation of the member genes in a pathway. In addition, we place structured spike-and-slab priors (which are encoded with the pathways and a gene-gene interaction (GGI) network) on the centroid matrix so that even a set of genes that is not initially contained in the pathways (due to the incompleteness of the current pathway database) can be involved in the factorization in a stochastic way specifically, if those genes are connected to the member genes of the pathways on the GGI network. We also present update rules for the posterior distributions in the framework of variational inference. As a full Bayesian method, our proposed method has several advantages over the current NMTF methods, which are demonstrated using synthetic datasets in experiments. Using the The Cancer Genome Atlas (TCGA) gastric cancer and metastatic gastric cancer immunotherapy clinical-trial datasets, we show that our method could identify biologically and clinically relevant pathways associated with the molecular subtypes and immunotherapy response, respectively. Finally, we show that those pathways identified by the proposed method could be used as prognostic biomarkers to stratify patients with distinct survival outcome in two independent validation datasets. Additional information and codes can be found at https://github.com/parks-cs-ccf/BayesianSNMTF.

Tree-Weighting for Multi-Study Ensemble Learners

Maya Ramchandran[1], Prasad Patil[1,2], Giovanni Parmigiani[1,2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health; Department of Biostatistics, Harvard T.H. Chan School of Public Health; [2]Department of Data Sciences, Dana-Farber Cancer Institute

Multi-study learning uses multiple training studies, separately trains classifiers on each, and forms an ensemble with weights rewarding members with better cross-study prediction ability. This article considers novel weighting approaches for constructing tree-based ensemble learners in this setting. Using Random Forests as a single-study learner, we compare weighting each forest to form the ensemble, to extracting the individual trees trained by each Random Forest and weighting them directly. We find that incorporating multiple layers of ensembling in the training process by weighting trees increases the robustness of the resulting predictor. Furthermore, we explore how ensembling weights correspond to tree structure, to shed light on the features that determine whether weighting trees directly is advantageous. Finally, we apply our approach to genomic datasets and show that weighting trees improves upon the basic multi-study learning paradigm. Code and supplementary material are available at https://github.com/m-ramchandran/tree-weighting.

PTR Explorer: An approach to identify and explore Post Transcriptional Regulatory mechanisms using proteogenomics

Arunima Srivastava[1], Michael Sharpnack[1], Kun Huang[2], Parag Mallick[3], Raghu Machiraju[1]

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

Integration of transcriptomic and proteomic data should reveal multi-layered regulatory processes governing cancer cell behaviors. Traditional correlation-based analyses have demonstrated limited ability to identify the post-transcriptional regulatory (PTR) processes that drive the non-linear relationship between transcript and protein abundances. In this work, we ideate an integrative approach to explore the variety of post-transcriptional mechanisms that dictate relationships between genes and corresponding proteins. The proposed workflow utilizes the intuitive technique of scatterplot diagnostics or scagnostics, to characterize and examine the diverse scatterplots built from transcript and protein abundances in a proteogenomic experiment. The workflow includes representing gene-protein relationships as scatterplots, clustering on geometric scagnostic features of these scatterplots, and finally identifying and grouping the potential gene-protein relationships according to their disposition to various PTR mechanisms. Our study verifies the efficacy of the implemented approach to excavate possible regulatory mechanisms by utilizing comprehensive tests on a synthetic dataset. We also propose a variety of 2D pattern-specific downstream analyses methodologies such as mixture modeling, and mapping miRNA post-transcriptional effects to explore each mechanism further. This work suggests that the proposed methodology has the potential for discovering and categorizing post-transcriptional regulatory mechanisms, manifesting in proteogenomic trends. These trends subsequently provide evidence for cancer specificity, miRNA targeting, and identification of regulation impacted by biological functionality and different types of degradation.

Network Representation of Large-Scale Heterogeneous RNA Sequences with Integration of Diverse Multi-omics, Interactions, and Annotations Data

Nhat Tran, **Jean Gao**

The University of Texas at Arlington

Long non-coding RNA (lncRNA), microRNA, and messenger RNA enable key regulations of various biological processes through a variety of diverse interaction mechanisms. Identifying the interactions and cross-talk between these heterogeneous RNA classes is essential in order to uncover the functional role of individual RNA transcripts, especially for unannotated and sparsely discovered RNA sequences with no known interactions. Recently, sequence-based deep learning and network embedding methods are gaining traction as high-performing and flexible approaches that can either predict RNA-RNA interactions from sequence or infer missing interactions from patterns that may exist in the network topology. However, most of the current methods have several limitations, e.g., the inability to perform inductive predictions, to distinguish the directionality of interactions, or to integrate various sequence, interaction, expression, and genomic annotation datasets. We proposed a novel deep learning framework, rna2rna, which learns from RNA sequences to produce a low-dimensional embedding that preserves proximities in both the interaction topology and the functional affinity topology. In this proposed embedding space, the two-part "source and target contexts" capture the receptive fields of each RNA transcript to encapsulate heterogeneous cross-talk interactions between lncRNAs and microRNAs. The proximity between RNAs in this embedding space also uncovers the second-order relationships that allow for accurate inference of novel directed interactions or functional similarities between any two RNA sequences. In a prospective evaluation, our method exhibits superior performance compared to state-of-art approaches at predicting missing interactions from several RNA-RNA interaction databases. Additional results suggest that our proposed framework can capture a manifold for heterogeneous RNA sequences to discover novel functional annotations.

Hadoop and PySpark for reproducibility and scalability of genomic sequencing studies

Nicholas R. Wheeler[1], Penelope Benchek[1], Brian W. Kunkle[2], Kara L. Hamilton-Nelson[2], Mike Warfe[1], Jeremy R. Fondran[1], Jonathan L. Haines[1], **William S. Bush**[1]

[1]Case Western Reserve University, [2]University of Miami

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 examine the use of the Spark infrastructure for the management, access, and analysis of genomic data in comparison to traditional genomic workflows on typical cluster environments. We validate the framework by reproducing previously published results from the Alzheimer's Disease Sequencing Project. Using the framework and analyses designed using Jupyter notebooks, Spark provides improved workflows, reduces user-driven data partitioning, and enhances the portability and reproducibility of distributed analyses required for large-scale genomic studies.

CERENKOV3: Clustering and molecular network-derived features improve computational prediction of functional noncoding SNPs

Yao Yao, Stephen A. Ramsey

Oregon State University

Identification of causal noncoding single nucleotide polymorphisms (SNPs) is important for maximizing the knowledge dividend from human genome-wide association studies (GWAS). Recently, diverse machine learning-based methods have been used for functional SNP identification; however, this task remains a fundamental challenge in computational biology. We report CERENKOV3, a machine learning pipeline that leverages clustering-derived and molecular network-derived features to improve prediction accuracy of regulatory SNPs (rSNPs) in the context of post-GWAS analysis. The clustering-derived feature, locus size (number of SNPs in the locus), derives from our locus partitioning procedure and represents the sizes of clusters based on SNP locations. We generated two molecular network-derived features from representation learning on a network representing SNP-gene and gene-gene relations. Based on empirical studies using a ground-truth SNP dataset, CERENKOV3 significantly improves rSNP recognition performance in AUPRC, AUROC, and AVGRANK (a locus-wise rank-based measure of classification accuracy we previously proposed).

PRECISION MEDICINE: ADDRESSING THE CHALLENGES OF SHARING, ANALYSIS, AND PRIVACY AT SCALE

PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

AnomiGAN: Generative Adversarial Networks for Anonymizing Private Medical Data

Ho Bae, Dahuin Jung, Hyun-Soo Choi, Sungroh Yoon

Seoul National University

Typical personal medical data contains sensitive information about individuals. Storing or sharing the personal medical data is thus often risky. For example, a short DNA sequence can provide information that can identify not only an individual, but also his or her relatives. Nonetheless, most countries and researchers agree on the necessity of collecting personal medical data. This stems from the fact that medical data, including genomic data, are an indispensable resource for further research and development regarding disease prevention and treatment. To prevent personal medical data from being misused, techniques to reliably preserve sensitive information should be developed for real world applications. In this paper, we propose a framework called anonymized generative adversarial networks (AnomiGAN), to preserve the privacy of personal medical data, while also maintaining high prediction performance. We compared our method to state-of-the-art techniques and observed that our method preserves the same level of privacy as differential privacy (DP) and provides better prediction results. We also observed that there is a trade-off between privacy and prediction results that depends on the degree of preservation of the original data. Here, we provide a mathematical overview of our proposed model and demonstrate its validation using UCI machine learning repository datasets in order to highlight its utility in practice. The code is available at https://github.com/hobae/AnomiGAN/

Frequency of ClinVar pathogenic variants in chronic kidney disease patients surveyed for return of research results at a Cleveland public hospital

Dana C. Crawford[1,2,3,], John Lin[1], Jessica N. Cooke Bailey[1,2], Tyler Kinzy[1], John R. Sedor[4,5], John F. O'Toole[5], William S. Bush[1,2,3]

[1]Cleveland Institute for Computational Biology, [2] Departments of Population and Quantitative Health Sciences, and [3]Genetics and Genome Sciences, Case Western Reserve University [4]Department of Physiology and Biophysics, Case Western Reserve University; and [5]Department of Nephrology and Hypertension, Glickman Urology and Kidney and Lerner Research Institute, Cleveland Clinic

Return of results is not common in research settings as standards are not yet in place for what to return, how to return, and to whom. As a pioneer of large-scale of return of research results, the Precision Medicine Initiative Cohort now known of All of Us plans to return pharmacogenomic results and variants of clinical significance to its participants starting late 2019. To better understand the local landscape of possibilities regarding return of research results, we assessed the frequency of pathogenic variants and APOL1 renal risk variants in a small diverse cohort of chronic kidney disease patients (CKD) ascertained from a public hospital in Cleveland, Ohio genotyped on the Illumina Infinium MegaEX. Of the 23,720 ClinVar-designated variants directly assayed by the MegaEX, 8,355 (35%) had at least one alternate allele in the 130 participants genotyped. Of these, 18 ClinVar variants deemed pathogenic by multiple submitters with no conflicts in interpretation were distributed across 27 participants. The majority of these pathogenic ClinVar variants (14/18) were associated with autosomal recessive disorders. Of note were four African American carriers of TTR rs76992529 associated with amyloidogenic transthyretin amyloidosis, otherwise known as familial transthyretin amyloidosis (FTA). FTA, an autosomal dominant disorder with variable penetrance, is more common among African-descent populations compared with European-descent populations. Also common in this CKD population were APOL1 renal risk alleles G1 (rs73885319) and G2 (rs71785313) with 60% of the study population carrying at least one renal risk allele. Both pathogenic ClinVar variants and APOL1 renal risk alleles were distributed among participants who wanted actionable genetic results returned, wanted genetic results returned regardless of actionability, and wanted no results returned. Results from this local genetic study highlight challenges in which variants to report, how to interpret them, and the participant's potential for follow-up, only some of the challenges in return of research results likely facing larger studies such as All of Us.

Network-Based Matching of Patients and Targeted Therapies for Precision Oncology

Qingzhi Liu[1], Min Jin Ha[2], Rupam Bhattacharyya[1], Lana Garmire[3], Veerabhadran Baladandayuthapani[1]

[1]Department of Biostatistics, University of Michigan; [2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center; [3]Department of Computational Medicine and Bioinformatics University of Michigan

The extensive acquisition of high-throughput molecular profiling data across model systems (human tumors and cancer cell lines) and drug sensitivity data, makes precision oncology possible – allowing clinicians to match the right drug to the right patient. Current supervised models for drug sensitivity prediction, often use cell lines as exemplars of patient tumors and for model training. However, these models are limited in their ability to accurately predict drug sensitivity of individual cancer patients to a large set of drugs, given the paucity of patient drug sensitivity data used for testing and high variability across different drugs. To address these challenges, we developed a multilayer network- based approach to impute individual patients' responses to a large set of drugs. This approach considers the triplet of patients, cell lines and drugs as one inter-connected holistic system. We first use the omics profiles to construct a patient-cell line network and determine best matching cell lines for patient tumors based on robust measures of network similarity. Subsequently, these results are used to impute the "missing link" between each individual patient and each drug, called Personalized Imputed Drug Sensitivity Score (PIDS-Score), which can be construed as a measure of the therapeutic potential of a drug or therapy. We applied our method to two subtypes of lung cancer patients, matched these patients with cancer cell lines derived from 19 tissue types based on their functional proteomics profiles, and computed their PIDS-Scores to 251 drugs and experimental compounds. We identified the best representative cell lines that conserve lung cancer biology and molecular targets. The PIDS-Score based top sensitive drugs for the entire patient cohort as well as individual patients are highly related to lung cancer in terms of their targets, and their PIDS-Scores are significantly associated with patient clinical outcomes. These findings provide evidence that our method is useful to narrow the scope of possible effective patient-drug matchings for implementing evidence-based personalized medicine strategies.

Phenome-wide association studies on cardiovascular health and fatty acids considering phenotype quality control practices for epidemiological data

Kristin Passero[1], Xi He[1], Jiayan Zhou[1], Bertram Mueller-Myhsok[2,3,4], Marcus E. Kleber[5], Winfried Maerz[5,6,7], Molly A. Hall[1]

[1]Penn State; [2]Max Planck Institute of Psychiatry; [3]Munich Cluster of Systems Biology; [4]University of Liverpool; [5]Heidelberg University; [6]SYNLAB Academy; [7]Medical University of Graz

Phenome-wide association studies (PheWAS) allow agnostic investigation of common genetic variants in relation to a variety of phenotypes but preserving the power of PheWAS requires careful phenotypic quality control (QC) procedures. While QC of genetic data is well-defined, no established QC practices exist for multi-phenotypic data. Manually imposing sample size restrictions, identifying variable types/distributions, and locating problems such as missing data or outliers is arduous in large, multivariate datasets. In this paper, we perform two PheWAS on epidemiological data and, utilizing the novel software CLARITE (CLeaning to Analysis: Reproducibility-based Interface for Traits and Exposures), showcase a transparent and replicable phenome QC pipeline which we believe is a necessity for the field. Using data from the Ludwigshafen Risk and Cardiovascular (LURIC) Health Study we ran two PheWAS, one on cardiac-related diseases and the other on polyunsaturated fatty acids levels. These phenotypes underwent a stringent quality control screen and were regressed on a genome-wide sample of single nucleotide polymorphisms (SNPs). Seven SNPs were significant in association with dihomo-γ-linolenic acid, of which five were within fatty acid desaturases FADS1 and FADS2. PheWAS is a useful tool to elucidate the genetic architecture of complex disease phenotypes within a single experimental framework. However, to reduce computational and multiple-comparisons burden, careful assessment of phenotype quality and removal of low-quality data is prudent. Herein we perform two PheWAS while applying a detailed phenotype QC process, for which we provide a replicable pipeline that is modifiable for application to other large datasets with heterogenous phenotypes. As investigation of complex traits continues beyond traditional genome wide association studies (GWAS), such QC considerations and tools such as CLARITE are crucial to the in the analysis of non-genetic big data such as clinical measurements, lifestyle habits, and polygenic traits.

aTEMPO: Pathway-Specific Temporal Anomalies for Precision Therapeutics

Christopher Michael Pietras, Liam Power, Donna K. Slonim

Tufts University

Dynamic processes are inherently important in disease, and identifying disease-related disruptions of normal dynamic processes can provide information about individual patients. We have previously characterized individuals' disease states via pathway-based anomalies in expression data, and we have identified disease-correlated disruption of predictable dynamic patterns by modeling a virtual time series in static data. Here we combine the two approaches, using an anomaly detection model and virtual time series to identify anomalous temporal processes in specific disease states. We demonstrate that this approach can informatively characterize individual patients, suggesting personalized therapeutic approaches.

