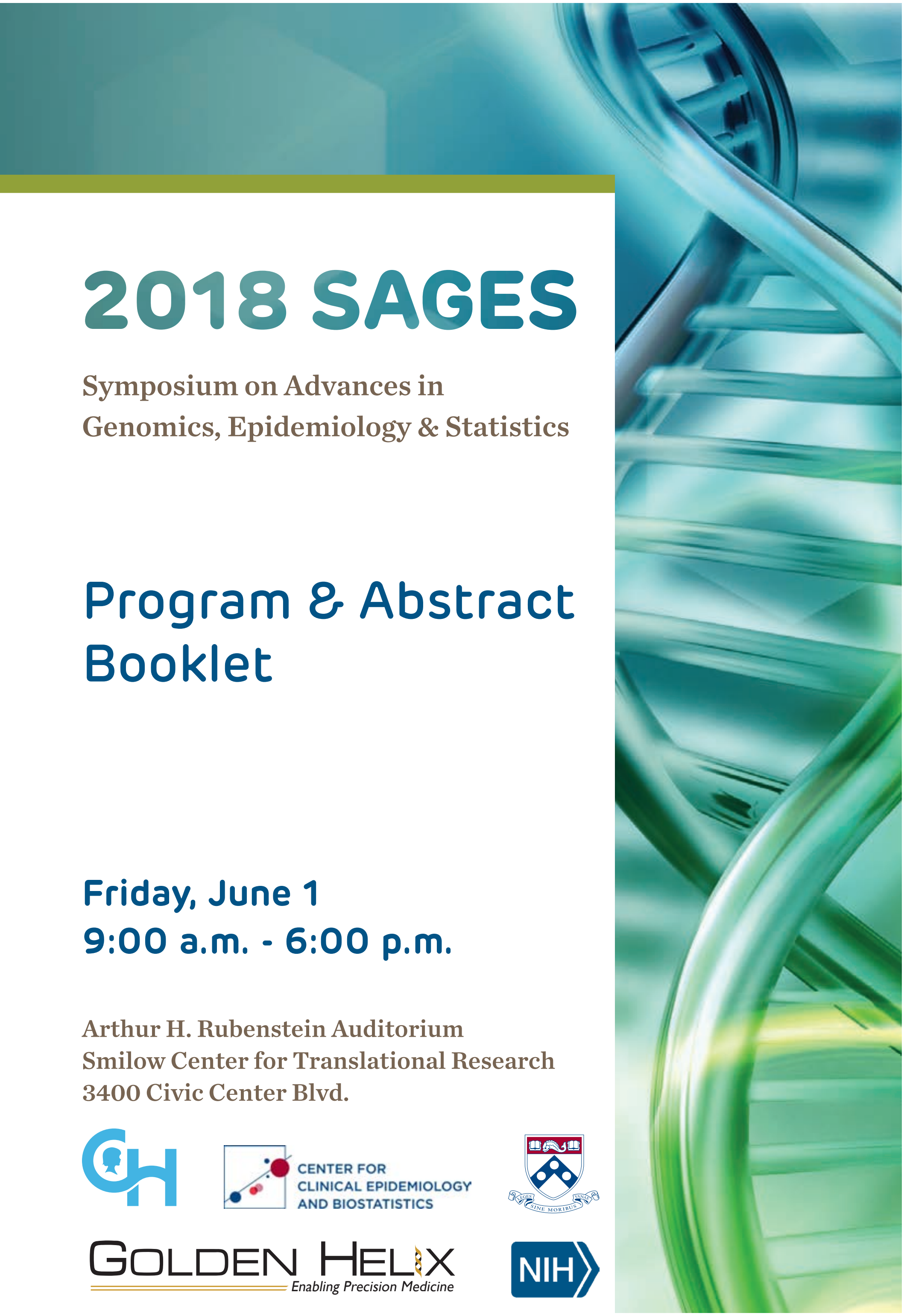

2018 SAGES

Symposium on Advances in Genomics, Epidemiology & Statistics

Program & Abstract Booklet

Friday, June 1
9:00 a.m. - 6:00 p.m.

Arthur H. Rubenstein Auditorium
Smilow Center for Translational Research
3400 Civic Center Blvd.

CENTER FOR
CLINICAL EPIDEMIOLOGY
AND BIOSTATISTICS

LEGES SINE MORIBUS VANAE

GOLDEN HELIX
Enabling Precision Medicine

NIH

Welcome

Advances in technology and significant decrease in the associated costs are driving progress in genomic studies. Studies of whole exome and genome sequences of complex traits in large samples are becoming increasingly common. Other sources of high-dimensional information, including expression, epigenetic, metabolic and microbiomic data, are also being collected in disease and control samples.

SAGES brings together an interdisciplinary group of scientists working in the fields of genomics, epidemiology, and statistics, to address these challenges. The forum provides an opportunity for scientists at all levels in their career to convene and review new developments in these areas of research. The symposium aims to facilitate exchange of ideas and promote interactions and collaborations among participants.

NIH
2017 SAGES
Symposium on Advances in Genomics, Epidemiology and Statistics
Friday, June 9,
9:00 a.m. - 6:00 p.m.
Arthur H. Rubenstein Auditorium
Smilow Center for Translational Research (SCTR)
3400 Civic Center Blvd.
Smilow Center for Translational Research

SAGES is supported by the Center for Clinical Epidemiology and Biostatistics (CCEB) of the Perelman School of Medicine at the University of Pennsylvania, and the Research Institute of The Children's Hospital of Philadelphia (CHOP).

The SAGES organizing committee is especially grateful to Jennifer Forbes-Nicotera (CCEB) and Juliet Kilcoyne (CHOP) for their invaluable effort in the organization of the symposium.

