2018 Biomedical Computation at Stanford Symposium

Abstract Book

Stanford University

April 19, 2018

Contents

Keynote Talk Titles

Keynote Talk Titles

Computational approaches for interpreting the epigenome and non-coding genome

Jason Ernst, PhD

University of California, Los Angeles

How evolutionary forces shape the genetic architecture of complex traits

Ryan Hernandez, PhD

University of California, San Francisco

Genetic insights into South Asian population history and disease

Priya Moorjani, PhD

University of California, Berkeley

Mining personal, dense, dynamic data clouds to enhance health and drive discovery

Nathan Price, PhD

Institute of Systems Biology

Student and Trainee Talk Abstracts

KLRD1 expressing NK cells may protect against influenza infection

Erika Bongen
Stanford University

Influenza infects tens of millions of people every year in the United States. Other than notable risk groups, such as children and the elderly, it is unclear what subpopulations are at higher risk of infection. Viral challenge studies, where healthy human volunteers are inoculated with live influenza virus, provide a unique opportunity to study infection susceptibility. Biomarkers predicting whether volunteers will be infected or not would be useful for identifying influenza risk groups and designing vaccines. Using our cell mixture deconvolution method, immunoStates, we estimated immune cell proportions from influenza challenge study whole blood transcriptomic data, and performed a multi-cohort analysis of differences in immune cell proportions in the blood between symptomatic-shedders and asymptomatic-nonshedders prior to influenza exposure. We found that proportions of natural killer (NK) cells were significantly lower in symptomatic-shedders at baseline in both discovery and validation cohorts. We observed that a gene used by immunoStates to define NK cells, *KLRD1*, was expressed at lower levels in symptomatic-shedders at baseline in discovery and validation cohorts. *KLRD1* expression in the blood at baseline negatively correlated with influenza infection symptom severity. *KLRD1* expression 8 hours post-infection in the nasal epithelium from a rhinovirus challenge study also negatively correlated with symptom severity. Our results support a model where an early response by *KLRD1* expressing NK cells can control viral infection.

Meta-analysis of sex-differential gene expression in human liver

Emily Flynn
Stanford University

Women are at more than 1.5-fold higher risk for clinically relevant drug adverse events, including serious adverse events such as drug induced liver injury. While the higher prevalence of adverse events in women is due in part to women being overdosed, research suggests that biological sex differences may impact drug response. Sex differences in the liver, which is the site of all xenobiotic metabolism, have been documented on the expression level; however, across studies there is a lack of consensus as to which genes have sex-biased effects. To address this problem, I applied meta-analysis to publicly available liver microarray datasets to identify sex-differential expression. Meta-analysis has the benefit of increasing reproducibility by aggregating data from multiple studies. I performed a systematic search of expression data from ArrayExpress and GEO to identify studies with sufficient male and female samples from normal human liver tissue. These studies were combined using inverse variance effects meta-analysis to identify genes with sex-differential expression. This analysis highlights the potential of meta-analysis techniques for investigating sex differences in expression data, and generates hypotheses that may help improve the understanding of the biology of sex differences in human liver.

Reconstructing Denisovan anatomy using DNA methylation maps

David Gokhman
Stanford University

The Denisovan is a human group unique for having its DNA sequence and methylation mapped, but whose morphology remains almost completely unknown. Here, we present a method to reconstruct anatomical profiles from DNA methylation patterns. This method is based on linking unidirectional promoter methylation changes to loss-of-function phenotypes of these genes. We tested its performance by assembling anatomical profiles of the Neanderthal and the chimpanzee, and comparing them to their known morphology. We demonstrate that this method reaches ∼87% precision in identifying divergent traits, and ∼89% in predicting their direction of change. We then reconstruct a putative anatomical profile of the Denisovan, and suggest that this group likely shared many traits with Neanderthals, including a projecting face, robust jaws, low forehead and wide pelvis. We also identify additional changes along the Denisovan lineage, such as increased length of the dental arch, and expanded biparietal width. We find that the vast majority of morphologies identified in the late Pleistocene Xuchang crania from China are included in our reconstruction of the Denisovan anatomical profile, providing first genetic support to their classification of these individuals as Denisovans. We conclude that DNA methylation maps provide means to predict morphology, and can be used to uncover anatomical features that do not survive in the paleontological record.