Feature Selection and Dimension Reduction of Social Autism Data

Peter Washington[1], Kelley Marie Paskov[1], Haik Kalantarian[1], Nathaniel Stockham[1], Catalin Voss[1], Aaron Kline[1], Ritik Patnaik[2], Brianna Chrisman[1], Maya Varma[1], Qandeel Tariq[1], Kaitlyn Dunlap[1], Jessey Schwartz[1], Nick Haber[1], Dennis P. Wall[1]

[1]Stanford University, [2]Massachusetts Institute of Technology

Autism Spectrum Disorder (ASD) is a complex neuropsychiatric condition with a highly heterogeneous phenotype. Following the work of Duda et al., which uses a reduced feature set from the Social Responsiveness Scale, Second Edition (SRS) to distinguish ASD from ADHD, we performed item-level question selection on answers to the SRS to determine whether ASD can be distinguished from non-ASD using a similarly small subset of questions. To explore feature redundancies between the SRS questions, we performed filter, wrapper, and embedded feature selection analyses. To explore the linearity of the SRS-related ASD phenotype, we then compressed the 65-question SRS into low-dimension representations using PCA, t-SNE, and a denoising autoencoder. We measured the performance of a multi-layer perceptron (MLP) classifier with the top-ranking questions as input. Classification using only the top-rated question resulted in an AUC of over 92% for SRS-derived diagnoses and an AUC of over 83% for dataset-specific diagnoses. High redundancy of features have implications towards replacing the social behaviors that are targeted in behavioral diagnostics and interventions, where digital quantification of certain features may be obfuscated due to privacy concerns. We similarly evaluated the performance of an MLP classifier trained on the low-dimension representations of the SRS, finding that the denoising autoencoder achieved slightly higher performance than the PCA and t-SNE representations.

ATRIFICIAL INTELLIGENCE FOR ENHANCING CLINICAL MEDICINE

POSTER PRESENTATIONS

Prioritizing Copy Number Variants using Phenotype and Gene Functional Similarity

Azza Althagafi, Jun Chen, Robert Hoehndorf

Computer, Electrical & Mathematical Science and Engineering Division (CEMSE), Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), 4700 KAUST, 23955-6900, Thuwal, Kingdom of Saudi Arabia

There are many types of genetic variation in the human genome, ranging from large chromosome anomalies to Single Nucleotide Variant (SNV). It is becoming necessary to develop methods for distinguishing disease-causing variants from a large number of neutral genetic variation in an individual. This problem is also relevant to Copy Number Variants (CNVs), which is a class of genetic variation where large segments of the genome differ in copy number amongst various individuals. Over the past several years, much progress has been made in the area of CNVs detection and understanding their role in human diseases. We now understand that CNVs account for much of human variability. Correspondingly, there have been several methods introduced to find disease-associated genes and SNVs. Different methods have been developed for predicting and prioritizing pathogenicity of SNVs found within a genome. Constructing similar methods for CNV is challenging due to the heterogeneity in variant size, type and the possibility of multiple genes being affected by large CNVs. CNV impact prediction methods should consider these factors in order to robustly prioritize pathogenic variants. We have built a method that incorporates biological background knowledge about the relation between phenotypes resulting from a loss of function in mouse genes, gene functions as described using the Gene Ontology (GO), as well as the anatomical site of gene expression along with a score that predicts the pathogenicity of CNV SVScore. We use this information to build a machine learning model that ranks CNVs based on their predicted pathogenicity and the relation between genes affected by the CNV and the phenotype we observe in affected individuals. Additionally, our approach considers several genomic features of each CNVs, such as the length of the coding sequence overlapping with the CNV, haploinsufficiency and triplosensitivity scores to measure the dosage-sensitivity for genes/regions, and GC content. Our results show that incorporating this information leads to improvement over a baseline model which uses only similarity scores between gene--phenotype associations and disease-associated phenotypes, as well as improvement over using only pathogenicity prediction methods for CNVs. Our method achieves an F-score of 80.85%, with 82.05% precision and 79.67% recall in our evaluation set. The results demonstrate that incorporating phenotype, functional, and gene expression information may be utilized to identify causative CNVs. Future work is required to evaluate and improve our model using patient-derived WGS data.

Inferring the Reward Functions that Guide Cancer Progression

John Kalantari[1], Heidi Nelson[2], Nicholas Chia[3]

[1]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA; [2]Colon and Rectal Surgery, Mayo Clinic, Rochester, MN, USA; [3]Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, MN, USA

Cancer can occur in patients with different genetic backgrounds via a multi-step evolutionary process, i.e., driven by modification and selection, that can accumulate different genetic alterations. Despite these differences, many cancer subtypes are unified by similar mechanisms or types of genetic changes. In other words, there are multiple etiological paths tied together by specific events that share commonality in their causal mechanism. Understanding these common mechanisms will enable the development of better therapies and preventative measures. It will also enable improved prediction of recurrence and metastatic advancement of cancer, directly impacting the 606,880 annual cancer deaths in the United States alone. Our work is built upon the central proposition that the Markov Decision Process (MDP) can better represent the process by which cancer arises and progresses. More specifically, by encoding a cancer cell's complex behavior as a MDP, we seek to model the series of genetic changes, or evolutionary trajectory, that leads to cancer as an optimal decision process. We posit that using an Inverse Reinforcement Learning (IRL) approach will enable us to reverse engineer an optimal policy and reward function based on a set of expert demonstrations extracted from the DNA of patient tumors. The inferred reward function and optimal policy can subsequently be used to extrapolate the evolutionary trajectory of any tumor. We introduce a novel data-agnostic artificial intelligence framework which can infer reward functions describing the causal mechanisms that best explain the observed behavior of an 'optimally-behaving agent'–the cancer cell. Using multi-omic data from 27 colorectal cancer (CRC) patients as proof-of-principle, we show that IRL provides a systematic and scalable approach to formally stating and solving the problem of cancer evolution. By providing a lineage path (i.e., sequences of alterations) obtained via subclonal reconstruction for each tumor, we are able to reduce this complex problem to the recovery of an associated reinforcement learning reward function. These reward functions have the potential to model unknown molecular mechanisms driving intratumor heterogeneity and to elucidate cancer etiologies.

Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach

Mohamad Koohi-Moghadam, Haibo Wang, Yuchuan Wang, Xinming Yang, Hongyan Li, **Junwen Wang**, Hongzhe Sun

Department of Chemistry, The University of Hong Kong, Hong Kong, China;
Department of Health Sciences, Mayo Clinic, Scottsdale, AZ, USA;
Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Scottsdale, AZ, USA;
Center for Individualized Medicine,Mayo Clinic, Scottsdale, AZ, USA;
College of Health Solutions, Arizona State University, Scottsdale, AZ, USA

Metalloproteins play important roles in many biological processes. Mutations at the metal-binding sites may functionally disrupt metalloproteins, initiating severe diseases; however, there seemed to be no effective approach to predict such mutations until now. Here we develop a deep learning approach to successfully predict disease-associated mutations that occur at the metal-binding sites of metalloproteins. We generate energy-based affinity grid maps and physiochemical features of the metal binding pockets (obtained from different databases as spatial and sequential features) and subsequently implement these features into a multichannel convolutional neural network. After training the model, the network can successfully predict disease-associated mutations that occur at the first and second coordination spheres of zinc-binding sites with an area under the curve of 0.90 and an accuracy of 0.82. Our approach stands for the first deep learning approach for the prediction of disease-associated metal-relevant site mutations in metalloproteins, providing a new platform to tackle human diseases.

GENERAL

POSTER PRESENTATIONS

Ranking RAS pathway mutations using evolutionary history of MEK1

Katia Andrianova, Igor Jouline

Ohio State University, Department of Microbiology, Columbus, Ohio 43210

The Ras/MAPK (rat sarcoma/mitogen-activated protein kinase) signaling pathway is involved in essentially all aspects of organismal development, from the first cell divisions in the early embryo to postnatal development and growth. Given its critical function, it is not surprising that deregulated Ras/MAPK signaling, resulting from either genetic or environmental perturbations, can lead to cancer and developmental abnormalities. A large class of such abnormalities, known as RASopathies, is associated with activating germ-line mutations in many components of the Ras pathway. Over the past decade when next generation sequencing (NGS) has become valuable and cost-effective tool for research applications and clinical diagnostics of Mendelian diseases, simultaneous sequencing of multiple genes in MAPK signaling pathways have yielded many reports with hundreds of mutations possibly associated with RASopathies and cancer. In particular, multiple new mutations were identified in MEK1 kinase. The majority of newly discovered coding variations neither have been described in other individuals nor have been studied or functionally analyzed in cellular or animal models, thus leaving clinicians to rely on in silico predictions of the "variants of uncertain significance" consequences with computational software, such as PolyPhen and SIFT. Automated sequence searches used in these methods do not distinguish possible duplication events in the genes' histories, hence multiple sequence alignment (MSA) sets usually include both ortholog and paralog copies. As purifying selection treads on one of the duplicate copy it can become associated with a different phenotype compared to its paralogous sibling and/or to the parental gene. In most cases of Mendelian diseases only one specific duplicate of the gene in the human genome results to be associated with a disease. This indicates the importance of considering both common ancestors and any gene's duplication history for the variants interpretation. The presence of seven human MEK proteins increases the chances of including paralogs into the analysis, and therefore, substantially limits mutation interpretation. In this study we established the first precise description of an evolutionary history of MEK kinases and identified potential duplication events. We determined that MEK1 is an ancestor of the entire MEK family. In depth analysis of the orthologous proteins showed that essentially all experimentally proven pathogenic mutations were predicted as "damaging" by our approach. By comparing our results with the predictions made by PolyPhen-2 and SIFT we showed how careful analysis of an evolutionary history of a gene may improve accuracy of missense mutations outcomes prediction.

Integrative Analysis of COPD and Lung Cancer Metadata Reveals Shared Alterations in Immune Response, PTEN and PI3K-AKT Pathways}

Dannielle Skander[1], Arda Durmaz[1], Mohammed Orloff[2], **Gurkan Bebek**[1]

[1]Case Western Reserve University, [2]University of Arkansas for Medical Sciences

Chronic obstructive pulmonary disease (COPD) and lung cancer are among the leading causes of death worldwide. While it is believed the two diseases are related, the mechanisms behind this relationship remain unclear. We investigate the relationship between COPD and lung cancer using an integrative -omics approach. Integration of epigenetic and mRNA gene expression data allows us to discover the functionally relevant genes, i.e., the genes crucial for disease development. Using this approach, our study suggests that the mechanisms driving the development of both diseases are related to the interleukin immune response (IL4 and IL17), PTEN and PI3K-AKT pathways. Understanding this relationship between COPD and lung cancer is crucial for future prevention and treatment options of both COPD and lung cancer.

Investigating sources of irreproducibility in analysis of gene expression data

Carly A. Bobak, Jane E. Hill

Dartmouth College

The use of big data promises to change the landscape of biomedical research; however, irreproducibility of results remains a problem. In this work, we set out to investigate proposed methods to increase reproducibility of gene expression results. Specifically, we test the following three hypotheses: Results from pathway enrichment will be more similar across datasets than results on differentially expressed (DE) genes Similarity across smaller datasets will be lower than similarity in larger datasets Results from multi-cohort data will be more similar than results from single cohort data We selected three unique datasets from the Gene Expression Omnibus that include active TB patients, spanning pediatric and adult patients. In each dataset we ranked DE genes as they were associated with TB vs other (healthy controls, other diseases, or latent tuberculosis infection). We then calculated the rank biased overlap (RBO) of the ranked genes across each dataset. RBO is a similarity measure scaled between 0 and 1 and can be interpreted as the average agreement between two lists. Gene set enrichment analysis (GSEA) was performed, and we calculated a rank for the pathway hits and compared RBO for associated pathways between datasets. On average, the RBO increased by a fold change of 1.83×10^4 when comparing similarity of associated pathways to similarity of DE genes. We then divided each dataset in half and repeated the analysis on all sub-datasets. Sub-datasets from the same parent dataset had similar results (mean RBO of 0.60, sd=0.24) as opposed to subsets from a different parent dataset (mean=0.10, sd=0.15). Contradicting our original hypothesis, overall RBO calculated between subsets from different parent datasets did not necessarily decrease compared to the initial RBO calculation – in fact, half of the RBO comparisons increased in the sub-datasets compared to using the whole datasets. To test the final hypothesis, we co-normalized, merged, and then randomly divided datasets into three approximately equal pieces. We repeated the DE analysis on each piece of the merged dataset. Across mixed datasets, the mean RBO was 0.023 (sd=0.43). Heterogeneous datasets were more alike than unique datasets, but less alike than a single divided dataset. However, the RBOs from mixed datasets compared to original datasets were not statistically significantly different from the RBOs comparing results from the original datasets. Thus, we demonstrated that associated pathways are greatly more reproducible than associated genes. Further study is necessary to investigate the conditions under which statistical power and heterogeneity of data influence reproducibility of findings from gene expression studies.