Funding for this conference was made possible in part by grant R13 HG007809 from the National Human Genome Research Institute. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

tional Research
cine
Identifying and inducing gene expression
patterns underlying cell identity

9:00-9:45am	**REGISTRATION & BREAKFAST**
9:45-10:00am	**Welcome and Opening Remarks** **Marcella Devoto,** *CHOP and University of Pennsylvania*
10:00-11:30am	**SESSION 1** Moderator: Barbara Engelhardt, Princeton University
10:00-10:30am	**Identifying disease-relevant cell types from GWAS data** **Hilary Finucane,** *Broad Institute and Harvard*
10:30-11:00am	**Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome** **Michael Hoffman,** *University of Toronto*
11:00-11:30am	**Omic data-based biomarkers for cancer immunotherapy** **Wei Sun,** *Fred Hutchinson Cancer Research Center*
11:30am-1:00pm	**LUNCH**
12:00-1:00pm	**POSTER SESSION 1** (odd numbered posters)
1:00-2:30pm	**SESSION 2** Moderator: Maja Bucan, University of Pennsylvania
1:00-1:15pm	**Identifying recurrent mutations from unphased population-level sequencing data** **Kelsey Johnson,** *University of Pennsylvania*
1:15-1:30pm	**A comparison of univariate and multivariate GWAS methods for analysis of multiple dichotomous phenotypes** **Yasmmyn Salinas,** *Yale University*
1:30-2:00pm	**Constrained statistical inference for the analysis of microbiome data** **Shyamal Peddada,** *University of Pittsburgh*
2:00-2:30pm	**Tracking complex traits over space and time with ancient DNA** **Iain Mathieson,** *University of Pennsylvania*
2:30-3:30pm	**COFFEE BREAK AND POSTER SESSION 2** (even numbered posters)
3:30-4:30pm	**SESSION 3** Moderator: Iuliana Ionita-Laza, Columbia University
3:30-4:00pm	**Genetic risk prediction for complex traits and its relationship to sub-phenotypes in vitiligo** **Stephanie Santorico,** *University of Colorado Denver*
4:00-4:30pm	**Improving the value of public genomic data with phenotype prediction** **Jeffrey Leek,** *Johns Hopkins*
4:30-5:30pm	**CONCLUSION & COCKTAIL RECEPTION**

Poster Numbers & Titles

1	The regulatory landscape of genetic variants associated with psychiatric disorders and neurodegenerative diseases *A Amlie-Wolf, L Qu, EE Mlynarski, CD Brown, GD Schellenberg, LS Wang*
2	Applying Next-Generation Sequencing to the Genetics and Ecology of Rhodnius pallescens, A Vector of Chagas Disease *F Bermudez*
3	More precise metagenomics classifications using unique k-mer counts *FP Breitwieser, SL Salzberg*
4	Widespread hyper RNA editing sites in bovine genome *W Cai, S Zhang, J Song*
5	BrainSeq Phase II: schizophrenia-associated expression differences between the hippocampus and the dorsolateral prefrontal cortex *L Collado-Torres, EE Burke, A Peterson, JH Shin, SA Semick, BrainSeq Consortium, R Tao, A Deep-Soboslay, TM Hyde, JE Kleinman, DR Weinberger, AE Jaffe*
6	Detection of de novo copy number deletions from targeted sequencing of trios *J Fu, E Leslie, A Scott, J Murray, M Marazita, T Beaty, R Scharpf, I Ruczinski*
7	Evaluating the contribution of cell-type specific alternative splicing to variation in lipid levels *K Gawronski, W Bone, E Pashos, Y Park, X Wang, W Yang, D Rader, K Musunuru, B Voight, C Brown*
8	Violence Exposure, Stress Biomarkers and Gender Differences in Buccal Telomere Length among African American Young Adults *L Jackson, F Saadatmand*
9	Assembling the building blocks for a unified splicing code *A Jha, M Gazzara, Y Barash*
10	Integration of Transcriptomic Data Identifies Global and Cell-Specific Asthma-Related Gene Expression Signatures *M Kan, M Shumyatcher, BE Himes*
11	Identifying the genetic and environmental determinants of gene expression variation in Africans *DE Kelly, R Ma, NG Crawford, Y Ren, RA Rawlings-Goss, GR Grant, M Yeager, S Chanock, A Ranciaro, S Thompson, JB Hirbo, W Beggs, TB Nyambo, SA Omar, DO Meskel, G Belay, CD Brown, H Li, SA Tishkoff*
12	Genomics of Addiction *T Koschitzky, L Almasy, COGA Collaborators*
13	HiPPIE2: Identifying the transcription factors mediating enhancer–target gene regulation in the human genome *YC Hwang, PP Kuksa, A Amlie-Wolf, BD Gregory, LS Wang*
14	Evaluation of PrediXcan capabilities to predict gene expression levels and prioritize variant-based associations using datasets with varied population background *B Li, S Verma, Y Veturi, A Verma, Y Bradford, D Haas, M Ritchie*
15	QC Software for Analysis of Sequence Data in Family-based Studies *Q Li, J Bailey-Wilson*