Gene annotation bias impedes biomedical research

Winn Haynes
Stanford University

Our results provide evidence of a strong research bias in literature that focuses on well annotated genes instead of the genes with the most significant disease relationship in terms of both gene expression and genetic variation. Despite the rise of high throughput technologies, annotation inequality has continued to grow over time. While focusing research on the best characterized genes may be natural because it is easy to formulate a mechanistic hypothesis of the gene's function in disease, we propose that omics-era researchers should instead allow data to drive their hypotheses. By focusing on genes with the strongest molecular evidence instead of the most annotations, researchers will break the self-perpetuating annotation inequality cycle that results in research bias.

The block bootstrap method for longitudinal microbiome data

Pratheepa Jeganathan
Stanford University

Microbial ecology serves as a foundation for a wide range of scientific and biomedical studies. Rapidly-evolving high-throughput sequencing technology enables the comprehensive search for microbial biomarkers using longitudinal experiments. Such experiments consist of repeated biological observations from each subject over time and are essential in accounting for the high between-subject and within-subject variability in microbiome data. Unfortunately, statistical tests based on parametric models rely on correctly specifying temporal dependence structure which is unavailable in most microbiome data. In this paper, we propose an extension of the nonparametric bootstrap method that enables inference on longitudinal microbiome data. The proposed moving block bootstrap (MBB) method accounts for within-subject dependency by identifying overlapping blocks of repeated observations within each subject to draw valid inferences. Our simulation studies show an increase in power compared to merge-by-subject (MBS) strategies. We also show that compared to tests that presume independence samples (PIS), our proposed method reduces false microbial biomarker discovery rates. We provide an open-source R package (https://github.com/PratheepaJ/bbootLong) to make our method accessible and the study in this paper reproducible.

Genetic regulatory mechanisms of smooth muscle cells map to coronary artery disease risk loci

Boxiang Liu
Stanford University

Coronary artery disease (CAD) is the leading cause of death globally. Genome-wide association studies (GWAS) have identified more than 100 independent loci that influence CAD risk, most of which resides in non-coding regions of the genome. We generated transcriptome and whole-genome datasets using human coronary artery smooth muscle cells (HCASMC) for 53 unrelated healthy donors, as well as ATAC-seq on a subset of 8 donors. By comparison to publicly available datasets in GTEx and ENCODE, we find genetic regulatory mechanism specific to HCASMC. We further validate the close relevance of HCASMC to CAD risk using on transcriptomic and epigenomic levels. By jointly modeling eQTL and GWAS datasets, we found five genes (*SIPA1*, *TCF21*, *SMAD3*, *FES*, and *PDGFRA*) that modulate CAD risk through HCASMC, all of which have biologically relevant function in blood vessel lesion repair. Comparison with GTEx shows that SIPA1 influence CAD risk predominantly through HCASMC.

Poster Abstracts

Genomic and leukocytic inferences of the tumor microenvironment in the context of node-negative and node-positive disease

Alborz Bejnood

Stanford University

There has been significant progress towards understanding the relationship between cancer and immune cells in the tumor microenvironment. Nevertheless, the apparent ability of malignant cells to suppress an immune response remains a major challenge in treatment design. Computational tools that address this issue by examining the metastatic transition from the primary tumor to lymph nodes have generated promising insights. Using publicly available resources annotated with clinical lymph node data, we apply the deconvolution algorithm CIBERSORT to infer the leukocyte representation from bulk tumor transcriptomes in the context of lymph node-positive and lymph node-negative samples. We use a pan-cancer approach to analyze changes across lymph node status and between tumor and matched normal samples, suggesting a role by tumor cells in macrophage activation. Corresponding prognostic significance was assessed by Cox regression.