Ethereum and MultiChain blockchains as secure tools for individualized medicine

Charlotte Brannon, Gamze Gursoy, Sarah Wagner, Mark Gerstein

Yale University Computational Biology and Bioinformatics Program

With the rapidly decreasing cost of genome sequencing and advent of individualized medicine, reliance on individual genomic data will soon be integral to medical treatment decisions. For example, a patient's personal genomic sequence will provide physicians with information on which to base tests and diagnoses. Similarly, pharmacogenomics data will reveal the most effective prescriptions for a particular patient. Genomic data will need to be shared efficiently among multiple parties. However, because these are sensitive personal data which will directly impact medical treatment decisions, they must be maintained in a secure, high-integrity fashion. Blockchain technology is one way to achieve secure, high-integrity data storage. We present two proof-of-concept solutions, one for storing and querying personal genomic sequence data in a MultiChain blockchain designed for direct sharing with physicians; and one for storing and querying gene-drug interaction data in an Ethereum blockchain smart contract designed for shared access among permissioned researchers and physicians. Despite the high security and integrity that comes with blockchain data storage, there is a trade-off with data access efficiency and storage costs. We overcome these challenges by developing novel storage techniques. When storing personal genomic sequence data, we do not store the actual sequence data but rather a set of meta-data which can be used in combination with a reference genome to reconstruct the original sequences. When storing pharmacogenomics data, we use an index-based, multi-mapping approach to provide time- and space- efficient insertion and querying.

Genomic predictors of L-asparaginase-induced pancreatitis in pediatric cancer patients

Britt I. Drögemöller, Galen E. B. Wright, Shahrad R. Rassekh, Shinya Ito, Bruce C. Carleton, Colin J. D. Ross, The Canadian Pharmacogenomics Network for Drug Safety Consortium

Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, BC, Canada; BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC, Canada; Department of Pediatrics, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada; Clinical Pharmacology and Toxicology, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada; Pharmaceutical Outcomes Programme, BC Children's Hospital, Vancouver, BC, Canada

Background: L-asparaginase is highly effective in the treatment of pediatric acute lymphoblastic leukemia. Unfortunately, the use of this treatment is limited by the occurrence of pancreatitis, a severe and potentially lethal adverse drug reaction, which occurs in 2-18% of patients. As previous studies have been unable to identify strong associations between clinical variables and susceptibility to L-asparaginase-induced pancreatitis, genetic factors are expected to play an important role this adverse drug reaction. Objectives: We sought to explore the role of these genetic susceptibility factors to L-asparaginase-induced pancreatitis in pediatric cancer patients. Methods: Patients who were treated with L-asparaginase were recruited from 13 pediatric oncology units across Canada (n=284) and extensive clinical data were collected for all patients. Genotyping was performed using the Illumina HumanOmniExpress and Global Screening Arrays and pancreatic gene expression profiles were imputed in these individuals using GTEx v7 and S-PrediXcan. Genome- and transcriptome-wide associations (GWAS and TWAS) were performed to identify associations with L-asparaginase-induced pancreatitis. Results: GWAS analyses identified significant associations between genetic variants in HLA-DQA1 and –DRB1 and pancreatitis, while TWAS revealed that individuals experiencing L-asparaginase-induced pancreatitis exhibited lower expression levels of HLA-DRB5. Further interrogation of the TWAS data revealed an enrichment in genes involved in the somatic diversification of immune receptors. Conclusions: These analyses uncovered an association between genetic variation in immune-related genes and the development of L-asparaginase-induced pancreatitis. These associations mirror previous associations with the HLA region and (i) pancreatitis induced by other drugs and (ii) L-asparaginase-induced hypersensitivity.

NITECAP: A novel method and interface for the identification of circadian behavior in highly parallel time-course data

Thomas G. Brooks[1], Cris W. Lawrence[1], Nicholas F. Lahens[1], Soumyashant Nayak[1], Dimitra Sarantopoulou[1], Garret A. FitzGerald[1,2], Gregory R. Grant[3]

[1]Institute for Translational Medicine and Therapeutics (ITMAT), University of Pennsylvania; [2]Systems Pharmacology and Translational Therapeutics; [3]Department of Genetics, University of Pennsylvania

We introduce a new tool called NITECAP for the task of identifying circadian behavior in massively parallel measurements of biological entities; for example, finding circadian genes from gene expression time course data measured by RNA-Seq or microarrays. NITECAP employs a permutation-based approach which uses a novel statistic designed to be sensitive to circadian behavior. NITECAP also uses an approach to multiple-testing which produces q-values directly without needing to first generate p-values which then need to be adjusted. Our approach has several advantages particularly when individual p-values are underpowered or unreliable. Importantly, we have developed an intuitive user-friendly web-based interface which enables investigators to perform robust circadian analyses of this type directly without expert informatics support. Users can quickly scroll through time course profiles sorted by effect size, greatly facilitating the choice of significance thresholds that currently require making blind choices of numerical cutoffs. Putting this type of analysis in the hands of the investigators can significantly streamline their research. The web site also enables the other standard significance tests such as JTK and ANOVA and provides tools to perform comparative studies, such as finding phase or amplitude differences between different conditions. NITECAP is freely available for public use at: http://www.nitecap.org

The Interplay of Obesity and Race/Ethnicity on Major Perinatal Complications

Yaadira Brown, MPH[1]; Olubode A. Olufajo, MD, MPH[2]; Edward E. Cornwell III, MD[2]; William Southerland, PhD[3]

[1]Research Centers in Minority Institutions: Howard University, Howard University College of Medicine; [2]Research Centers in Minority Institutions: Howard University, Clive Callender Howard-Harvard Health Sciences Outcomes Research Center; [3]Research Centers in Minority Institutions: Howard University

Background: It has been established that a significant disparity exists in the rates of adverse perinatal outcomes across different racial/ethnic groups, with non-Hispanic Black women generally being most impacted. There is also evidence that obesity is associated with adverse perinatal outcomes. Although some studies have examined the impact of race/ethnicity and obesity on adverse perinatal outcomes, most studies have done so using local or statewide data. This study aims to use a national sample to determine the role of obesity in the racial/ethnic disparities seen in adverse perinatal outcomes in the United States. Methods: Data from the National Inpatient Sample was utilized in selecting pregnant women admitted for delivery between 2010 and 2014. Demographics (race/ethnicity, insurance type, household income, co-morbidities) and hospital characteristics were extracted. Race/ethnicity was categorized as Non-Hispanic Whites (NHW), Non-Hispanic Blacks (NHB), and Hispanics. Outcomes of interest were gestational diabetes, pre-eclampsia, pre-term birth, and hospital mortality. Multivariate logistic regressions were performed to determine the independent predictors of the outcomes, using two sets of models; one which included obesity as a variable in the model and one which did not. The differences between the two sets of models were compared by performing the Wald Test. Results: Our cohort consisted of 15,561,942 pregnant individuals admitted for delivery. There were 9,247,729 (59.43%) NHW, 2,552,569 (16.4%) NHB, and 3,761,644 (24.17%) Hispanic. Compared to other groups, NHB had significantly higher rates of pre-eclampsia (5.1%), pre-term birth (9.4%), and hospital mortality (.11%). They also had the highest rates of obesity (9.0%). On multivariate analysis, NHB were more likely to have pre-eclampsia (Adjusted Odds Ratio [aOR] 1.26; 95% Confidence Interval [CI] 1.23-1.29), pre-term birth (aOR 1.38; 95% CI 1.34-1.41), and hospital mortality (aOR 2.05; 95% CI 1.2-3.38) when compared to NHW. However, they had a similar risk for gestational diabetes (aOR 0.94; 95% CI 0.91-0.96) as NHW. Obesity was significantly associated with gestational diabetes (aOR 3.08; 95% CI 3.02-3.15), pre-eclampsia (aOR 2.14; 95% CI 2.09-2.19), and pre-term birth (aOR 1.04; 95% CI 1.01-1.06). Although the differences were minimal, the regression models that included obesity as a variable better predicted the outcomes than those that did not when assessing gestational diabetes, pre-eclampsia, and pre-term birth. Conclusion: These findings further confirm that racial/ethnic disparities exist amongst adverse perinatal outcomes, with NHB being disproportionately affected. They also suggest that obesity plays a significant role in the racial/ethnic disparities that do exist for the adverse perinatal outcomes measured, other than hospital mortality. These data suggest that addressing obesity in the population may be beneficial in improving perinatal outcomes, but they also suggest that more research is needed to identify the major factors that drive the racial/ethnic disparities that exist amongst perinatal outcomes in the United States.

A Comparison of Pharmacogenomic Information in FDA-Approved Drug Labels and CPIC Guidelines

Katherine I. Carrillo[1], Teri E. Klein[2]

[1]Henry M. Gunn High School, Palo Alto, CA; [2]Stanford University, Stanford, CA

Pharmacogenomics (PGx) is useful in helping to predict a patient's likely reaction to a medication based on their genotype, allowing for personalized medicine. The FDA maintains a "Table of Pharmacogenomic Biomarkers in Drug Labeling" (https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling) consisting of pharmacogenomic information found in the drug labeling. However, many labels on the list do not contain advice for a clinician about how or when to use a patient's genetic information. Guidelines created by the Clinical Pharmacogenetic Implementation Consortium (CPIC; https://cpicpgx.org/) contain information about how to use patient genetic information when prescribing drugs. Also, CPIC provides guidelines for some drugs not currently on the FDA biomarker list, though it does not provide guidelines for every drug on the biomarker list. Using PharmGKB annotated FDA-approved labels (through October 2019), we evaluated label information to determine (1) which labels contained any kind of prescribing information including a suggested alternate drug, dosing information or special considerations based on the patient's genotype/metabolizer status, (2) which PharmGKB annotated labels were present on the FDA biomarker list, and (3) what genes were involved. We did not include FDA labels annotated for genetic variation in cancer cells; only germline variation was included. We compared all available CPIC guideline recommendations to the information from the labels. We identified where the labels and guidelines are similar or not. PharmGKB has 223 annotations (not including 82 annotations for cancer cell DNA variation) based on 219 FDA-approved drug labels. Of these, 199 labels are currently on the biomarker list and 17 were on the biomarker list at one time but have been removed by the FDA. Twenty labels have dosing information and 35 recommend an alternate drug based on genotype/metabolizer phenotype. Another 34 labels have some other special consideration, but most labels on the biomarker list (136) have no guidance for clinicians about what to do about the biomarker, if anything. There are 45 drugs with published CPIC guidelines (https://cpicpgx.org/genes-drugs/). Thirty-six of the drugs have a label on the FDA biomarker list but the information on the label does not always match the guideline. Only 21 of the CPIC drugs have labels with guidance. For some drugs, the PGx information on the labels is similar to the CPIC guidelines but different for many others. The FDA biomarker list has more drugs than CPIC guidelines written and in some cases the labels tell clinicians when they should test a patient while CPIC doesn't talk about testing. However, for most drugs, the labels don't give the clinicians a lot of information about what to do with their patients' genetic test results. For the drugs with CPIC guidelines, there is more information about how to use genetic test results and why. Funded by NIH/NIGMS R24 GM61374.

xTEA: a transposable element insertion analyzer for genome sequencing data from multiple technologies

Chong Chu[1], Rebeca Monroy[2], Soohyun Lee[1], **E. Alice Lee**[2], Peter J. Park[1]

[1]Harvard Medical School, [2]Boston Children's Hospital

Transposable elements (TEs) comprise nearly 50% of the human genome. Although most of the TEs are now silent, several types of retrotransposons including LINE-1, Alu, and SVA are still active. Somatic TE insertions have been shown to occur frequently in multiple tumor types [1,2] and at a low rate in neurons of phenotypically normal individuals [3]. Multiple tools have been developed to call TE insertions from genome sequencing data, but an efficient tool that can identify both germline and somatic TE insertions with high sensitivity and specificity is still lacking. Moreover, newer technologies such as 10X Linked-Read and PacBio or Nanopore long read sequencing provide an unprecedented opportunity to study TEs; however, current methods do not take advantage of these data types. Here, we present a new computational tool xTEA, building on our previous algorithm TEA [1]. This tool identifies TE insertions from Illumina paired-end reads, 10X Linked-Reads, long reads, or a combined dataset. xTEA outperforms MELT [4] and Traffic-mem [5] on normal and tumor Illumina data, respectively. A comparison of different sequencing platforms reveals that the analysis of long reads had greater sensitivity and specificity, especially in repetitive regions. Both 10X Linked-Reads and long reads demonstrated clear advantages over short reads in constructing full length TE insertions. Better performance was achieved on hybrid data compared to single platform data. Using 22 human samples with either PacBio or Nanopore long reads and matched short reads, we uncovered LINE-1 internal SV hotspots and SVA internal VNTR expansion. xTEA is a comprehensive cross-platform TE insertion-calling tool. It can be deployed on a computing cluster, AWS, and Google Cloud, and is efficient for large cohort analysis. xTEA is publicly available at https://github.com/parklab/xTEA. References [1] Lee, Eunjung, et al. "Landscape of somatic retrotransposition in human cancers." Science 337.6097 (2012): 967-971. [2] Rodriguez-Martin, Bernardo, et al. "Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours." BioRxiv (2017): 179705. [3] Evrony, Gilad D., et al. "Cell lineage analysis in human brain using endogenous retroelements." Neuron 85.1 (2015): 49-59. [4] Gardner, Eugene J., et al. "The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology." Genome research 27.11 (2017): 1916-1929. [5] Tubio, Jose MC, et al. "Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes." Science 345.6196 (2014): 1251343.

Go Get Data (GGD): simple, reproducible access to scientific data

Michael Cormier[1], Jon Belyeu[1], Brent Pedersen[1], Joe Brown[1], Johannes Koster[2], **Aaron R. Quinlan**[1]

[1]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA; [2]Algorithms for reproducible bioinformatics, Institute of Human Genetics, University of Duisburg-Essen, Essen, NRW, Germany

Genomics research is complicated by the difficulty of identifying, collecting, and integrating the numerous datasets and annotations germane to our experiments. Furthermore, these data exist in disparate sources, and are stored in diverse, often abused formats pertaining to different genome builds. These complexities waste time, inhibit reproducibility, and curtail research creativity. Inspired by the success of software package managers, we have developed Go Get Data (GGD; https://gogetdata.github.io/) as a fast, reproducible approach to install standardized packages of data and annotations for genomics research.