Poster Numbers & Titles

16	Analysis of differential abundance of taxa in microbiome studies using an off-set based linear regression *H Lin, S Peddada*
17	Transcriptome-Guided Imaging Genetic Analysis via a Novel Sparse CCA Algorithm *K Liu, X Yao, J Yan, K Nho, SL Risacher, AJ Saykin, JH Moore, L Shen*
18	Determining and inducing gene expression patterns underlying cell identity *IA Mellis, H Edelstein, R Truitt, PP Shah, W Yang, R Jain, A Raj*
19	Genetic Discrimination between LADA and Type 1 Diabetes within the MHC *R Mishra, JP Bradfield, DL Cousminer, A Chesi, KM Hodge, H Hakonarson, D Mauricio, NC Schloot, KB Yderstræde, B Voight, S Schwartz, BO Boehm, RDG Leslie, SFA Grant*
20	Nonparametric Survival Analysis with Delayed Treatment Effect *K Nam, NC Henderson, D Feng*
21	MAJIQ-HET robustly detects changes in RNA splicing between large heterogeneous sample groups *SS Norton, J Vaquero-Garcia, Y Barash*
22	Life History of Metastatic Breast Cancer Reveals Promising Therapeutic Targets *MR Paul, T Pan, D Pant, N Shih, Y Chen, LA Lee, A Solomon, D Lieberman, JJD Morrissette, D Soucier-Ernst, W Stavropoulos, KN Maxwell, C Clark, GK Belka, M Feldman, A DeMichele, LA Chodosh*
23	DNA methylation changes in Alzheimer's disease across multiple brain regions implicate ANKRD30B *S Semick, R Bharadwaj, L Collado-Torres, R Tao, JH Shin, A Deep-Soboslay, J Weiss, D Weinberger, T Hyde, J Kleinman, A Jaffe, V Mattay*
24	Bivariate GWAS scan identifies six novel loci associated with lipid levels and coronary artery disease *K Siewert, B Voight*
25	Carpe D.I.E.M: A Data Integration Expectation Map of Multi-`Omics Data In Complex Disease Disparities *T Tate Hudson, C Williams-DeVane*
26	Calculating Overall Biological Process Dysfunction Related to Autism Risk Genes Identifies Clinically-Meaningful Genetic Information *OJ Veatch, DR Mazzotti, JS Sutcliffe, RS Schultz, T Abel, B Tunc, SG Assouline, E Brodkin, JJ Michaelson, TK Nickl-Jockschat, ZE Warren, BA Malow, AI Pack*
27	Multiplexed in situ analysis of the human pancreas using imaging mass cytometry *YJ Wang, D Traum, J Schug, K Kaestner*
28	Bulk Tissue Gene Expression Deconvolution Using Single Cell RNA-seq Data *X Wang, M Li, N Zhang*
29	Identifying Tissue-Specific Functional Interaction Modules: An Amygdala Imaging Genetic Studye *X Yao, K Liu, J Yan, K Nho, S Risacher, C Greene, J Moore, A Saykin, L Shen*
30	Generalized Integration Model for Improved Statistical Inference by Leveraging External Summary Data *H Zhang, L Deng, M Schiffman, J Qin, K Yu*

Selected Abstracts & Poster Abstracts

Selected Abstract 1

Identifying recurrent mutations from unphased population-level sequencing data

KE Johnson[1], BF Voight[2,3,4]

1. Genetics & Epigenetics Graduate Group, Perelman School of Medicine, University of Pennsylvania.
2. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania.
3. Department of Genetics, Perelman School of Medicine, University of Pennsylvania.
4. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania.

Recurrent mutations are a hallmark of Mendelian and complex disease. Studies identifying genes enriched for recurrent mutation have improved our understanding of the genetic basis of disease. Furthermore, as population resequencing continues to expand and approach mutation saturation, accounting for recurrent mutations will help to improve estimates of the site frequency spectrum from these data. Given the importance of recurrent mutation in health and evolutionary studies, we require an approach to discover recurrent mutations in genetic data that lacks familial relationships. Here, we present a method to infer recurrent mutations in population-level sequencing data without pedigree data or phased haplotypes. The key intuition underlying the method is that the time to the recent common ancestor (TMRCA) differs between recurrent and identical-by-descent (IBD) mutations. As a summary statistic for the local TMRCA around a site, we measure the 'obligate recombination distance': the distance to the nearest opposite homozygote genotype on either side of the target allele for each pair of carriers. We calculate the likelihood of this collection of pairwise distances under IBD or recurrent mutation scenarios, and identify alleles whose measurements are inconsistent with recent IBD as candidate recurrent mutations. Simulation studies indicate reasonably powered performance of our approach. We applied our method to whole-genome sequencing data from the UK10K project. We observe enrichment of putative recurrent mutations at CpG sites, consistent with the elevated mutation rate at CpGs relative to other contexts. Future applications of this method include incorporating recurrent mutation into tests of rare variant burden in disease.

Selected Abstract 2

A comparison of univariate and multivariate GWAS methods for analysis of multiple dichotomous phenotypes

Y Salinas[1], A DeWan[2], Z Wang[3]

1. Department of Chronic Disease Epidemiology, Yale School of Public Health
2. Department of Chronic Disease Epidemiology, Yale School of Public Health
3. Department of Biostatistics, Yale School of Public Health

Analysis of multiple phenotypes in genome-wide association studies (GWASs) has the potential to enhance statistical power and allows for exploration of pleiotropy. Multi-trait analyses can be conducted using both univariate and multivariate methods. To select an analytic approach, it is important to understand the performance of available methods. However, comparative evaluations of multi-trait methods have primarily focused on the analysis of quantitative traits. Therefore, this study aimed to evaluate the performance of multivariate GWAS methods for analysis of dichotomous (case/control) phenotypes using simulated data. We focused on three methods implemented through R statistical packages-MultiPhen, generalized estimating equations (GEEs), and generalized linear mixed models (GLMMs)-and also compared them to the standard univariate GWAS. We simulated data (N=20,000) for one bi-allelic SNP and two case/control phenotypes assuming a classical liability threshold model, and varied the number of traits associated with the SNP, degree of association, trait-specific prevalences, and cross-phenotype correlation. We generated 10,000 replicates and evaluated power using a genome-wide significance level of 5x10-8. Our results show that, in the absence of pleiotropy, multivariate methods outperform the univariate when there are strong, positive cross-phenotype correlations, but that, in the presence of pleiotropy, the univariate approach tends to outperform multivariate methods when the cross-phenotype correlation is positive. GEEs outperformed MultiPhen and GLMMs across most scenarios. This suggests that, to maximize GWAS discovery, the use of univariate and multivariate (GEE-based) approaches in parallel can be recommended. This study provides researchers with empirical guidelines for the application of these methods to real data.