Identifying a predictive gene expression signature for lymph node involvement in cervical cancers

Joshua Bloomstein, Rie von Eyben, and Elizabeth Kidd

Stanford University

Cervical cancer is the 4th most common cause of cancer death in women worldwide. A major cause of mortality is metastasis, as patients who present with lymph node involvement have significantly worse 5-year overall survival compared to those with localized disease. Lymph node involvement is usually determined by surgical pathology or imaging, such as FDG-PET/CT. To date, there is no simple lab test that accurately predicts cervical cancer metastasis. Though machine learning is being widely used for image analysis, the technology has yet to be applied to gene expression analysis for prediction of cervical cancer metastasis. A tendency of only certain tumors to spread, coupled with the relatively non-invasive nature of biopsy for nucleic acid isolation, make cervical cancer a suitable candidate for metastasis prediction by sequencing. RNA-sequencing data was collected from the primary tumors of 74 cervical cancer patients, with or without lymph node tumor involvement. Least Absolute Shrinkage and Selection Operator (LASSO) was used to select an informative set of genes for prediction. Using a linear support vector machine, a model was developed to classify tumors based on lymph node involvement. A 10-gene panel, which includes Arginase 1, classified training set cases with 93% accuracy and test set cases with 72% accuracy. The TCGA Cervical cancer RNA-seq dataset was used to further validate the prediction model. We are also evaluating the previously described hypothesis of upregulation of hypoxia-inducible genes in lymph node-positive cervical cancer.

Age prediction using plasma global metabolic profiles

Bryan Bunning

Stanford University

Rationale: Untargeted liquid chromatography mass spectrometry (LCMS) provides highly dimensional data on a patients metabolic profile. Before metabolite identification, we used a feature selection algorithm on our unidentified metabolic data to model basic clinical traits, giving confidence that further pathway analysis on more nuanced clinical traits post identification could be feasible and worth further investigation. Methods: Untargeted metabolic profiling was performed using a broad coverage platform involving HILIC- and RPLC-MS (Contrepois et al. 2015) on a twin cohort (N=330, 432 total plasma samples including 3mo-2yr follow up visits, 65% female, aged 0-82yr). 1326 metabolic features were extracted after data pre-processing and were used for modeling age as a continuous response variable using the R package randomForest. We also performed a categorical model by using a subset of adults, aged 18-35 (N=101) and 60+ (N=85). Results: The regression model resulted in a mean of squared residuals of 84.52 years with 83.03% of variance explained. 164 significant metabolic features were used in the model. The classification model yielded an out of bag error of 3.23% using 142 significant features. 71 (30.2%) features overlapped between the two models. Conclusion: Random Forest was capable of predicting age using unidentified plasma metabolic profiles independent of gender. Interestingly, different metabolites were selected as predictors modeling continuous age versus categorical age. Currently, we are working to annotate relevant peaks to allow for various pathway analyses and provide context to the features most important to our models.

Inferring microbiome networks

Claire Donnat

Stanford University

From longitudinal biomedical studies to social networks, graphs have emerged as a powerful framework for describing evolving interactions between agents in complex systems. In particular, recent works in microbial ecology have advocated using graphs to represent the various symbiotic or competitive interactions between communities of bacteria, yielding an alternative and –for the purpose of some analyses– more informative representation of the data than the raw bacterial abundance counts themselves. Capturing these interactions is key to gain a deeper understanding of the inner workings of the microbiome, which has been shown to be related to various medical conditions ranging from obesity to preterm-birth or antibiotics resistance. A crux of the analysis thus lies in the processing of the data into a meaningful graph. However, typically modeled as zero-inflated negative binomial distributions and shown to be compositional, microbiome studies are often considered as statistically atypical and challenging datasets thus calling for the design of tailored statistical tools and making the inference of significant interactions between bacteria an interesting open-ended challenge. Here we propose investigating several frameworks for inferring such microbiome graphs. In particular, after analyzing several current popular approaches, we argue that these rely on heavy parametric assumptions on the data or use unsatisfactory estimators (e.g. standardized proportions). We sugggest an alternative approach to this problem by tackling it through the lens of association rule mining. Our discussion is based on both a set of synthetic examples as well as applications to a set of real-life microbiome studies.