Global epigenomic regulation of gene expression and cellular proliferation in T-cell leukemia

Sinisa Dovat, Yali Ding, Bo Zhang, Jonathon L. Payne, Feng Yue

Pennsylvania State University College of Medicine, Hershey, PA, USA

Ikaros encodes a DNA-binding protein that functions as a tumor suppressor in T-cell acute lymphoblastic leukemia (T-ALL). Deletion and/or functional inactivation of Ikaros results in the development of high-risk leukemia. The mechanisms through which Ikaros regulates gene expression and tumor suppression in T-ALL are unknown. Ikaros haplo-knockout mice develop T-ALL with 100% penetrance with arrest of T-cell differentiation. During the process of malignant transformation to T-ALL, Ikaros haploinsufficient thymocytes lose their remaining wildtype Ikaros allele. Re-introduction of Ikaros into Ikaros-null T-ALL cells results in cessation of cellular proliferation and induction of T-cell differentiation. Thus, this is an optimal system for studying Ikaros tumor suppressor function because it captures the role of Ikaros in the transition from a malignant state (Ikaros-null T-ALL) to a non-malignant state (following Ikaros re-introduction). We used ATAC-seq and ChIP-seq of H3K4me1, H3K4me3, H3K27ac, and Ikaros to perform dynamic, global epigenomic and gene expression analyses at several time points in Ikaros-null T-ALL and following Ikaros re-introduction in order to determine the mechanisms of Ikaros' tumor suppressor activity. Expression analysis identified a large number of novel signaling pathways that are directly regulated by Ikaros and Ikaros-induced enhancers, and that are responsible for the cessation of proliferation and induction of T-cell differentiation in T-ALL cells. Epigenomic analysis identified novel Ikaros functions in the epigenetic regulation of gene expression: Ikaros directly regulates de novo formation and depletion of enhancers; de novo formation of active enhancers and activation of poised enhancers; and Ikaros directly induces the formation of super-enhancers. Global analysis of chromatin accessibility revealed that Ikaros binding resulted in the opening of over 3400 previously-inaccessible chromatin sites. This is accompanied by de novo enrichment of H3K4me1 and H3K4me3 modifications and formation of de novo enhancers and promoters. These data demonstrate that Ikaros has pioneer activity and triggers coordinated regulation of gene expression. Ikaros pioneering activity was further determined by direct binding of ikaros to reconstituted nucleosomes by electromobility shift assay. Dynamic analyses demonstrate the long-lasting effects of Ikaros' DNA binding on enhancer activation, de novo formation of enhancers and super-enhancers, and chromatin accessibility. In conclusion, our results establish that Ikaros' tumor suppressor function occurs via global regulation of the enhancer and super-enhancer landscape, along with regulation of chromatin accessibility, and identified novel tumor suppressor regulatory pathways in T-ALL.

A pharmacogenomic investigation of the cardiac safety profile of ondansetron in children and in pregnant women

Galen E. B. Wright, Britt I. Drögemöller, Jessica Trueman, Kaitlyn Shaw, Michelle Staub, Shahnaz Chaudhry, Sholeh Ghayoori, Fudan Miao, Michelle Higginson, Gabriella S.S. Groeneweg, James Brown, Laura A. Magee, Simon D. Whyte, Nicholas West, Sonia Brodie, Geert 'tJong, Howard Berger, Shinya Ito, Shahrad R. Rassekh, Shubhayan Sanatani, Colin J. D. Ross, Bruce C. Carleton

British Columbia Children's Hospital Research Institute, Vancouver, British Columbia, Canada; Pharmaceutical Outcomes Programme, British Columbia Children's Hospital, Vancouver, British Columbia, Canada; Division of Translational Therapeutics, Department of Pediatrics, University of British Columbia, Vancouver, British Columbia, Canada; Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada; Clinical Research Unit, Children's Hospital Research Institute of Manitoba, Winnipeg, Manitoba, Canada; Division of Clinical Pharmacology and Toxicology, The Hospital for Sick Children, Toronto, Ontario, Canada; British Columbia Women's Hospital and Health Centre, Vancouver, British Columbia, Canada; Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia, Vancouver, British Columbia, Canada; School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College, London, United Kingdom; Department of Pediatric Anesthesia, British Columbia Children's Hospital, Vancouver, British Columbia, Canada; Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; Department of Obstetrics and Gynecology, St. Michael's Hospital, Toronto, Ontario, Canada; Epi Methods Consulting, Toronto, Ontario, Canada; Division of Cardiology, Department of Pediatrics, Children's Heart Centre, BC Children's Hospital, University of British Columbia, Vancouver, Canada

Background: 5-HT3 receptor antagonists, such as ondansetron, are highly effective medications for the treatment of nausea and vomiting. However, these medications are also associated with prolongation of the QT interval, placing patients at risk of cardiac adverse events. Pharmacogenomic information for therapeutic response to ondansetron exists, particularly pertaining to CYP2D6, but no study has been performed on genetic factors that influence the cardiac safety of this medication. Objectives: Determine ondansetron-induced cardiac electrophysiological changes in three unique patient cohorts and identify pharmacogenomic predictors of QT interval prolongation. Methods: Three patient groups receiving ondansetron for the prevention of nausea and vomiting were recruited and followed prospectively (pediatric post-surgical patients n=101; pediatric oncology patients n=98; pregnant women n=62). Electrocardiograms were conducted at baseline and post-ondansetron administration. Pharmacogenomic associations were then assessed via analyses of comprehensive CYP2D6 genotyping data and genome-wide association analyses. Results: In the entire cohort, 62 patients (24.1%) were defined as cases based on Bazett-corrected QTc values. The most significant shift from baseline occurred at five minutes post-administration (P=9.8 x10-4). Genome-wide analyses identified novel candidate genes for this drug-induced phenotype. The two most significant associations were observed for a missense variant in TLR3 (rs3775291; P=2.00 x 10-7) and an eQTL for SLC36A1 (rs34124313; P=1.97 x 10-7). These genes are implicated in serotonin- and QT-related traits and therefore likely represent biologically relevant findings. CYP2D6 activity score was not associated with case-control status. Conclusions: The results of this study provide the first step towards understanding the genomic basis of cardiac changes occurring after ondansetron use in children and pregnant women, with the overall goal to improve the safety of these commonly used antiemetic medications.

TREND: a platform for exploring protein function in prokaryotes using phylogenetics, domain architectures, and gene neighborhoods information.

Vadim M. Gumerov, Igor B. Zhulin

The Ohio State University

Key steps in a computational study of protein function involve analysis of (i) relationships between homologous proteins, (ii) protein domain architecture, and (iii) gene neighborhoods the corresponding proteins are encoded in. Each of these steps requires a separate computational task and sets of tools. Combining the results into a complete analysis is usually done by hand, which is time-consuming and error-prone. Here we present a new platform, TREND (tree-based exploration of neighborhoods and domains), which can perform all the necessary steps in automated fashion and put the derived information into phylogenomic context, thus making evolutionary based protein function analysis more efficient. TREND is freely available at http://trend.zhulinlab.org. TREND consists of two pipelines: (1) Domains, which identifies protein domains, transmembrane regions and low-complexity segments, and maps this information on the phylogenetic tree, and (2) Neighborhoods, which identifies gene neighborhoods for the given set of protein sequences, clusters the genes based on shared domains of the encoded proteins, identifies operons and puts the derived data into phylogenomic context. Locally stored databases of the Pfam profile Hidden Markov models (HMMs) and CDD position-specific scoring matrices are used as a source of models for domains identification. Another source is a rich collection of signal-transduction specific profile HMMs derived from MiST database. The pipelines are highly customizable. On start, both pipelines first align provided proteins and build phylogenetic trees. These steps can be skipped if a researcher already has an alignment or a tree and would like to use them instead. Optionally redundancy of the sequences can be reduced. Instead of protein sequences, protein identifiers can be provided as input; corresponding sequences will be fetched from RefSeq and MiST databases. Results of the pipelines are presented as interactive pictures with cross-links to Pfam, CDD, RefSeq and MiST databases. All produced results can be downloaded for subsequent analysis.

TrackSigFreq: subclonal reconstructions based on mutation signatures and allele frequencies

Caitlin F. Harrigan[1,2,4], Yulia Rubanova[1,2,4], Quaid Morris[1,2,3,4,5,6], Alina Selega[2,4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada; [2]Donnelly Centre for Cellular and Biomolecular Research , University of Toronto, Toronto, Canada; [3]Department of Molecular Genetics, University of Toronto, Toronto, Canada; [4]Vector Institute, Toronto, Canada; [5]Ontario Institute for Cancer Research, Toronto, Canada; [6]Memorial Sloan Kettering Cancer Centre, New York, USA (pending)

Mutational signatures are patterns of mutation types, many of which are linked to known mutagenic processes. Signature activity represents the proportion of mutations a signature generates. In cancer, cells may gain advantageous phenotypes through mutation accumulation, causing rapid growth of that subpopulation within the tumour. The presence of many subclones can make cancers harder to treat and have other clinical implications. Reconstructing changes in signature activities can give insight into the evolution of cells within a tumour. Recently, we introduced a new method, TrackSig, to detect changes in signature activities across time from single bulk tumour sample. By design, TrackSig is unable to identify mutation populations with different frequencies but little to no difference in signature activity. Here we present an extension of this method, TrackSigFreq, which enables trajectory reconstruction based on both observed density of mutation frequencies and changes in mutational signature activities. TrackSigFreq preserves the advantages of TrackSig, namely optimal and rapid mutation clustering through segmentation, while extending it so that it can identify distinct mutation populations that share similar signature activities.

A Flexible Pipeline for the Prediction of Biomarkers Relevant to Drug Sensitivity

V. Keith Hughitt[1], Sayeh Gorjifard[1], Aleksandra M. Michalowski[1], John K. Simmons[2], Ryan Dale[1], Eric C. Polley[3], Jonathan J. Keats[4], Beverly A. Mock[1]

[1]NCI, [2]Personal Genome Diagnostics, [3]Mayo Clinic, Rochester, [4]TGen

Recent years have seen an explosion in the availability of paired molecular profiling and drug screen data, providing an unprecedented opportunity for the development of targeted therapies based on an individual's genetic background. Despite a number of recent successes in diseases ranging from cystic fibrosis to cancer, significant hurdles remain in our ability to accurately predict treatments based on molecular profiling data. In particular, few such tools exist that allow the integration of heterogeneous data types (e.g. genomic, transcriptomic, and somatic mutations), along with high-throughput drug screen data to make predictions about treatment efficacy. Here, we describe a generalized open-source pipeline developed for the analysis of precision medicine data, Pharmacogenomics Prediction Pipeline, or "P3". The modular design of P3 enables the inclusion of arbitrary input data types and the selection from multiple alternative machine learning algorithms, while automated statistical and visualization reporting steps incorporated throughout the pipeline assist in parameter tuning and early detection of problematic data elements. By incorporating external biological annotations from sources such as The Molecular Signatures Database (MSigDB), Drug Signatures Database (DSigDB), and DrugBank, P3 is able to detect important pathways correlated with drug sensitivity, while the inclusion of molecular profiling and clinical data from external patient and cell lines datasets allows P3 to focus its efforts on genes which are most likely to play a role in therapeutic response. To demonstrate the use of P3 for preclinical biomarker prediction, we applied P3 to an unpublished multiple myeloma dataset consisting of exome, RNA-Seq, and drug screen data for 1900 compounds across 45 tumor cell lines. Furthermore, gene expression and clinical data from 20 additional publically-available patient and cell line multiple myeloma datasets (>5,500 samples in total), along with data from the GDSC and CCLE drug sensitivity experiments were also analyzed, providing a rich source of information with respect to the biological relevance of putative biomarkers detected by the pipeline.

Creating a Metabolic Syndrome Research Resource (MetSRR)

Willysha Jenkins[1], Christian Richardson[2], ClarLynda Williams-DeVane PhD[1]

[1]Fisk University Nashville TN, [2]Duke University Durham NC

Metabolic syndrome (MetS) is a multifaceted syndrome. Risk factors include visceral adiposity, dyslipidemia, hyperglycemia, hypertension, and environmental factors. An established component of chronic disease sequela, MetS leads to an increased risk of cardiovascular disease and type 2 diabetes. MetS also leads to an increased risk of stroke. Comparative studies have identified heterogeneity in the pathology of MetS across groups, however, the etiology of these differences has yet to be elucidated. Despite the presence of public repositories of biological MetS-related data, the ability to access and work said data has its challenges. The process of querying databases, wrestling with software and wrangling data into workable formats prior to analysis is both cumbersome and time consuming. The Metabolic Syndrome Research Resource (MetSRR) is a curated database that provides access to MetS associated biological and ancillary data. It is an amalgamation of current and potential biomarkers of MetS extracted from relevant National Health and Nutrition Examination Survey (NHANES) data from 1999-2016. Each potential biomarker selection was driven by insights elucidated by the review of over 100 peer-reviewed articles. It includes 28 demographic, survey and known MetS related variables. There are 9 curated categorical variables and 42 potentially novel biomarkers. All measures are captured from over 90, 000 individuals. This biocuration effort will provide increased access to curated MetS related data. It will also serve as a hypothesis generation tool for disparate MetS etiology discovery, providing the ability to generate; and export ethnic group/race, sex, and age-specific curated datasets. MetSRR seeks to broaden participation in research efforts to identify clinically evaluative disparate MetS biomarkers. To the best of our knowledge, MetSRR is the only MetS specific database targeted at uncovering the disparate etiology of MetS through biocuration.

Utilizing cohort information to find causative variants

Senay Kafkas, Robert Hoehndorf

Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University Science and Technology, 4700 KAUST, Thuwal, 23955-6900 Saudi Arabia

Identification of causative variants in genomic data is challenging. Current studies focus on prioritizing variants within individual genomes, or apply statistical methods (e.g.GWAS) to large cohorts. With the rapid advancements and cost decrease in NGS, scientists are able to produce sequence data from large disease cohorts and healthy population. For example, UK Biobank makes available genotype to phenotype relations for >500,000 individuals and whole exome sequencing (WES) data for 50,000 individuals. Patients with the same/similar set of phenotypes may share the same/ biologically related genetic abnormalities and risk factors. The availability of these datasets may allow us to stratify individuals by their phenotype and use this information to identify causative variants within large cohorts. We propose a new method that stratifies patients by their phenotypes and identifies the set of causative variants which can explain phenotypes in most individuals within a cohort from WES/WGS. First, we generated and used synthetic disease cohorts to evaluate our method. We used the human genotype-phenotype associations from ClinVar and the sequence data from 1000 Genomes and generated synthetic cohorts with different population sizes for 200 randomly selected diseases from ClinVar. To generate a synthetic disease cohort of size N, first we picked randomly N individuals from 1000 Genomes and then for each individual, we picked randomly one of the variants of the given disease and added it to the genotype of the given individual. We pre-processed the sequence data by annotating with CADD and selecting only the most deleterious variant of a given gene for each individual. Furthermore, we “normalize” pathogenicity scores based on their frequencies within a population in order to account for different distribution within genes based on their length. We then apply our method on UK Biobank. We developed a method that identifies causative variants by utilizing information about shared phenotypes within a cohort and compared them against individually prioritizing variants using WES/WGS data and average gene ranks. Our approach relies on a machine learning model trained on a pathogenicity prediction score (e.g. CADD), the frequency of observing a pathogenicity score above a certain threshold in the same gene within a population, and uses this cohort and phenotype-derived information as feature to predict causative variants within individual genome sequences. Our method can identify causative variants in small and medium-sized cohorts (2 to 100 individuals). As the disease becomes more complex (i.e. involving harmful variants in multiple genes), our machine learning model improves over established methods in particular in larger cohorts (>80 individuals). Currently, we applied our method on UK Biobank and suggest candidate causative variants for 1499 complex diseases.