Abstract 1

The regulatory landscape of genetic variants associated with psychiatric disorders and neurodegenerative diseases

A Amlie-Wolf[1,2], L Qu[2], EE Mlynarski[2], CD Brown[1,2,3], GD Schellenberg [1,2,3], LS Wang[1,2,3]

1. Genomics and Computational Biology.
2. Penn Neurodegeneration Genomics Center; Department of Pathology and Laboratory Medicine.
3. Department of Genetics; Perelman School of Medicine, University of Pennsylvania.

To characterize common noncoding regulatory mechanisms underlying genetic susceptibility to brain-related phenotypes, we applied our INFERNO (http://inferno.lisanwanglab.org/) pipeline to GWAS data for 6 neurodegenerative diseases (Alzheimer's, ALS, FTD, CBD, PSP, Parkinson's) and 5 psychiatric disorders (schizophrenia, bipolar, ADHD, depression, autism). INFERNO defines sets of potentially causal variants underlying noncoding GWAS signals by population-specific LD structure, annotates them with binding sites for 332 transcription factors (TF) and active enhancer annotations in 239 tissues and cell types, empirically quantifies the enrichment of tissue-specific enhancer overlaps, and applies a Bayesian co-localization model to GWAS summary statistics and GTEx eQTL data from 44 tissues to identify target genes regulated by variants overlapping enhancers in matching tissue categories.

INFERNO identified significant enhancer overlaps for several phenotypes in relevant categories including brain (CBD, PD, ALS, PSP, bipolar, schizophrenia), muscle in ALS, and blood/immune in PD, AD, and schizophrenia. Enhancer activity and allelic differences in 3 enhancers identified in the AD analysis have been validated by luciferase, with more experiments underway. INFERNO identified 3,004 strongly colocalized GWAS-eQTL signals for 590 genes across these phenotypes, including 44 genes targeted in more than one phenotype. We also identified 234 TFs with disrupted binding sites in more than one phenotype including 22 observed in all 11 phenotypes, and 125 microRNAs with disrupted seed sites in more than one phenotype. Thus, INFERNO enables the inference of common tissue contexts and regulatory mechanisms underlying genetic susceptibility to neurodegenerative diseases and psychiatric traits, prioritizing signals for post-GWAS research.

Abstract 2

Applying Next-Generation Sequencing to the Genetics and Ecology of Rhodnius pallescens, A Vector of Chagas Disease

F Bermudez

Princeton University.

Chagas disease caused by the protozoan Trypanosoma cruzi, which is transmitted to humans and other mammals through the feces of domestic and sylvatic triatomine bugs. Past studies have mainly targeted well-known domestic triatomine species; however, there are several primarily sylvatic species, like Rhodnius pallescens, that also continue to transmit the parasite to humans. R. pallescens poses a significant threat to reducing vector-borne Chagas disease transmission in Panama, yet it remains underinvestigated.

For this project, we applied the sequenced Rhodnius prolixus genome and Next-Generation Sequencing techniques to detect and quantify parasites, to measure genetic diversity, and to identify blood meal sources among R. pallescens samples. Our results detected the presence of both T. cruzi and T. rangeli within R. pallescens samples, estimated a within-R. pallescens genetic diversity of 0.463% and a divergence from R. prolixus of 4.53%, and confirmed multiple mammalian blood meal sources within each sample. These findings highlight factors related to the vector-parasite-host interactions of R. pallescens that could influence vector population dynamics, the frequency and distribution of infected triatomine species, as well as T. cruzi transmission to humans. They also set the stage for the use of NGS to study other domestic and sylvatic triatomine bug species. Looking ahead, a comprehensive understanding of the genetics and ecology of all relevant domestic and sylvatic triatomine species will allow us to better characterize T. cruzi parasite transmission and the risk for human infection. This knowledge will ultimately help improve vector control and Chagas disease prevention strategies.

Abstract 3

More precise metagenomics classifications using unique k-mer counts

FP Breitwieser[1], SL Salzberg[1]

1. Johns Hopkins University.

False positive identifications are a significant problem in metagenomics classification and are often hard to identify without re-alignment of the reads. However, re-alignment is expensive, and genomic coverages are difficult to summarize up a taxonomy tree. We present KrakenHLL, an ultra-fast k-mer based classifier that reports, in addition to read counts, the number of unique k-mers matching to each taxon. False-positive identifications tend to pile-up in limited genomic regions, and can thus be identified by low unique k-mer counts. KrakenHLL is based on Kraken but achieves higher recall and sensitivity while maintaining the same ultra-fast performance. KrakenHLL in addition supports mapping against multiple databases and an extended taxonomy for identifying strains and plasmids.

Abstract 4

Widespread hyper RNA editing sites in bovine genome

W Cai, S Zhang, J Song

1. Department of Animal Science, University of Maryland, College Park 20740, USA.
2. College of Animal Science and Technology, China Agricultural University, Beijing 100193, China.