Ascribing an etiology to a mutation signature through logistic regression of cancer genomic RNA expression data

Noah Dove
University of California, Santa Cruz

Individual mutagenic processes produce distinct patterns of base substitutions known as mutation signatures. Of 30 known ones, 13 still have no proposed etiology. We hypothesize that alterations in gene expression linked to increased mutagenesis can be stable in tumors over long periods of time because they would positively contribute to carcinogenesis. Therefore, we investigate whether we could link the presence of a mutation signature to alterations in gene expression in tumors using linear regression. As proof of principle, we selected a mutation signature associated with mismatch repair (MMR) deficiency, signature 15. In stomach adenocarcinoma, we initially produced a model with an AUC of 0.995 (with 5-fold cross-validation), verified by an aggregate model with an F1 score of 0.96 after 1000-fold repetition leaving 1/6 of the sample out each time. Consistent with the nature of signature 15, we identified reduced expression of MLH1, which is part of the core MMR machinery, as the second highest predictor. Random permutations of the signature labels produced an AUC of 0.54. As a control for signature specificity, we applied our analysis to two signatures that are caused by APOBEC activity (signatures 2 and 13) in a variety of tumors and found no correlation with MMR genes. Instead, in one case we obtained a modestly accurate predictive model (aggregate F1 score 0.76), with APOBEC3B ranking second, again consistent with the proposed etiology. Our ability to link mutation signatures with expression data establishes our linear regression approach as a possible strategy to ascribe etiology to orphan signatures.

Automatically detecting disease markers with optical coherence tomography

Nicholas Dwork
Stanford University

Optical Coherence Tomography is a medical imaging modality that records the amount of light reflected from different depths; in this way, it is able to construct sub-surface images of a sample. The amount of light that penetrates through the sample is governed by the Beer-Lambert law and quantified with the attenuation coefficient. This parameter is a relevant biomarker for several diagnostic applications including cancer staging, glucose diffusion, scar assessment, edema detection, and glaucoma diagnosis. This work presents a method to automatically quantify the attenuation coefficient of each voxel of an OCT image. We present results on phantom data, ex-vivo biological data, and in-vivo clinical data.

Uncovering tissue-specific mediators of disease-causing genotypes

Michael Gloudemans
Stanford University

Genome-wide association studies (GWAS) have implicated a myriad of genetic variants in causing human disease. For most of these variants, their precise roles remain cryptic. However, recent characterization of many GWAS loci as gene regulatory variants is a step towards unraveling these variants molecular mechanisms. Here, I show how we and others have pinpointed genes that mediate genotype-phenotype interactions, and I present a comparison of these existing methods. I showcase a few applications of these methods to highly specialized cell types to identify tissue-specific mediators of common human diseases, such as coronary artery disease and age-related macular degeneration. Finally, I describe my work to apply this method across the compendium of publicly available GWAS and eQTL studies.

Flexible record linkage tools for data integration of Chinese electronic health records

Charles Li
University of California, Berkeley

Record linkage, the process of linking records corresponding to the same individual within or between databases, is a core component of medical informatics, allowing for the consolidation of patient information across registries and through time. In many cases, record linkage is a nontrivial problem due to the absence of unique identifiers and likelihood of errors in data entry, which obscure the true match status of records. Measures of similarity for character fields, such as patient names or addresses, can help mitigate the impact of data entry error on record linkage and subsequent analyses. While many similarity measures have been developed and validated for English language data, their utility in addressing errors common to non-Latin languages has not been established in a record linkage context. We develop and adapt functions to quantify the phonetic and visual similarity of Chinese character data and evaluate their contribution to probabilistic linkage of records from China's National Infectious Disease Reporting System.

Phenotype prediction using a vectorized representation of genomic variants

Greg McInnes
Stanford University

Inaccurate treatment of patients imposes a substantial burden on the healthcare system. Part of this problem is due to genetic variants that lead to variability in patient response to drugs. Pharmacogenomics plays a crucial role in understanding the relationship between genes and drugs and a better understanding of the relationship between them could lead to better prescribing practices, helping to reduce waste and improve patient outcomes in the healthcare system. We propose a new approach to predicting drug response using a vectorized representation of genomic variants. By encoding genomic variants as a vector of their functional annotations we can inform predictive modeling with biological knowledge. Early results show that informing models with biologically relevant annotations improves the phenotypic predictive performance. I will discuss the motivation behind this project, the details of constructing a variant vector, and results predicting warfarin dosage from exome data.