Integrated analysis of JAK-STAT pathway in homeostasis, simulated inflammation and tumour

Milica Krunic[1], Anzhelika Karjalainen[1], Mojoyinola Joanna Ola[1], Stephen Shoebridge[1], Sabine Macho-Maschler[1], Caroline Lassnig[1], Andrea Poelzl[1], Matthias Farlik[2], Nikolaus Fortelny[2], Christoph Bock[2], Birgit Strobl[1], Mathias Mueller[1]

[1]Institute of Animal Breeding and Genetics and Biomodels Austria University of Veterinary Medicine Vienna Austria; [2]CeMM – Center for Molecular Medicine Austrian Academy of Sciences Vienna Austria

Janus kinases (JAKs) and signal transducers and activators of transcription (STATs) play a key role in cytokine signalling and in the defence against infection and cancer. JAK-STAT signalling components interact with chromatin remodelling proteins and change chromatin architecture/landscape during cell differentiation and recognition and elimination of pathogens. Using different sequencing approaches (ATAC-Seq, ChIPmentation, single-cell RNA-Seq, RNA-Seq), our goal is to untangle the roles of JAK-STAT proteins in shaping chromatin landscapes of myeloid and lymphoid cells in homeostasis, sterile (simulated) inflammation and within tumour microenvironment. Additionally, we are investigating how evolutionary conserved STAT protein isoforms interact with chromatin and co-regulatory proteins to induce cell type- and gene-specific responses. The poster shows our summarised findings as a result of integration of different approaches.

BEERS 2: The Next Generation of RNA-Seq Simulator

Nicholas F. Lahens[1], Thomas G. Brooks[1], Dimitra Sarantopoulou[1], Soumyashant Nayak[1], Cris W. Lawrence[1], Anand Srinivasan[2], Jonathan Schug[3,4], Garret A. FitzGerald[1,5], John B. Hogenesch[6], Yoseph Barash[4], Gregory R. Grant[1,4]

[1]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [2]PMACS Enterprise Research Applications and High Performance Computing, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [3]Institute for Diabetes, Obesity, and Metabolism, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [4]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [5]Department of System Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [6]Division of Human Genetics, Department of Pediatrics, Center for Chronobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

The accurate interpretation of RNA-Seq data presents a moving target as scientists continue to introduce new experimental techniques and analysis algorithms. This challenge has led researchers to perform a substantial number of benchmarking studies in order to determine best analysis practices. Simulated datasets have proven to be an invaluable tool in these efforts. Despite this strong need for simulated data, only a few RNA-Seq simulators have been released in the public domain, and all of them are based on simplifying assumptions that limit their utility. To address these shortcomings and generate realistic simulated data we are developing the Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEERS) 2: an open-source, modular simulator that models each step in the process of converting RNA molecules into sequencing reads. We take an empirical approach to generating realistic RNA samples reflecting biological variability, alternative splicing, and allele-specific expression, which uses real data to train the parameters. Next, we model biochemical reactions and biases from each step in library construction as separate modules. Using an object-oriented paradigm, each module has well-defined inputs and outputs allowing users to easily substitute new modules. This design gives BEERS 2 the flexibility to model changes to library construction and sequencing protocols, evolving in parallel with sequencing technology. BEERS 2 is open source, freely available, and will be a crucial tool for the community as we continue to develop standards for transcriptome analysis.

Effect Modification by Age on a Diagnostic Three-Gene-Signature in Patients with Active Tuberculosis

Lauren McDonnell[1], Carly A. Bobak[1,2], Matthew Nemesure[1], Justin Lin[1], Jane E. Hill[1]

[1]Thayer School of Engineering at Dartmouth College, [2]Geisel School of Medicine at Dartmouth College

Introduction Tuberculosis (TB) is the leading cause of death from a single infectious agent worldwide (1). In 2017, there were 10 million reported cases of TB and another 1.3 million deaths from the disease (1). It is currently the leading killer for individuals who are HIV positive (1). In 2014, the WHO developed the ambitious Sustainable Development Goals (SDGs) which included "End TB", a major program aiming to eradicate the TB epidemic by 2030 (2). Accomplishing this will require more advanced diagnostics that are less invasive and determine the disease status more quickly and more reliably. In our analysis, we aim to model risk factors associated with the development of TB. Here, we are looking at demographic features from multi-cohort studies pulling data from thirty different countries from the Gene Expression Omnibus examining patients with active TB, latent TB, other diseases, and healthy controls. The data is pulled predominantly from developing countries, but also includes samples from developed countries, including the UK, France, Germany, and the United States. In total, the dataset includes 3,096 participants. Meta analysis of similar datasets have proposed a three-gene-score as a "global" tuberculosis metric (3). This type of analysis suggests that all active TB patients, regardless of other factors, will express this gene score. Our hypothesis is that this active TB will be additionally mediated by demographic factors such as age and HIV status that are associated with TB. Methodology We performed a multivariate logistic regression analysis to identify demographic features associated with culture-confirmed Tuberculosis. The model features included age, HIV status, and gene expressions for each gene individually (GBP5, DUSP3, and KLF2), as well as an interaction term for HIV and age with each of the three genes. Results The results of our multivariate logistic regression suggest that age modifies all three genes in the proposed global gene signatures (p-values of 5.38e-05, 6.75e-05, and, 0.01012, for GBP5, KLF2 and DUSP3 respectively). Initial findings also indicate that HIV status is a mediator of the effect of GBP5 (p-value of 0.03437). Knowing that the relationship between the gene expression of these three genes varies by demographics may change the way that a diagnostic is implemented in clinic. Our hope is that this analysis will be used to further refine the three-gene signature for specific demographic groups where it may be most effective in diagnosing active TB. Citations (1) WHO Global Tuberculosis Report 2018 www.who.int/tb/publications/global_report/en/ (2) Ending Tuberculosis by 2030: Can We Do It? A. B. Suthar, R. Zachariah, Harries https://www.ingentaconnect.com/contentone/iuatld/ijtld/2016/00000020/00000009/art00007?crawler=true (3) Genome-Wide Expression for Diagnosis of Pulmonary Tuberculosis: a Multicohort Analysis https://www.ncbi.nlm.nih.gov/pubmed/26907218

Classification and mutation prediction from gastrointestinal cancer histopathology images using deep learning

Sung Hak Lee[1], Hyun-Jong Jang[2]

[1]Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, [2]Department of Physiology, College of Medicine, The Catholic University of Korea

BACKGROUND: Although microscopic analysis of tissue slides has been the basis for disease diagnosis for decades, intra- and inter-observer variabilities remain issues to be resolved. The recent introduction of digital scanners has allowed for researchers to use deep learning in the analysis of tissue images because many H&E whole slide images (WSIs) are available. In the present study, we investigated the possibility of a deep learning-based, fully automated, computer-aided diagnosis system with WSIs from a gastric adenocarcinoma (STAD) dataset. In addition, we trained the network to predict several commonly mutated genes in STAD. Furthermore, we showed that deep learning can predict MSI directly from H&E images. MATERIALS AND METHODS: We studied the automatic classification of 'normal' and 'tumor' regions using a total of 432 H&E- stained WSIs from TCGA gastric cancer image dataset. The slides were tiled in non-overlapping 360x360 pixel windows at a magnification of 20x. We used 70% of those tiles for training, 15% for validation, and 15% for final testing. The deep learning with convolutional neural networks was performed based on inception v3 architecture. To study the prediction of gene mutations from H&E images, average area under the curve (AUC) values for KRAS and SMAD4 mutation (93 and 88 cases, respectively) were calculated using our automatic tumor classification deep-learning approach. To study the prediction of MSI (MSS vs. MSI-H) from H&E images, 383 cases were enrolled using the same approach. RESULTS: The performance of our method is comparable to that of pathologists, with an AUC of up to 0.999. Furthermore, we trained the network to predict two commonly mutated genes in STAD (KRAS and SMAD) and investigated whether they can be predicted from pathology H&E images. We found that KRAS and SMAD mutation can be predicted from pathology images, with AUCs of 0.711 to 0.737, similar results from previous studies with non-small cell lung cancer histopathology images using deep learning. For the prediction of MSI, patch-level and patient-level AUCs were 0.843 and 0.912, respectively, which is superior to the previous studies with TCGA-COAD and -STAD histopathology images. CONCLUSIONS: These findings suggest that deep-learning models can assist pathologists in the detection of cancer subtypes and in the prediction of gene mutations and MSI status. After training on larger datasets and prospective validation, this approach has the potential to provide immunotherapy to a much broader subset of patients with STAD.

Mapping the Emergence and Migration of Hematopoietic Stem Cells and Progenitors During Human Development at Single Cell Resolution

Feiyang Ma, Vincenzo Calvanese, Sandra Capellera-Garcia, Sophia Ekstrand, Matteo Pellegrini, Hanna K. A. Mikkola

Department of Molecular, Cell and Developmental Biology, UCLA, Los Angeles, CA, USA

Hematopoiesis is established during development through multiple waves of blood cell production, starting with lineage-primed progenitors required for the embryos needs, and culminating in the generation of self-renewing hematopoietic stem cells (HSCs) for life-long hematopoiesis. Although hematopoietic ontogeny has been studied extensively in mice, we lack knowledge of the anatomical, temporal and molecular map for hematopoietic development in human. Prior studies suggest that HSCs emerge from hemogenic endothelium in the aorta-gonad-mesonephros (AGM) region between 4-6 weeks of human gestation. Extraembryonic sites including the placenta, umbilical and vitelline arteries, and the yolk sac, have been proposed to generate HSCs in the mouse. However, whether the same sites generate HSCs in human is unclear, mainly due to the limited access to developmental tissues and lack of reliable methods to identify developing human HSCs. We created a single-cell transcriptome map of hemato-vascular cells (CD34+ and/or CD31+) from human hematopoietic tissues at 1st and 2nd trimester. Using a molecular signature of self-renewing HSCs defined in our previous molecular and functional studies, we could identify CD34+Thy1+RUNX1+HOXA7+MLLT3+HLF+ cells as HSCs throughout development. Analyses of 5-wk AGM revealed a distinct population of newly emerged HSCs that vanished by 7 wks. HSCs colonized the fetal liver by 6 wks, where they expanded and differentiated beyond 15 wks. Small but distinct population expressing HSC molecular markers was reproducibly detected in 5 wk placentas. At this time, the heart, umbilical cord and fetal liver lacked clear HSC populations, implying minimal spreading through circulating blood. Interestingly, preceding HSC colonization, the 5 wk fetal liver already harbored CD34+Thy1-RUNX1+HOXA7-MLLT3-HLF- progenitors that co-expressed markers associated with erythro-myeloid and lympho-myeloid potential. Comparable populations were abundant in the yolk sac, suggestive of their origin. This data-set provides an unprecedented resource to dissect the dynamics and molecular pathways governing the emergence and progression of distinct waves of hematopoietic cells during human development, and serves as a reference map for the generation of HSCs in vitro for therapeutic purposes.

Large-scale Machine Learning and Graph Analytics for Functional Prediction of Pathogen Proteins

Jason McDermott[1], Song Feng[1], William Nelson[1], Joon-Yong Lee[1], Sayan Ghosh[1], Ariful Khan[1], Mahantesh Halappanavar[1], Justine Nguyen[2], Jonathan Pruneda[2], David Baltrus[3], Joshua Adkins[1]

[1]Pacific Northwest National Laboratory, [2]Oregon Health & Science University, [3]University of Arizona

Proteins enact the functionality encoded by genomes and so understanding protein function is critical to many areas of biology. Prediction of protein function from sequence is possible because of evolutionary relationships between proteins with similar functions, and existing algorithms can identify the corresponding sequence similarity. However, many proteins have similar functions but diverse sequences, which thwart existing methods, and driven by advances in sequencing technology the number of protein sequences with no known function or similarity to proteins of known function is large and growing rapidly. We use reduced amino acid alphabet mapping and kmer-based protein sequence representation to detect functional similarities between proteins and apply this method to bacterial and viral proteins that mimic eukaryotic ubiquitin ligases and deubiquitinases and classes of bacteriocins. These models allow prediction of novel examples that are not detected by traditional sequence similarity, and can provide insight into active sites or other functional domains for the proteins. To explore sequence space in a more discovery-oriented way we have applied this approach to a very large set of bacterial protein sequences (>20 million sequences) and use a GPU-based algorithm to quickly calculate a similarity graph based on protein features beyond traditional sequence similarity. Exascale graph analytics methods are used to identify groups of closely related sequences from the similarity graph. We show that this method can recapitulate known relationships between proteins, highlight inconsistencies in the underlying protein database, and provide hypotheses for functions of novel proteins thus providing a large-scale sequence landscape.

Gene-set analysis using GWAS summary statistics and GTEx database

Masahiro Nakatochi

Department of Nursing, Nagoya University Graduate School of Medicine

Recently, sample sizes of genome-wide association studies (GWASs) are rapidly increasing. Consequently, many genetic loci associated with traits have been identified. It is difficult to interpret how these many loci identified by GWAS contribute to the traits. As a function of SNP, regulation of gene expression level is considered. The SNP is called as expression quantitative trait loci (eQTLs). The GTEx project revealed many eQTLs in many tissues of human. In this study, I propose an approach of a gene set analysis using GWAS summary statistics and GTEx database to investigate how the genetic loci identified by GWAS contribute to the trait. This approach has three steps. At first, trait-associated SNPs are identified by GWAS. Second, genes whose expression level was associated with trait-associated SNPs in at least one tissue in the GTEx database are searched. These genes were classified into either of positively or negatively correlated genes. Finally, gene set enrichment analyses of positively correlated genes and negatively correlated genes are performed with the modified Fisher's exact test to identify trait-associated pathways or gene sets. Using this approach, I found serum uric acid (SUA)-associated gene sets based on a SUA GWAS. Gene set enrichment analysis of UniProt terms found the terms "Williams-Beuren syndrome", "sodium", "transport", "sodium transport", and "alternative splicing" were enriched for the positively correlated genes. This approach provides another insight into the SNPs identified by GWAS.