RNA editing is an important way to modify nucleotides at specific sites within RNA molecules at a post-transcriptional level in many species. Although it has been systematically studied in many mammalian, little is known about RNA editing in bovine. Here we carefully align and examine the unmapped reads, and detect the hyper RNA editing sites using 9 tissue datasets in bovine. A total of 1,756,773 unique hyper editing sites were detected at 32,455 cluster regions, with 96.12% of the hyper-editing clusters being A-to-G. The brain had the highest enrichment factor, as well as the largest number of unique hyper-edited regions and sites. The 37% of the hyper-edited sites overlapped with 10,951 genes, including 3,234 hyper RNA editing sites being missense variants. Interestingly, we found 92.5% of RNA editing sites were located in repetitive element region. Most of hyper-edited were significantly enriched in Bov-tA repeats. These Bov-tA repeats likely form a dsRNA structure, the ADAR target, by hybridizing with nearby, oppositely oriented Bov-tAs. In addition, we found a significant correlation between the expression of ADAR gene and the number of hyper RNA editing sites in different tissue, suggesting that ADAR expression may contribute to hyper editing occurrence. These results provide a landscape of hyper RNA editing in different tissue. Taken together, the widespread RNA editing clusters and their specificity in different tissues demonstrate evidence for RNA editing likely involve in regulatory mechanism and suggest that the RNA editome should be further studied in bovine.

Abstract 5

BrainSeq Phase II: schizophrenia-associated expression differences between the hippocampus and the dorsolateral prefrontal cortex

L Collado-Torres, EE Burke, A Peterson, JH Shin, SA Semick, BrainSeq Consortium, R Tao, A Deep-Soboslay, TM Hyde, JE Kleinman, DR Weinberger, AE Jaffe

Lieber Institute for Brain Development.

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

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

Results: We identified widespread transcriptional regulation in the DLPFC and the hippocampus over development, with 10,807 genes differentially expressed across age and brain regions that are nominally replicated in BrainSpan. Of these genes, ~55% contain differentially expressed exons and splice junctions that replicated in BrainSpan. We found 48 genes differentially expressed (at FDR < 5%) in hippocampus between schizophrenia patients and non-psychiatric controls, with low replication with DLPFC, suggesting regional heterogeneity of the molecular correlates of schizophrenia diagnosis. We characterized extensive genetic regulation of gene expression with significantly different effects across the two brain regions: we identified 115,787 region-dependent eQTLs (at FDR < 1%), corresponding to 1,484 genes, at the gene, exon, or splice junction level (99 across all three). These region-dependent eQTLs included clinically relevant risk variants – five schizophrenia risk loci showed significant differential regional regulation.

Abstract 6

Detection of de novo copy number deletions from targeted sequencing of trios

J Fu[1], E Leslie[2], A Scott[1], J Murray[3], M Marazita[4], T Beaty[1], R Scharpf[1], I Ruczinski[1]

1. Johns Hopkins University.
2. Emory University.
3. University of Iowa.
4. University of Pittsburgh.

De novo copy number deletions have been implicated in many diseases, but there is no formal method to date however that identifies de novo deletions in parent-offspring trios from capture-based sequencing platforms. We developed Minimum Distance for Targeted Sequencing (MDTS) to fill this void. MDTS has similar sensitivity (recall) compared to adaptions of existing methods. However, MDTS has a much lower false positive rate compared to existing methods, which results in a much higher positive predictive value (precision). Our method also exhibited much better scalability, having been designed to handle more than a thousand trios. MDTS is available as open source software through Bioconductor.

Abstract 7

Evaluating the contribution of cell-type specific alternative splicing to variation in lipid levels

K Gawronski[1], W Bone[1], E Pashos[1], Y Park[1], X Wang[2], W Yang[3], D Rader[1,4], K Musunuru[1,4], B Voight[5], C Brown[1]

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
2. Cardiovascular Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
3. Institute for Regenerative Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
4. Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
5. Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Lipid levels are heritable traits associated with cardiovascular disease risk, and previous genome-wide association studies (GWAS) have identified 150+ loci associated with these traits – however, the genetic mechanisms underlying most of these loci are not well understood. Recent research indicates that changes in the abundance of alternatively spliced transcripts may be an important mechanism contributing to complex traits. With the increased viability of induced pluripotent stem cells (iPSCs) and iPSC-derived hepatocyte-like cells (HLCs), understanding whether these models can be used to interrogate lipid biology is of increasing interest. Consequently, identifying genetic loci that associate with alternative splicing (i.e., sQTLs) in these cells and determining the degree to which these loci are informative for lipid biology would be ideal, but has not been described to date.

We present sQTL discovery efforts using data from sample-matched iPSC and HLC lines, as well as from a separate set of primary liver samples. Genes that are differentially spliced between iPSC and HLC cells are enriched for insulin signaling and lipid metabolism pathways. HLC sQTLs co-localize with GWAS lipid loci unexplained using HLC eQTL data alone, and at least 20% of sQTLs discovered in one stem cell type are not identified in the other cell type. Further, replication analysis indicates that HLC sQTLs more closely represent primary liver sQTLs compared to iPSC sQTLs. Our results provide an important foundation for efforts that use iPSC and iPSC-derived cells to evaluate genetic mechanisms influencing cardiovascular disease risk and complex traits in general.

Abstract 8

Violence Exposure, Stress Biomarkers and Gender Differences in Buccal Telomere Length among African American Young Adults

L Jackson[1, 2], F Saadatmand[1]

1.Howard University, Department of Pediatrics, Washington, DC.
2. Howard University, W. Montague Cobb Research Lab, Washington, DC.