A novel multi-cohort framework for analysis of host response signatures

Aditya Rao
Stanford University

Public data repositories and other data sharing platforms have been a massive boon to researchers in the biomedical sciences. Data sharing can reduce costs, save valuable time, and helps ensure the transparency of public research. Furthermore, the robustness that we can achieve from integrating large amount of heterogeneous data has allowed us to make findings that are both significant and durable. However, designing projects around public data can be a double-edged sword. When relevant data is available, it can allow for more efficient and more powerful analyses. But when the data is not present or lacking in quality, it can impose severe limitations on the types of analyses that can be done and on the types of questions that can be asked. In order to increase our ability to create diagnostic signatures from public data, we have created a generalizable statistical framework based around conormalization of gene expression data. This pipeline incorporates many of the strategies employed by our multi-cohort meta-analysis pipeline, but because it works with pooled data, many more datasets qualify for usage. This framework performs well across a variety of diseases and analysis conditions, and has been used to generate a 10 gene diagnostic signature that can differentiate between bacterial and viral infections, as well as a 9 gene signature for distinguishing between malaria and other febrile infections.

R2C2: increasing accuracy of MinION sequencing reads

Roger Volden

University of California, Santa Cruz

Oxford Nanopore Technology's (ONT) long-read MinION sequencer can sequence hundreds of thousands of long DNA molecules. ONT's main limitation however, is raw read accuracy, which is about 90%. As a result, ONT has developed 2D/1D2 library preparation protocols which make it possible to read both template and complement of a DNA doublestrand, which increases the read accuracy to about 93%. To further increase read accuracy, we propose our new method, called R2C2, which uses rolling circle amplification (RCA) of circularized DNA. RCA produces long molecules containing tandem repeats of the original DNA molecule. To analyse these long molecules, we have also developed an analysis pipeline using Smith-Waterman alignments to detect repeats, and partial order alignments to combine these repeats into a consensus sequence. Depending on how many repeats are read, consensus reads can reach accuracy greater than 99%.

Linked-read, whole genome sequencing reveals pervasive chromosomal-level instability and novel rearrangements in brain metastases from colorectal cancer

Li Xia

Stanford University

Little is know about the genomic features of brain metastases from colorectal cancer (CRC) and the contributing genetic factors. We report the results of analysis of a series of brain metastasis from colorectal cancers. Our analysis involves comprehensive characterization of large-scale structural aberrations, including chromosome aneuploidies with a method called linked-read whole genome sequencing. Using a new somatic rearrangement caller (ZoomX), we resolved on average 145 distal inter- and intra-chromosomal somatic junctions per sample with their exact haplotype information and basepair level breakpoints. In addition, we used a new haplotyping method based on these imbalances to generate cancer chromosome haplotypes of up to 146Mbp long. We identified extensive chromosomal-level instability (CIN) in these cancer genomes, with an average of 90 large-scale copy number aberrations per sample in sizes ranging from hundreds of Kbp to hundreds of Mbp. A significant fraction of CIN is likely attributable to earlier chromothripsis events. For example, we identified multiple chromothripsis events disrupting the loci of known cancer genes such as TP53, an essential colon cancer driver. We identified novel rearrangements including an oncogenic gene fusion; e.g. SET/DPP10 among these metastatic samples. Significantly, a majority of chromosome arms demonstrated an allelic imbalance across all of the samples. In summary, the analysis revealed pervasive chromosome-level genome instability as a potential contributor to devastating brain invasion by CRC. It also demonstrated the advantages of linked-read whole genome sequencing. This approach is cost-effective and represents a high-resolution tool to assess genome-wide rearrangements and to generate megabase-scale haplotypes.