Targeting Cancer via Signaling Pathways: A Novel Approach to the Discovery of Gene CCDC191's Double-agent Function using Differential Gene Expression, Heat Map Analyses through AI Deep Learning, and Mathematical Modeling

Annie Ostojic

Purdue University

According to a recent Johns Hopkins University study posted in May of 2018, the number of total genes in the genome was recalculated to be 43,162 genes comprised of 21,306 protein-coded genes and 21,865 non-coded genes. With completion of base pair sequencing in the Human Genome Project back in 2003, hope existed for acceleration of new medical treatments and disease intervention. However, earlier bioinformatic processes were unable to produce results quickly enough, so many gene functions remain unknown to date. A need exists to analyze gene functions in pathways to meet a changing medical industry of pharmacogenomics, personalized medicine, and cancer treatments relative to gene expression patterns. New methodology for determining functions of unstudied genes to rapidly extrapolate, classify, and correlate their gene expressions to biological pathways is at the forefront of bioinformatic studies. This research discovered the function of gene CCDC191, a coiled-coil domain-containing protein-coding gene, whose function had not been fully studied nor defined. A novel approach was utilized to determine the function of CCDC191 by combining gene expression analysis, patient survival analysis, differential gene expression, heat map with AI deep learning, and reverse engineering mathematical modeling. This study presents analyses and insights into gene CCDC191 which have not been performed prior, and it provides a replicable methodology which incorporates AI deep learning image classification, and reverse engineering mathematical modeling to determine gene functions in pathways and cancer connectedness.

RFEX: Simple Random Forest Model and Sample Explainer for non-Machine Learning experts

Dragutin Petkovic, Ali Alavi, DanDan Cai, Jizhou Yang, Sabiha Barlaskar

San Francisco State University (all authors)

Machine Learning (ML) is becoming an increasingly critical technology in many areas. However, its complexity and its frequent "non-transparency" create significant challenges, especially in the biomedical and health areas. One of the critical components in addressing the above challenges is the explainability or transparency of ML systems, which refers to the model (related to the whole data) and sample explainability (related to specific samples). Our research focuses on both model and sample explainability of Random Forest (RF) classifiers. Our RF explainer, RFEX, is designed from the ground up with non-ML experts in mind, and with simplicity and familiarity, e.g. providing a one-page tabular output and measures familiar to most users. In this paper we present significant improvement in RFEX Model explainer compared to the version published previously, a new RFEX Sample explainer that provides explanation of how the RF classifies a particular data sample and is designed to directly relate to RFEX Model explainer, and a RFEX Model and Sample explainer case study from our collaboration with the J. Craig Venter Institute (JCVI). We show that our approach offers a simple yet powerful means of explaining RF classification at the model and sample levels, and in some cases even points to areas of new investigation. RFEX is easy to implement using available RF tools and its tabular format offers easy-to-understand representations for non-experts, enabling them to better leverage the RF technology.

Apparent bias toward long gene misregulation in MeCP2 syndromes disappears after controlling for baseline variations

Ayush T. Raman[1,2], Amy E. Pohodich[2], Ying-Wooi Wan[2], Hari Krishna Yalamanchili[2], William E. Lowry[3], Huda Y. Zoghbi[2], Zhandong Liu[2]

[1]Broad Institute of MIT and Harvard, [2]Baylor College of Medicine, [3]University of California Los Angeles

Background: Rett syndrome is a neurodevelopmental disorder caused by mutations in MECP2, a methyl-binding protein whose task is to orchestrate gene expression, and MeCP2 mutations disrupt the expression of several thousand genes. Over the past ten years, a number of studies observed that Rett syndrome and other disorders that affect neuronal synapses seem to preferentially dysregulate genes that are longer than 100 Kb. These length-dependent transcriptional changes in MeCP2-mutant samples are modest, but, given the low sensitivity of high-throughput transcriptome profiling technology, here we re-evaluate the statistical significance of these results. Results: We develop a robust statistical approach to estimate noise accurately and identify statistically significant gene length-dependent changes. We find that the apparent length-dependent trends previously observed in MeCP2 microarray and RNA-sequencing datasets disappear after estimating baseline variability (i.e., intra-sample differences) from randomized control samples across publically available 17 different MeCP2 datasets. We show that even MAQC/SEQC Phase-III benchmark datasets are prone to the long gene bias, which does not include MeCP2 or its effects on expression — suggesting that the bias is not an inherent feature of gene expression following MeCP2 disruption. We hypothesized that PCR amplification, a process shared by both microarray and RNA-seq technologies, might introduce the observed bias in long gene expression. We find no bias with nanoString technology, a technique that does not use PCR amplification, for SEQC/MAQC samples or Mecp2 mutant samples. This confirmed our notion that the previous observations of long-gene bias resulted from amplification-based technologies and the failure to establish a proper baseline. Conclusions: We conclude that accurate characterization of length-dependent (or other) trends requires establishing a baseline from randomized control samples. We propose that smaller fold changes in transcription observed after PCR amplification leads to an overestimation of long gene expression levels.

Prediction of chronological and biological age from laboratory data

Luke Sagers[1], Luke Melas-Kyriazi[2], Chirag J. Patel[3], Arjun K. Manrai[1]

[1]Boston Children's Hospital Computational Health Informatics Program, [2]Harvard University Department of Mathematics, [3]Harvard Medical School Department of Biomedical Informatics

Aging has pronounced effects on blood laboratory biomarkers used in the clinic. Prior studies have largely investigated a single biomarker or population at a time, limiting a comprehensive view of biomarker variation and aging across different populations. Here we develop a supervised machine learning approach to study the aging process using 356 blood biomarkers measured in 67,536 individuals across demographically diverse populations. Our model predicts age with a mean absolute error (MAE) in held-out data of 4.76 years and an R2 value of 0.92. Age prediction was highly accurate for the pediatric cohort (MAE = 0.87, R2 = 0.94) but inaccurate for ages 65+ (MAE = 4.30, R2 = 0.25). Extensive variability was observed in which biomarkers carry the most predictive power across different age groups, genders, and race/ethnicity groups, and novel candidate biomarkers of aging were identified for specific age ranges (e.g. Vitamin E for ages 18-45). We further show that predictors accurate for one age group may fail to generalize to other groups, and find that nearly a third of all biomarkers exhibit non-linearity near adulthood. As populations worldwide undergo major demographic changes, it will be increasingly important to catalogue biomarker variation across age groups and discover new biomarkers to distinguish chronological and biological aging.

Whole genome sequencing analysis of influenza C virus in Korea

Sooyeon Lim, Han Sol Lee, Ji Yun Noh, Joon Young Song, Hee Jin Cheong, Woo Joo Kim

Division of Infectious Diseases, Department of Internal Medicine, Korea University College of Medicine, Seoul, South Korea; Division of Brain Korea 21 Program for Biomedicine Science, College of Medicine, Korea University, Seoul, South Korea; Asia Pacific Influenza Institute, Korea University College of Medicine, Seoul, South Korea

Through the Hospital-based Influenza Morbidity and Mortality (HIMM) surveillance system, 973 nasopharyngeal swab specimens from children under 2 years of age were collected and tested for influenza viruses using real-time PCR. Among the tested specimens, 383 were positive for influenza A and / or B virus. Influenza C virus was confirmed in five specimens. In this study, we used five influenza C virus positive specimens and a cell-cultured influenza C virus. Viral RNA was isolated using the QIAamp viral RNA mini kit (Qiagen, Hilden, Germany) following a manufacturer's instructions. All isolated RNA was finally eluted with 60 ul of distilled water. Reverse transcription reaction was performed by Primescript 1 st strand cDNA synthesis kit (Takara, Shiga, Japan) using uni-5' primer. The genome-wide amplification of the influenza C virus was performed using taq polymerase. The amplified gene fragments were performed using the Nextera XT DNA library Prep kit (Illumina), according to the manufacturer's protocol.This study was the first report of influenza C virus using NGS analysis in South Korea. In this study, young children with influenza C virus infections had acute respiratory illnesses, such as fever, rhinorrhea, and cough, but no pneumonia or severe respiratory illness was observed. Based on NGS analysis, we can expand our understanding various symptoms of influenza C virus.

Mining the Humuhumunukunukuapua and the Shaka of Autism with Big Data Biomedical Data Science

Peter Washington, Brianna Chrisman, Kaiti Dunlap, Aaron Kline, Arman Husic, Michael Ning, Kelley Marie Paskov, Nathaniel Stockham, Maya Varma, Emilie LeBlanc, Jack Kent, Yordan Penev, Min Woo Sun, Jae-Yoon Jung, Catalin Voss, Nick Haber, **Dennis P. Wall**

Departments of Pediatrics (Systems Medicine) and Biomedical Data Science, Stanford University

Mental health is arguably at the core of all health, and early childhood mental health predicts a long term healthy life course. Yet, finding, treating, and preventing mental health disorders in children is limited by reach and scalable methods. Thankfully, advances in AI and ubiquitous technology have marshaled in unparalleled opportunities for scalable mobile health. We have constructed a series of mobile solutions that treat and track while simultaneously building novel computer vision libraries for precision models. These solutions function as mobile games that are highly engaging and designed for the individual, encouraging compliance with the required "dose" while passively collecting metrics to measure, and ultimately predict outcomes. We can quantify or digitize a child's phenotype through these passively collected data, not just once, but many times, as the child plays our games and learns through playing. These games engender trust and as they do, we "crowd" build a community of stakeholders that not only shares Phenome data, but also data on their Genome and the Environment. With the 3 modalities, we use data fusion multivariate techniques to resolve the G+E=P equation for autism and set the stage for doing the same in other spectrum disorders across mental health.

Development of a recurrence prediction model for early lung adenocarcinoma using radiomics-based artificial intelligence

Hee Chul Yang, Gunseok Park, Ji Eun Oh

Division of Convergence Technology, National Cancer Center Research Institute

Purpose: This study aimed at predicting the recurrence after curative resection for the patients with lung adenocarcinoma (ADC) using the phenotypic radiomics features obtained from the CT images. Material: From January 1, 2010, to December 31, 2015, a total of 604 primary lung ADC patients who had the tumor size of 1-3cm underwent curative resection at a single institution. Method: A total of 604 patients' preoperative CT images were used for feature extraction. The final dataset was randomized into a training set (n=424) and a test set (n=180) with the ratio of 7:3. Radiomics features were selected from t-test ($P<0.05$) and a radiomics signature was classified by the logistic regression model. The optimal model was evaluated through a ROC curve. Result: In a logistic regression analysis, 6 radiomics features were finally selected from 51 features to build a radiomics signature that was significantly associated with recurrence. The optimal model was built with features associated with the dependent variable. They presented good performance in the prediction of recurrence alone with an AUC of 76.2% accuracy. The test set validated 72.2% accuracy. Conclusion: The radiomics signature can be a useful recurrence prediction tool even in small-sized lung ADC.

DRLPC: Dimension Reduction of Sequencing Data using Local Principal Components

Yun Joo Yoo[1], Fatemeh Yavartanu[1], Shelley B. Bull[2]

[1]Seoul National University, [2]The Lunenfeld-Tanenbaum Research Institute

Genome-wide association studies (GWAS) using single nucleotide polymorphism (SNP) data usually have millions of variables with complex correlation structure resulting from linkage disequilibrium. When multi-SNP joint analysis using multiple regression is applied, a dimension reduction method such as principal component analysis can be considered. Replacing SNP data with principal components can resolve multi-collinearity which often occurs in regression using high-density sequencing or imputed SNP data. However, the principal components constructed from all SNP variables in a region are hard to interpret as a biological entity and are not useful for localization and fine mapping. In this study, we propose an algorithm DRLPC (Dimension Reduction using Local Principal Components) to reduce the dimension for regression analysis by selecting clusters of SNPs in high correlation and replacing each cluster by a local principal component constructed from the SNPs in the cluster. The algorithm aims to resolve multicollinearity between updated variables by considering variance inflation factor (VIF) and removing variables with high VIF. We examined the behaviour of DRLPC by applying the algorithm to the 1000 Genomes Project data. Chromosome 22 SNP sets of three populations (EUR, ASN, AFR) were dimension reduced for each gene region separately comparing several choices of threshold values for clustering and principal components selection. When averaged across the genes, the ratio of the number of final variables over the number of original variables was 50 % for the genes with 5~10 SNPs and as low as 10% for the genes with more than 1,000 SNPs. The reduction rate was smaller for the AFR population compared to the other populations EUR and ASN, possibly due to weaker LD in the African population. We also compared the power of multi-SNP tests constructed based on regression results obtained from the original data and dimension reduced data. These tests include generalized Wald, LC (linear combination) tests, and MLC (Multi-bins linear combination) tests. LC tests and MLC tests are also dimension reduction techniques in the sense that LC combines all individual effects into a one degree of freedom test and and MLC combines the individual effects into a linear combination within a bin (cluster) and constructs a test with degrees of freedom equal to the number of clusters. Since DRLPC uses the same clustering algorithm based on clique partitioning as MLC we compared results of MLC with original data to DRLPC Wald test with processed data under the same clustering threshold and found that they yield similar power. We conclude that DRLPC can provide efficient dimension reduction while resolving multi-collinearity and also lessens the problem of interpretability because these principal components represent smaller sized regions, possibly short haplotypes.