Violence exposure has long-lasting social and biological impacts on African American (AA) health and can lead to physiological changes, including shorter telomere length (TL). While studies have investigated the relationship between TL and life stress, few focus on AA young adults and their direct nexus to violence. This study examines the effect of violence and gender on both stress biomarkers and TL in AA young adults. We examine the relationship between violence exposure, seven stress biomarkers (IgA/G/E/M, C Reactive Protein, Cortisol, Epstein Barr Virus Antigen) and TL in a cross sectional analysis of 50 buccal samples (N=25 males & 25 females) of AA 18-25 years old in Washington DC who experience differential violence exposure (physical, threat, witnessed, and sexual). Average TL was measured by qPCR. Mann-Whitney tests identified differences between males and females in exposure to violence, stress biomarkers, and the measures of TL. Correlations were calculated between TLs, biomarker levels, and violence measures. Elevated sexual violence exposure was positively correlated (RRANGE= 0.22 to 0.55) with all elevated stress biomarkers except IgE. TL in the high violence exposure group was negatively correlated to all stress biomarker levels (RRANGE= -0.09 to -0.31) except for IgG. Violence exposed females had a longer TL than men (MeanF:M=1.87 v 1.62, p≤.054). Male sexual violence exposure was correlated to TL (R=0.575, p≤ 0.03). High violence levels correlate to shorter TLs and higher stress biomarker levels in AA young adults.

Abstract 9

Assembling the building blocks for a unified splicing code

A Jha[1], M Gazzara[2,3], Y Barash[1,2,*]

1. Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States.
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States.
3. Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States.

*To whom correspondence should be addressed: yosephb@upenn.edu

Background

Alternative splicing has a key role in increasing transcriptome diversity. Since differential splicing is prevalent among tissues, cell types and developmental stages, its misregulation can lead to diseases. This motivates computational research efforts to uncover splicing regulatory mechanisms. Splicing codes are probabilistic graphical models that predict splicing outcome in different conditions. The connections in these models can be queried to understand the contribution of different regulatory mechanisms towards the splicing outcome. A key limitation of previous splicing codes is that they model only cassette/exon-skipping events.

Results

Here, we propose a computational framework that extends the work from Jha et al. 2017 in three directions. First, we introduce a unified framework for splicing code for alternative 3' and 5' events in addition to the exon-skipping events. Second, we improve the framework to handle inherent structure in the splicing data, modeling it explicitly. Finally, we develop a convolutional neural network that learns motifs from the RNA sequence de-novo while making use of the existing and newly added features. We evaluate the new framework on diverse tissue datasets from human and mouse and demonstrate its improvement compared to previous models.

Abstract 10

Integration of Transcriptomic Data Identifies Global and Cell-Specific Asthma-Related Gene Expression Signatures

M Kan, M Shumyatcher, BE Himes

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA.

Over 140,000 transcriptomic studies performed in healthy and diseased cell and tissue types, at baseline and after exposure to various agents, are available in public repositories. Integrating results of transcriptomic datasets has been an attractive approach to identify gene expression signatures that are more robust than those obtained for individual datasets, especially datasets with small sample size. We used Reproducible Analysis and Validation of Expression Data (RAVED), a pipeline that facilitates the creation of R Markdown reports detailing reproducible analysis of publicly available transcriptomic data, to analyze asthma and glucocorticoid response microarray and RNA-Seq datasets. Subsequently, we used three approaches to integrate summary statistics of these studies and identify cell/tissue-specific and global asthma and glucocorticoid-induced gene expression changes. Transcriptomic integration methods were incorporated into an online app called REALGAR, where end-users can specify datasets to integrate and obtain instant results that may facilitate design of experimental studies.

Abstract 11

Identifying the genetic and environmental determinants of gene expression variation in Africans

DE Kelly[1], R Ma[1], NG Crawford[1], Y Ren[1], RA Rawlings-Goss[1], GR Grant[1], M Yeager[2], S Chanock[3], A Ranciaro[1], S Thompson[1], JB Hirbo[4], W Beggs[1], TB Nyambo[5], SA Omar[6], DO Meskel[7], G Belay[7], CD Brown[1], H Li[3], SA Tishkoff[1]

1. University of Pennsylvania.
2. National Institutes of Health, Division of Cancer.
3. Frederick National Laboratory for Cancer Research.
4. Vanderbilt University.
5. Muhimbili University.
6. Kenya Medical Research Institute.
7. Addis Abada University.

Gene regulation plays a predominant role in human evolution and complex traits, and high throughput methods are making the measurement of expression variation routine. However, studies to date have failed to capture the breadth of global genetic and environmental diversity by focusing primarily on Western individuals of European descent. Studies of Africans, who harbor the most genetic variation in the world and are exposed to a diversity of environmental variables and diets, are necessary to complete our understanding of how evolution has shaped human genetic and phenotypic diversity. To identify genetic and environmental contributors to gene expression variation in whole blood we have collected RNA sequencing data from 171 individuals representing 9 diverse African populations. Differential expression analyses uncover genes correlated with ancestry and environmental variables, clustering individuals by ancestry and diet. Combining expression data with SNP genotypes, we also map cis-eQTLs in our samples. The majority of identified eQTLs replicate in Europeans, though they can often be mapped to a more narrow credible set owing to the shorter tracks of linkage-disequilibrium in Africa. Conversely, those that fail to replicate are enriched for variants that are at moderate frequency in Africa and are low frequency or monomorphic in Europe. Using these eQTLs, allele frequency differentiation and scans of natural selection identify candidate genes and pathways that may have undergone positive selection in these populations. This study represents the most diverse study of gene expression in Africans to date and highlights the need to extend genomics studies to non-European populations.