Assessing the effects of non-heritable factors on immunological phenotypes

Zheng Yan
Stanford University

The human immune system evolved to provide an intricate and powerful set of defences against a world of pathogens, toxins, and allergenic substances. Much of what comprise the immune system signaling proteins, immune cells, and even organs, have been found to vary considerably among individuals. A majority of this variance has been linked to the influence of non-heritable factors rather than genetics. Thus, uncovering more specific interactions between specific non-heritable factors and components of the immune system could be key to discovering novel insights for immunotherapy and personalized medicine. To address such needs, we performed systems-level analyses of 60 healthy monozygotic twins between 12 and 59 years of age. We evaluated 166 different cell type frequencies, serum cytokines, and signalling pathway responses and analyzed their relationships with several non-heritable factors hypothesized to be particularly influential. We found that human cytomegalovirus infection correlated with especially high levels of CD8+ T cells expressing CD94 or CD85j, both subsets often linked to controlling autoimmune damages. We found similar correlations between cytomegalovirus infection with significantly higher levels of cell types associated with memory inflation. These results support previous concerns that latent viral infections, while appearing benign, have very significant effects on adaptability and aging of the immune system. We also found significant correlations between CD161- CD45RA- regulatory T cells, central memory CD4+ T cells, and effector CD4+ T cells with aging. These results confirm the highly adaptive nature of the immune system, as well as provide more specific insights into the impacts of aging and persistent viral infection on immune phenotype.

Characterizing heterogeneity in the tumor immune microenvironment in triple negative breast cancer by multiplexed imaging

Leeat Yankielowicz-Keren
Stanford University

The immune system plays a critical role in modulating cancer progression. However, knowledge of the composition, organization, and interactions between immune and tumor cells is limited. Cancer development is a complex process that involves multiple cell types, each defined by co-expression of multiple proteins, and depends on the interplay between individual cells in the tumor and microenvironment. This complexity is not captured with current diagnostic methods, which evaluate expression of 1-2 proteins in-situ. We developed MIBI (Multiplexed ion beam imaging), in which antibodies are labeled with metals and detected by secondary ion mass spectrometry. We used MIBI to measure single-cell expression levels of 36 proteins in-situ at sub-cellular resolution in 41 triple-negative breast cancer patients. The complex data generated by multiplexed imaging presents a unique analytical challenge. We developed a multi-step image analysis pipeline for standardized processing of a multiplexed imaging cohort, including low-level data processing, deep-learning-based segmentation, identification of cell types by clustering, and spatial enrichment analysis. Our analysis reveals correlated heterogeneity in the amount and composition of immune infiltration, with enriched co-occurrence of specific immune populations, suggesting organization in the tumor-related immune response. Tumors differ in the degree of spatial mixing between tumor and immune cells. This histological organization is correlated with expression of checkpoint molecules on specific cell populations, enriched along the tumor-immune border. We elucidate organizational features of the tumor-immune microenvironment and demonstrate that high-dimensional imaging combined with our analytical pipeline allow to synergize single-cell data on morphology, expression, and spatial location to dissect tumor complexity and heterogeneity.

Facile generation of single-cell transcriptome and immune repertoire from clinical tumor specimens

Junjie Zhu

Stanford University

Immunotherapies including cell based therapies generate deep and durable responses in patients with chemotherapy-refractory cancers. However, in solid tumors, particularly those with stereotypic driver mutations and resultant neoantigens, the clonality and cell states of tumor infiltrating lymphocytes (TILs) remain poorly understood. Here, we use a droplet-based 5 single cell RNA-sequencing (scRNA-seq) to simultaneously profile transcriptome and immune repertoire of the same cells, enabling the phenotypic characterization of each clonotype. We performed scRNA-seq on unsorted and CD45+-FACS sorted cells from fresh clinical samples of multiple tissue types including lung, liver and kidney, and obtained on average tens of thousands of cells per sample. We demonstrated that high quality single cell suspension can be rapidly and reliably generated from clinical samples. We devised diversity metrics to appropriately classify the immune repertoire within samples, and cross samples of different tissues. We observed distinct cell type compositions, and more importantly context-specific T-cell clonal expansion patterns, suggesting the activation of different molecular programs in these tumors. In summary, we provide a proof of concept for rapid generation of large number of single cell transcriptomes of TILs paired with their corresponding TCR cDNA sequence in fresh tumor samples across different tissue types. Further analysis with this methodology on larger clinical cohorts will provide robust correlative prognostic markers of clinical phenotypes of immunotherapy responses. Our analysis strategy of TCR sequences among large clinical cohorts across multiple tumor types will facilitate cell based therapeutic efforts including CAR T cell or autologous T cell therapies.

Author Index