Meta-analysis in exhausted T cells from Homo sapiens and Mus musculus provides novel targets for immunotherapy

Lin Zhang[1], Yicheng Guo[2], Hafumi Nishi[1]

[1]Tohoku University Graduate School of Information Sciences, [2]Columbia University, Department of Systems Biology

Antibody target immune checkpoint inhibitors to reverse T cell exhaustion is a promising approach for immunotherapy of cancers. However, the therapeutic efficacy is still low for known immune checkpoint inhibitors, such as PD1 and CTLA4. T cell exhaustion is a state of T cell dysfunction during chronic infections and cancers. It exhibits several characteristic features, such as poor effector functions in a hierarchical manner, impaired memory T cell potential, sustained upregulation and co-expression of multiple inhibitory receptors. The mechanism and pathways for T cell exhaustion remain to be fully described. In this study, we performed meta-analysis with 7 datasets from both humans and mice, to uncover the molecular mechanism of T cell dysfunction. Through gene set enrichment analysis, the predefined exhaustion gene sets were observed to be significant enrichment in the exhausted T cells. The different expression analyses showed an overlap of 21 upregulation and 37 downregulation genes shared by exhausted T cells in humans and mice. These genes were significantly enriched in exhaustion response-related pathways, such as signal transduction, immune system process, and regulation of cytokine production. Besides, co-expression analysis identified 175 genes were highly correlated with exhaustion trait in humans and mice. Above all, our study revealed that TOX and CD200R1 might be considered as potential and high-efficient targets for immunotherapy.

INTRINSICALLY DISORDERED PROTEINS (IDPS) AND THEIR FUNCTIONS

POSTER PRESENTATIONS

Disordered Function Conjunction: On the in-silico function annotation of intrinsically disordered regions

Sina Ghadermarzi, Akila Katuwawala, Christopher J. Oldfield, Amita Barik, Lukasz Kurgan

Virginia Commonwealth University

Intrinsically disorder regions (IDRs) lack a stable structure, yet perform biological functions. The functions of IDRs include mediating interactions with other molecules, including proteins, DNA, or RNA and entropic functions, including domain linkers. Computational predictors provide residue level indications of function for disordered proteins, which contrasts with the need to functionally annotate the thousands of experimentally and computationally discovered IDRs. In this work, we investigate the feasibility of using residue-level prediction methods for region-level function predictions. For an initial examination of the multiple function region-level prediction problem, we constructed a dataset of (likely) single function IDRs in proteins that are dissimilar to the training datasets of the residue-level function predictors. We find that available residue-level prediction methods are only modestly useful in predicting multiple region-level functions. Classification is enhanced by simultaneous use of multiple residue-level function predictions and is further improved by inclusion of amino acids content extracted from the protein sequence. We conclude that multifunction prediction for IDRs is feasible and benefits from the results produced by current residue-level function predictors, however, it has to accommodate inaccuracy in functional annotations.

MUTATIONAL SIGNATURES

POSTER PRESENTATIONS

Transcription-associated regional mutation rates and signatures in regulatory elements across 2,500 whole cancer genomes

Jüri Reimand

Ontario Institute for Cancer Research, University of Toronto

The genomes of healthy and cancerous cells accumulate somatic mutations over time with complex variations across tissues and genomic contexts. Certain classes of functional elements of the genome are subject to differential mutation rates due to regionalized activities of mutational processes. To investigate regional mutations, we developed RM4RM, a statistical framework for detecting differential mutation rates and trinucleotide signatures in sets of genomic regulatory elements. To validate our model, we first analyzed CTCF binding sites across >2,500 whole cancer genomes of 39 cancer types of the ICGC-TCGA PCAWG cohort. We found significant mutation enrichments in CTCF sites in liver, esophageal, breast and other cancer types that was primarily driven by T>C/G mutations and multiple rare mutation signatures of unknown etiology. Transcription start sites of protein-coding genes and a broader set of experimentally-defined regulatory elements derived from primary tumors of the TCGA project also showed significantly elevated regional mutation rates in multiple cancer types. TSS-specific regional mutation enrichment was particularly dominant in highly transcribed genes of matching tumors while none was apparent in silenced genes. In contrast, no mutation enrichment dependency on transcript abundance was observed in distal regulatory elements. These data indicate a transcription initiation-coupled mutational process active in multiple cancer types supported by multiple mutational processes and trinucleotide signatures specifically enriched in highly-transcribed TSSs. Our findings and statistical model enable detailed studies of the mechanisms of somatic mutagenesis and advances our understanding of genetic drivers of disease.

Complex mosaic structural variations in human fetal brains

Shobana Sekar[1], Livia Tomasini[2], Maria Kalyva[3], Taejeong Bae[1], Logan Manlove[1], Bo Zhou[4], Jessica Mariani[2], Fritz Sedlazeck[5], Alexander E. Urban[4], Christos Proukakis[3], Flora M. Vaccarino[2], **Alexej Abyzov**[1]

[1]Mayo Clinic, [2]Yale University, [3]University College London, [4]Stanford University, [5]Baylor College of Medicine

Somatic mosaicism in cells of the human brain is common and may have functional consequences that lead to diseases including neurological ones. Mosaic variations in brain can be point mutations, insertions of mobile elements, and structural changes. Previously we detected and described 200-400 mosaic point mutations per single cell clones from cortices of three human fetuses (15 to 21 weeks postconception). Here we describe four mosaic structural variations (SVs) in the same brains. The SVs were of kilobase scale and complex, i.e., consisting of deletion(s) and a few rearranged genomic fragments that sometimes originated from different chromosomes. Sequences at breakpoints at the rearrangements had microhomologies suggesting their origin from replication errors. One SV was found in two clones and we timed its origin to ~14 weeks postconception. Our study reveals the existence of mosaic SVs, likely arising from cell proliferation, in the human brain in mid-neurogenesis.

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

POSTER PRESENTATIONS

Stratification of kidney transplant recipients based on temporal disease trajectories

Isabella Friis Jørgensen PhD[1], Søren Schwartz Sørensen PhD[2], Søren Brunak PhD[1]

[1]Novo Nordisk Foundation Center for Protein Research - Faculty of Health and Medical Sciences - University of Copenhagen - Blegdamsvej 3B - DK-2200 Copenhagen N -Denmark; [2]Department of Nephrology - Rigshospitalet - Copenhagen University Hospital - Blegdamsvej 9 - DK-2100 Copenhagen Ø -Denmark

Organ transplantations often improve the life of chronically sick patients. However, immune-suppressive medication given to transplant recipients increase the risk of complications, especially infections and infection-related death. One in five kidney transplant recipients die from infection. We want to stratify kidney transplant recipients into groups of patients with different patterns of infectious diseases and mortality to predict which patients have higher risk of specific infections. We use the Danish National Patient Registry (DNPR) that contains hospital diagnoses for 6.9 million patients from the entire Danish population from 1994 to 2018. We use a previously published method to identify significant time-dependent disease trajectories for all patients with a kidney transplantation. Subsequently, we use hierarchical clustering of Jaccard distances between the disease trajectories to find distinct groups of trajectories from kidney transplant recipients. In the DNPR, we identified 5,644 patients with a kidney transplantation resulting in 43 significant disease trajectories that consist of three consecutive diseases including several infectious-related diagnoses. More than 87% of the kidney transplantation recipients follow at least one of these trajectories; hence are diagnosed with the three diseases in the order the trajectory specifies. Clustering reveals two main groups of temporal disease trajectories. We identify patients following the two groups of disease trajectories and discover significant differences in mortality after kidney transplantation between patients following different disease trajectories. This study used previous disease history from large-scale hospital diagnoses to stratify common, temporal disease trajectories into two distinct groups. Depending on the type of trajectory kidney transplantation recipients follow significant differences in mortality are seen. These methods can be used to guide clinicians about higher risks of certain infections and mortality of certain groups of kidney transplant recipients.

Modeling Gene Expression Levels from Epigenetic Markers Using a Dynamical Systems Approach

James Brunner[1], Jacob Kim[2], **Kord M. Kober**[3]

[1]Mayo Clinic, Rochester, MN; [2]Columbia University, New York, NY; [3]University of California, San Francisco, CA

Gene regulation is an important fundamental biological process and involves a number of complex biological processes that are essential for development and adaptation to the environment. Understanding the role of epigenetic changes in gene expression is a fundamental question of molecular biology. Predicting gene expression from epigenetic data is an active area of research and previous studies have used statistical approaches for building prediction models. Dynamical systems can be used to generate a model to predict gene expression using epigenetic data and a gene regulatory network (GRN). By dynamically simulating hypothesized mechanisms of transcriptional regulation, we provide predictions based directly on these biological hypotheses. Furthermore, a stochastic dynamical system provides us with a distribution of gene expression estimates, representing the possibilities that may occur within the cell. The purpose of this study is to develop and evaluate a stochastic dynamical systems model predicting gene expression levels from epigenetic data for a given GRN. We model gene regulation using a piecewise-deterministic Markov process (PDMP) where transcription factor (TF) binding is a Boolean random variable representing the bound/unbound state of a binding site region of DNA. TF binding is given as the difference of two Poisson jump processes (i.e., binding and unbinding), so that time between binding and unbinding events is exponentially distributed with propensities taken to be linear functions of the available TF. Epigenetic modification of the TF binding site impacts the binding propensity of TF and is measured as the percentage of methylated bases (i.e., beta). We use a linear ordinary differential equation based on the underlying GRN to determine the value of the transcript between TF binding or unbinding events. We include baseline transcription and decay and are able to solve exactly between jumps of binding/unbinding events. In a discrete space, continuous time Markov process, the equilibrium distribution can be estimated by sampling from a realization of the process. For our continuous space PDMP we can estimate the equilibrium distribution in a similar manner using kernel density estimation with a Gaussian kernel. We estimate the marginal distributions of various gene variables with a 1-dimensional kernel. We use a GRN assume to be known to create a model of gene regulation that includes TF binding dynamics. We associate binding sites with the genes that they regulate and use these associations to create a bipartite graph. The GRN and training/testing data are created from publicly available data. The epigenetic parameter is assumed to be measurable. The remaining parameters are estimated using a negative log-likelihood minimization procedure. We can compute a log-likelihood for a set of paired epigenetic and transcription samples by time averaging a sample path against a Gaussian kernel. We report on the design and evaluation of the model's performance.

Translating Big Data neuroimaging findings into measurements of individual vulnerability

Peter Kochunov[1], Paul Thompson[2], Neda Jahanshad[2], Elliot Hong[1]

[1]University of Maryland School of Medicine, Maryland, USA; [2]University of Southern California, California, USA

We propose an intuitive anatomically informed approach to derive an index of similarity between individual brain patterns and the expected patterns of neuropsychiatric disorders based on Big Data neuroimaging studies. Big Data neuroimaging studies, such as these performed by Enchancing Neuro Imaging Genetics Meta Analysis (ENIGMA) consortium provided scientific community with the regional patterns of effect sizes in common neuropsychiatric disorders such as schizophrenia (SZ), bipolar and major depressive disorders (BP and MDD), epilepsy (EP), Alzheimer's dementia (AD), mild cognitive impairment (MCI) and others. These patterns describe regional deficit using standardized sMRI, dMRI and rsfMRI workflows. They are derived from statistically powerful and inclusive samples and are highly reproducible (r=0.8-0.9) in independent samples. We developed "Regional Vulnerability Index" (RVI) to measure similarity between an individual and the expected pattern of the patient-control differences RVI can be calculated for a single or across imaging modalities. For a single modality RVI, example uses Fractional Anisotropy (FA) measure from dMRI, is calculated as following. FA for each of the 23 major white matter regions, as defined by ENIGMA atlas, in an individual is converted to z-values by (A) calculating the residual values after regressing out age and sex effects for this region and (B) subtracting the average value for a region and (C) dividing by the standard deviation calculated from the healthy controls. This produces a vector of 23 z-values (one per region) for each individual in the sample. RVI is calculated as the correlation coefficient between 23 region-wise z values for the subject and the patient-controls effect sizes in ENIGMA. RVI takes values from 1 (individual pattern is aligned with disorder pattern) to -1 (individual pattern is in anti-alignment). For cross-modality research, RVI can be expanded hierarchically by building a combined vector that includes multiple phenotypes. For example, the RVI-White Matter calculation uses a vector of 69 values that combine tract-wise FA, radial (RaD) and axial (AxD) diffusivity values per person. To merge effect sizes across diverse domains, we use a pseudo-ordinary transformation that maps effect sizes between 0 and 1 while preserving the relative distance between them. We first demonstrated that RVI-SZ values are significantly elevated in patients with SZ and are also predictive of treatment resistance. That is subjects who developed resistance to modern antipsychotic medications had significantly higher RVI-SZ values than these who responded to treatment. We next demonstrated that RVI for SZ were significantly correlated with RVI for AD but not MCI due to significant overlap in deficit patterns between these disorders. We next showed that calculating RVI across multiple modalities produces vulnerability measures that are more sensitive to patient control differences in the independent datasets and showed stronger sensitivity to cognitive deficits and negative symptoms. The RVI calculator tools are distributed with solar-eclipse software (www.solar-eclipse-genetics.org)

Automating new-user cohort construction with indication embeddings

Rachel D. Melamed

Department of Computational Biomedicine and Biomedical Data, University of Chicago

The electronic health record is a rising resource for quantifying medical practice and discovering adverse effects of drugs. One of the challenges of health care data is the high dimensionality of the health record. Any study of patterns in health data must account for tens of thousands of potentially relevant diagnoses or treatments. In this work, we develop indication embeddings, a way to reduce the dimensionality of health data while capturing the information relevant to treatment decisions. We demonstrate that these embeddings recover therapeutic uses of drugs. Then we use these embeddings as an informative representation of relationships between drugs, between health history events and drug prescriptions, and between patients at a particular time in their health history. We show the application of these embeddings in areas of current research. For drug safety studies, particularly retrospective cohort studies, our low-dimensional representation helps in finding comparator drugs and constructing comparator cohorts. This enables us to develop an automated approach to choose comparator cohorts for a treated population.

Reproducibility-optimized statistical testing for omics studies

Tomi Suomi, **Laura Elo**

Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland

Differential expression analysis is one of the most common types of analyses performed on various biological and biomedical data, including e.g. RNA-sequencing and mass spectrometry proteomics. It is the process that detects features, such as genes or proteins, showing statistically significant differences between the sample groups under comparison. However, as different test statistics perform well in different datasets, the choice of an appropriate test statistic has remained a major challenge. To address the challenge, our reproducibility-optimized test statistic (ROTS) optimizes the statistic on the basis of the data by maximizing the reproducibility of the top-ranked features through a bootstrap procedure. Finally, it provides a ranking of the features according to their statistical evidence for differential expression between the sample groups. We have shown the robust performance of ROTS in a range of studies from transcriptomics to proteomics, covering both bulk and single cell measurements. ROTS is freely available as an R package in Bioconductor.