Genomics of Addiction

T Koschitzky[1], L Almasy[2,3], COGA Collaborators[4]

1. Penn Undergraduate Research Mentoring Program, University of Pennsylvania.
2. Department of Genetics, University of Pennsylvania.
3. Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia

Approximately 16 million people in the United States are diagnosed with Alcohol Use Disorder. Though it is established that inherited factors play a significant role in one's risk of developing AUD, the specific genes and polymorphisms that influence AUD remain unknown. COGA seeks to identify these genes by utilizing family groups for genome analysis. Previous COGA studies have found that the brain's P300 response is weaker in individuals who have AUD and in their family members. We identified specific chromosomal regions with high LOD scores, indicating a high probability that genes in those regions are linked to the P300 phenotype. We analyzed the SNPS in those regions, focusing on chromosome 6: bps 11614637 - 22747355 in one family and on chromosome 8: bps 29414980 - 80591207 in the second family. We ran a measured genomic association analysis to determine the correlation between each SNP and phenotype. In each family, we found variants in the linkage peak with strong signals (chr6: two variants with p =.004. chr8: one variant with p = .0009 and one with p = .008.) Though it seems unlikely that these SNPs are directly causal to the P300 phenotype, their strong association suggests that they are linked to what is actually contributing to it. A likely possibility is that the variant is in a non-coding region. Though we are still in the process of analysis, our research suggests promising association between AUD and polymorphisms in chromosomes 6 and 8.

Abstract 13

HiPPIE2: Identifying the transcription factors mediating enhancer–target gene regulation in the human genome

YC Hwang[1,2], PP Kuksa[2,3], A Amlie-Wolf[1,2], BD Gregory[1,4], LS Wang[1,2,3]

1. Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine;
2. Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine.
3. Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine.
4. Department of Biology, University of Pennsylvania, Philadelphia, PA.

The majority of disease/trait-associated genetic variations reported by GWAS are located in non-coding regions. A major class of non-coding elements, enhancers, regulate gene expression by recruiting transcription factors and forming long-range interactions with protein-coding promoters. To identify all possible enhancers and genes they regulate genome-wide, we developed a novel method (HiPPIE2) and analyzed latest ultra-high read-depth Hi-C datasets (~20B reads) across human cells (GM12878, IMR90, K562) for physical DNA–DNA interactions. Unlike the standard genome binning approach for mapping chromatin interactions, our method aims at identifying the precise loci of physically interacting DNA regions (PIRs). To more accurately delineate borders of physically-interacting DNA regions, HiPPIE2 utilizes multiple sources of information including (1) read strandness; (2) read distances to the closest restriction sites; and (3) DNA ligation constraints. PIRs are on average 978bps in length and covered 51.4% of the genome. 77% of CTCF binding sites are covered by PIRs, indicating DNA-interacting sites are associated with enhancer–promoter interactions and insulators. Additionally, 86% of PIRs are covered with known open chromatin regions, while 61% of open chromatin regions are covered by PIRs. This suggests that DNA-interacting sites have the binding affinity for transcription regulation but not all open chromatin regions are involved in long-range regulation. Using a spline-based model (Fit-Hi-C), we recovered 1,193,987 significant PIR-PIR interactions. Our novel analysis identified 338,791 DNA-DNA interactions with the evidence of protein binding. We further discovered the motifs involved in DNA-DNA interactions and revealed the transcription factor complexes that may suggest mechanisms underlying long-range regulation.

Abstract 14

Evaluation of PrediXcan capabilities to predict gene expression levels and prioritize variant-based associations using datasets with varied population background

B Li[1], S Verma[1], Y Veturi[1], A Verma[1], Y Bradford[1], D Haas[2,3], M Ritchie[1]

1. The University of Pennsylvania.
2. Vanderbilt University School of Medicine.
3. Meharry Medical College.

Genome-wide association studies (GWAS) have successfully identified numerous disease susceptibility loci, however, the interpretation of discovered disease-related loci remains challenging. To address this issue, functional genomics knowledge is integrated into GWAS, for example, expression quantitative trait loci (eQTLs), which influence gene expression activities, in the anticipation that obtained associations will indicate certain functional relationships between single nucleotide polymorphism (SNP) loci and the diseases of interest. PrediXcan is a computational algorithm developed to exploit eQTL data. We evaluated the PrediXcan capabilities to predict gene expression levels and prioritize genome-wide association study results using three datasets, which have similar or different population structure to the PrediXcan training cohorts. Our results showed that PrediXcan had similar prediction performance for the test datasets having similar or different population background to the model training cohort and was able to prioritize insignificant GWAS associations onto the gene levels, which contributes to disease mechanistic studies.

Abstract 15

QC Software for Analysis of Sequence Data in Family-based Studies

Q Li, J Bailey-Wilson

NHGRI, National Institute of Health.

In recent years, with the increased popularity of DNA sequence data, the research community has paid more attention to family based studies of complex traits. Pedigree data pose new challenges in the quality control steps for sequence variants. Not only do we need to screen data based on various genotyping calling metrics, we also need to use the pedigree structure to detect any markers with inconsistent of inheritance patterns. In this work, we present our QC pipeline software, written in R, for sequence data analysis. It includes functions to extract various genotyping calling measurements, to detect Mendelian inconsistency, and to detect any excessive allele sharing due to remote inbreeding.

Abstract 16

Analysis of differential abundance of taxa in microbiome studies using an off-set based linear regression

H Lin, S Peddada

University of Pittsburgh.

Given that we have at least 10 times more microbial cells than human cells and over 10 times more microbial genes than human genes, it is not surprising that increasingly researchers are finding associations between human diseases and human microbiome. Consequently, researchers are often interested in identifying taxa that are differentially abundant between the healthy and patients with disease and determining the biological functions and processes associated with such taxa. Determination of differentially abundant taxa with a small false discovery rate (e.g. controlled at 5%) is therefore an important first step. A number of procedures have been introduced and used in the literature for identifying differentially abundant taxa. Unlike many of the commonly seen genomic data, the microbiome data are count data that are constrained by a simplex and potentially have a large number of zeros. Consequently, it is not appropriate to apply standard methods of analysis when identifying differentially abundant taxa. Recently a method called “Analysis of Composition of Microbiome (ANCOM)” (Mandal et al., 2015) was introduced to address this problem. Although ANCOM is a general methodology that can be used in a wide range of contexts, it can be computationally intensive. For this reason, in this presentation, we introduce a more straightforward, offset-based regression methodology that is computationally more efficient than ANCOM while retaining all the positive features of ANCOM. Thus, similar to ANCOM, it controls the FDR and maintains similar power as ANCOM. The proposed methodology is illustrated using simulations studies and real data.