Data Integration Expectation Maps: Towards more informed 'omic data integration

Tia Tate[1], Christian Richardson[2], **ClarLynda Williams-DeVane**[3]

[1]University of North Carolina-Charlotte, [2]Duke University, [3]Fisk University

Innovative data technologies and decreasing costs have expanded the scope of available data relating to various diseases. A vast amount of -omics data generated at diverse levels (DNA, RNA, protein, metabolite and epigenetic) have revealed relationships of various biological processes. Generally, these diverse data types are considered independently while combinations of two or more data types are less explored. This narrow approach often fails to identify the intricate interactions responsible for the etiology of complex disease. Complete biological models of complex diseases are only likely to be discovered if the various levels of -omic mechanisms are considered from an integrative perspective. Integrative models often require the integration of biological, computational, mathematical, and statistical domains. However, a well-documented shortage of researchers with a command of multiple domains exists. Thus, we have proposed the use of Data Integration Expectation Maps (DIEMs) as visual tools for facilitating the understanding of integrating various -omic data types to understand complex diseases by filling in gaps in biological knowledge. DIEMs provide a user-friendly format for understanding integrative model development in complex diseases by 1)identifying data formats that can and/or have been integrated, 2) providing guidance on the best method to integrate the data, and 3) providing an expectation of biological insight to be gained from the integration.

PRECISION MEDICINE: ADDRESSING THE CHALLENGES OF SHARING, ANALYSIS, AND PRIVACY AT SCALE

POSTER PRESENTATIONS

Integrated omics data mining of synergistic gene pairs for cancer precision medicine

Euna Jeong, Choa Park, Sukjoon Yoon

Sookmyung Women's University

Current high-throughput technologies enable simultaneous acquisition of multi-level omics and RNAi/chemical screening data in cancers. Production and integration of these data help identifying associations of drug targets and synergistic biomarkers (mutations or gene expression), thus accelerating their clinical applications and patient stratification. We have extensively carried out cancer big data mining and phenotypic siRNA library screening for finding the optimal combination of targets and biomarkers for advanced cancer therapies such as regulating cancer stem-like cells (CSLCs) and oncogenic transcription factors. Our multiplexed screening dissect phenotypic responses into sensitivity and resistancy to the target knockdown. Combined with mutaome and transciptome data of screened cell lines, targetome-wide knockdown data reveal the functional aspect of synergistic effects between target siRNAs and mutation/transcription signatures, leading to the discovery of novel synthetic lethal gene pairs. Production and integration of these data enabled us to identify target-biomarker combinations for accelerating their clinical applications and patient stratification.

The power of dynamic social networks to predict individuals' mental health

Shikang Liu[1], David Hachen[1], Omar Lizardo[2], Christian Poellabauer[1], Aaron Striegel[1], Tijana Milenkovic[1]

[1]University of Notre Dame, [2]University of California at Los Angeles

Precision medicine has received attention both in and outside the clinic. We focus on the latter, by exploiting the relationship between individuals' social interactions and their mental health to predict one's likelihood of being depressed or anxious from rich dynamic social network data. Existing studies differ from our work in at least one aspect: they do not model social interaction data as a network; they do so but analyze static network data; they examine "correlation" between social networks and health but without making any predictions; or they study other individual traits but not mental health. In a comprehensive evaluation, we show that our predictive model that uses dynamic social network data is superior to its static network as well as non-network equivalents when run on the same data.

Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data

Jiayi Tong[1], Rui Duan[1], Ruowang Li[1], Martijn J. Scheuemie[2], Jason H. Moore[1], Yong Chen[1]

[1]University of Pennsylvania, [2]Janssen Research and Development LLC

Electronic Health Records (EHR) contain extensive patient data on various health outcomes and risk predictors, providing an efficient and wide-reaching source for health research. Integrated EHR data can provide a larger sample size of the population to improve estimation and prediction accuracy. To overcome the obstacle of sharing patient-level data, distributed algorithms were developed to conduct statistical analyses across multiple clinical sites through sharing only aggregated information. However, the heterogeneity of data across sites is often ignored by existing distributed algorithms, which leads to substantial bias when studying the association between the outcomes and exposures. In this study, we propose a privacy-preserving and communication-efficient distributed algorithm which accounts for the heterogeneity caused by a small number of the clinical sites. We evaluated our algorithm through a systematic simulation study motivated by real-world scenarios and applied our algorithm to multiple claims datasets from the Observational Health Data Sciences and Informatics (OHDSI) network. The results showed that the proposed method performed better than the existing distributed algorithm ODAL and a meta-analysis method.

PharmGKB: Automated Literature Annotations

Michelle Whirl-Carrillo[1], Li Gong[1], Rachel Huddart[1], Katrin Sangkuhl[1], Ryan Whaley[1], Mark Woon[1], Julia M. Barbarino[2], Jake Lever[3], Russ B. Altman[4], Teri E. Klein[5]

[1]Department of Biomedical Data Science, Stanford University; [2]Formerly Department of Biomedical Data Science, Stanford University; [3]Department of Bioengineering, Stanford University; [4]Departments of Bioengineering, Medicine and Genetics, Stanford University; [5]Departments of Biomedical Data Science and Medicine, Stanford University

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 scientistS manually curate the primary literature to capture details of published pharmacogenomic studies such as variant-gene-drug-phenotype associations, statistical significance, study size and population characteristics. PharmGKB refers to these manually created annotations as “Variant Annotations.”

PACKAGING BIOCOMPUTING SOFTWARE TO MAXIMIZE DISTRIBUTION AND REUSE

WORKSHOP POSTER PRESENTATIONS

Apollo provides Collaborative Genome Annotation Editing with the power of JBrowse

Nathan Dunn[1], Colin Diesh[2], Robert Buels[2], Helena Rasche[3], Anthony Bretaudeau[4], Nomi Harris[1], Ian Holmes[2]

[1]Lawrence Berkeley National Lab, [2]UC Berkeley, [3]University of Freiburg, [4]INRA

Genome annotation projects involve multi-step workflows that are largely automated. However, even with a fully automated annotation pipeline visual inspection and refinement of diverse types of information such as genomic and transcriptome alignments and predictive models based on sequence elements are critical to assure and improve the accuracy of the genome annotations prior to publication. To this end, Apollo (https://github.com/GMOD/Apollo/) is a web application that provides responsive and customizable visualization and editing of genomic elements. Built on top of the JBrowse genome browser (http://jbrowse.org/) and its large registry of plugins (https://gmod.github.io/jbrowse-registry/), Apollo supports efficient annotation curation through drag-and-drop editing, a large suite of automated structural edit operations, the ability to pre-define curator comments and annotation status to maintain consistency, attribution of annotation authors, fine-grained user and group access and edit permissions, and a visual history of revertible annotation edits. Setting up a new genome annotation in Apollo is straightforward. Apollo can be run from Docker or from provided AWS instances, and genomes with feature evidence can be retrieved from an existing JBrowse directory. We have also recently enabled researchers to upload their genome sequence and features (in FASTA, VCF, BAM, or GFF3 format) directly to Apollo, minimizing the need for scripting or server access. It is also possible to create annotations on the fly from BLAT or BLAST search results, which provides a way to initiate a gene previously annotated on a closely related species.. Apollo provides a Python library that wraps the web-services (https://github.com/galaxy-genome-annotation/python-apollo) so that workflow environments such as Galaxy can be automated so that the output of an automated workflow can directly create genome projects, provide evidence, and manage access to an Apollo instance. Apollo supports several popular formats for data export. Structural genome annotations can be exported as FASTA, GFF3, or VCF (if annotating variants) along with any associated metadata. Functional annotations mapped to Gene Ontology terms can be exported in GPAD2 or GPI2 format. Apollo is an open-source tool used in over one hundred genome annotation projects around the world, ranging from the annotation of a single species to lineage-specific efforts supporting the annotation of dozens of genomes. https://github.com/GMOD/Apollo/ https://genomearchitect.readthedocs.io/

g:Profiler - One functional enrichment analysis tool, many interfaces serving life science communities

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, Hedi Peterson

University of Tartu

Making sense of gene lists plays an important role in majority of biological and biomedical experiments. There are several methods and tools that help the scientists to carry out the computational load of these tasks. One of such is g:Profiler (https://biit.cs.ut.ee/gprofiler), a widely used toolset for functional interpretation and conversion of gene lists from hundreds of species. g:Profiler has served the community since 2007 and continues to provide life scientists with the most up-to-date data and methods to this day. Keeping the service trustworthy, the results reproducible and transparent has been the main goal of the team developing g:Profiler. The success in this end is indicated in the increasing number of user requests per year, which already in 2019 alone is close to 9 million queries. These millions of queries originating across the world reflect the diversity of usage preferences, skill sets and research goals of the scientific community. We, as the developers of g:Profiler, have taken this into account by developing and supporting different access options which, in hindsight, has been a huge factor in the increasing user traffic. On the one hand, g:Profiler web application provides researchers, who want quick and easily interpretable results, with nice visualizations, searchable tables and data export possibilities. On the other hand, there is a large bioinformatics community, whose members prefer to analyze gene lists in an automated manner. We support them by offering a standardized access through public APIs. And, as R and Python are the most popular programming languages among life scientists with informatics expertise, we have simplified the usage of APIs by wrapping them into corresponding packages named gprofiler2 and gprofiler-official, respectively. For the users somewhere in between, g:Profiler is also available from the Galaxy platform, which is a popular framework for data intensive biomedical research pipelines run in a graphical user interface. It is clear that the tools in such an interdisciplinary field need to be flexible in order to fully benefit the research community. However, from our experience, the complexity of providing a widely distributed toolset lies in the maintenance of the services rather than in the development, and this is the core reason for depreciation of tools. In g:Profiler the separate interfaces all use the data and methods from a shared hub making them reliable and consistent with each other even after the frequent data updates. We are positive that g:Profiler has been able to help thousands of researchers across the life science community because our priorities have been to reuse high quality and regularly updated data, and to maximize the access options so that we would not leave any life science subcommunity behind.

Increasing usability and dissemination of the PathFX algorithm using web applications and docker systems

Jennifer Wilson[1], Nicholas Stepanov[2], Ajinkya Chalke[2], **Mike Wong**[3], Dragutin Petkovic[2], Russ B. Altman[4]

[1]Department of Chemical & Systems Biology at Stanford University; [2]Computer Science Dept at San Francisco State University; [3]COSE Computing for Life Sciences at San Francisco State University; [4]Helix Group at Stanford University

Limited efficacy and unacceptable safety confound therapeutic development. Identifying potential liabilities earlier in drug development could significantly improve success rates. Recently, in collaboration with the US FDA, we developed the PathFX algorithm and openly available PathFX web application for better understanding pathway-level safety and efficacy phenotypes associated with a drug's target(s). Running PathFX algorithm locally would enable improved efficiency, security, and privacy, however installation of PathFX and its dependencies is challenging for non-computational scientists and prevents dissemination. In addition, while PathFX-web quickly analyzes network associations, the phenotype clustering feature has high computational costs that limit the efficiency of the shared cloud server. To resolve these challenges, we developed PathFX-web Docker container which provides an easy-to-install, easy-to-use web interface, a standalone command-line formulation to PathFX, added security/privacy and allows leveraging of the computational power of the user's hardware.

TRANSLATIONAL BIOINFORMATICS WORKSHOP: BIOBANKS IN THE PRECISION MEDICINE ERA

WORKSHOP POSTER PRESENTATIONS

Identification of biomarkers related to autism spectrum disorder using genomic information

Leena Sait, Martha Gizaw, Iosif Vaisman

School of Systems Biology, George Mason University

Autism spectrum disorder (ASD) is one of the most common neurodevelopmental disorders. Worldwide, ASD tends to have a prevalence of one per 132 persons, with an estimated prevalence of 1 in 59 children, according to CDC's Autism and Developmental Disabilities Monitoring Network. To date, no effective medical treatments for the core symptoms of ASD exists. However, biomarkers capable of detecting and diagnosing ASD can help to translate experimental research results to bench side clinical practices. Biomarker discovery in ASD is complicated by the diversity of core symptoms which comprise deficits in social communication, presence of rigid, repetitive and stereotypical behaviors, and comorbid medical (e.g., epilepsy) or psychiatric symptoms. The EU-AIMS Longitudinal European Autism Project (LEAP), the largest consortia made a great advancement in the discovery of biomarkers for ASD. It seeks to identify stratification biomarkers using neurobiological or neurocognitive measures, neuroimaging, electrophysiology, biochemistry and genetics. This work is aimed at the identification of single nucleotide polymorphisms (SNPs) based on SNP genotyping in genomic DNA in a large cohort of ASD patients and unaffected related individuals to help understand the exact genetic causes of ASD. We hypothesized that ranking the genes based on distance in the space of the alleles frequencies between affected and unaffected populations can be used to identify new putative biomarkers. The dataset retrieved from the Gene Expression Omnibus database (GSE6754) contains more than 6000 samples from 1,400 families. Our results show that the SNPs that are highly ranked by the distance in three-dimensional genotype count space between all the affected and unaffected subjects in the cohort are more likely to be linked to ASD. These results can open new possibilities for further investigation in identifying the genetic mechanisms of ASD.

A pan-cancer 3-gene signature to predict dormancy

Ivy Tran[1], Anchal Sharma[2], Subhajyoti De[2]

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

Tumor dormancy is characterized by the dissemination of hibernating tumor cells that do not proliferate until years after apparently successful removal of patients' primary cancer, resulting in the late relapse of the cancer. Distinguishing between the risk of early (£ 8 months) and late (³ 5 years) relapse in cancer patients is important for the targeted treatment of the tumor. In this study, we identified 53 genes that were significantly up-regulated or down-regulated in dormant cells, from which three genes, CD300LG, OCIAD2, VSIG4, were determined by recursive feature elimination to be the most important features in predicting tumor dormancy. Using this three gene signature, we trained a Random Forest algorithm on a cross-validated (10 fold repeated 3 times) dataset (n=422) randomly subsetted into training data (75%) and test data (25%), consisting of seven different tumor types - testicular cancer, breast cancer, glioblastoma multiforme, lung cancer, colon rectal cancer, kidney cancer and melanoma. The tuned prediction model yielded 80.19% prediction accuracy using confusion matrix analysis, and 82.74% prediction accuracy when using AUC of a ROC curve as the accuracy metric. When independently testing the model on a validation set (n=44) of liver cancer downloaded from ICGC, confusion matrix analysis yielded a 67.44% accuracy and AUC of a ROC curve yielded a 60.48% accuracy. This identified 3-gene signature can be useful in predicting early or late relapse of cancer in patients in clinical practice.

AUTHOR INDEX

E

F

G

H

I

J

O

P

Q

R

S

T

U

V

W

X

Y

Z