Abstract 17

Transcriptome-Guided Imaging Genetic Analysis via a Novel Sparse CCA Algorithm

K Liu, X Yao, J Yan, K Nho, SL Risacher, AJ Saykin, JH Moore, L Shen

1. University of Pennsylvania
2. Indiana University

A fundamental problem in brain imaging genetics is to investigate the association between genetic variations and quantitative traits (QTs) extracted from brain imaging data. Various prior knowledge such as group (e.g., linkage disequilibrium block in genome) and network structure has been incorporated in existing association studies. Given the high dimensionality of imaging and genetic data, these priors can improve the stability in variable selection and interpretability of the results. However, majority of priors used only impose constraints on the genetic or imaging side, but not jointly connect imaging with genetics. To bridge this gap, we propose to use the brain wide gene expression profile available in Allen human brain atlas as a two-dimensional prior to regularize the selection of both brain regions and genetic markers in the association analysis between the SNPs and imaging QTs from multiple modalities such as FDG and amyloid PET scans. With this prior, we expect to explore a set of genes jointly affecting a set of brain regions in both genetic and transcriptomic level. An alternating optimization algorithm is used to solve the formulated transcriptome-guided multimodel SCCA problem. Although the problem is not biconcave, a closed-form solution has been found for each of the two subproblems at the iteration. The empirical results on synthetic and real data show the proposed SCCA framework improves the robustness against noise and false positives/negatives, and facilitates the detection of relevant genes not only associated with the identified brain regions, but also differentially expressed there.

Abstract 18

Determining and inducing gene expression patterns underlying cell identity

IA Mellis, H Edelstein, R Truitt, PP Shah, W Yang, R Jain, A Raj

The University of Pennsylvania.

It is extremely difficult to change differentiated cells of one type into another; transdifferentiation protocols for human fibroblasts to cardiomyocytes, for example, convert only ~1% of cells and often with low fidelity.

We aim to better understand and engineer cell type-intrinsic patterns of gene regulation underlying the maintenance of cell identity for two predefined human cell types of interest: fibroblasts and cardiomyocytes. Toward understanding cell type-intrinsic patterns of gene regulation, we have conducted transcriptome-wide expression profiling using RNAtag-seq on panels of hundreds of samples of drug-perturbed fibroblasts and cardiomyocytes. We find that the expression of genes encoding transcription factors (TFs) known to regulate cardiomyocyte lineage-specific functions is more frequently perturbed, in particular more frequently up-regulated, across a broad range of drug conditions (i.e., “perturbable”) in cardiomyocytes than that of other expressed transcription factor genes. Further, in fibroblasts, we find that some known barriers to transdifferentiation and iPSC-reprogramming are more perturbable than other TF genes.

In the next stage of this project, we will identify previously uncharacterized transcription factor genes that display similar patterns of perturbability in cardiomyocytes and in fibroblasts. Then, we will attempt to develop a more efficient fibroblast to cardiomyocyte transdifferentiation protocol using inhibitors of fibroblast-specific perturbable factors and overexpression of cardiomyocyte-specific perturbable factors. Additionally, we will attempt to improve iPSC reprogramming efficiency of fibroblasts by inhibiting fibroblast-specific perturbable factors. If successful, we hope to establish gene expression perturbability as a useful property for experimentally identifying transcription factors important for the maintenance of cell identity.

Abstract 19

Genetic Discrimination between LADA and Type 1 Diabetes within the MHC

R Mishra[1,2], JP Bradfield[2], DL Cousminer[1,2], A Chesi[2], KM Hodge[2], H Hakonarson[2], D Mauricio[3], NC Schloot[4], KB Yderstræde[5], B Voight[1], S Schwartz[6], BO Boehm[7], RDG Leslie[8], and SFA Grant[1,2]

1. University of Pennsylvania.
2. The Children's Hospital of Philadelphia.
3. Hospital Universitari Germans Trias i Pujol, Spain.
4. German Diabetes Center, Düsseldorf, Germany.
5. Odense University Hospital, Odense, Denmark
6. Main Line Health System.
7. Ulm University Medical Centre, Ulm, Germany.
8. Queen Mary University of London, London, UK.

Studies on type 1 diabetes (T1D) and type 2 diabetes have revealed significant insights into novel biological mechanisms underlying diabetes, yet the genetic etiology of latent autoimmune diabetes in adults (LADA) remains largely unknown; furthermore, improved biomarkers of LADA are required to optimize diagnosis. Our genome-wide association study shows that the major histocompatibility complex (MHC) harbors the strongest association with LADA; however, the association is attenuated compared to observations in T1D cohorts. While T1D susceptibility has been known to be principally harbored within the MHC Class II genes HLA-DQB1 and HLA-DRB1, variation in the MHC class I genes HLA-A and HLA-B have been subsequently shown, through conditional analysis, to increase T1D risk further. We attempted recapitulation of findings in Nejentsev et al. using an imputation-based approach and performing forward stepwise conditional logistic regression of the MHC region in T1D cases (n=1,990) and controls (n=2,856) from the Wellcome Trust Case Control Consortium (WTCCC), and in a LADA cohort (n = 978) using population-based controls (n=1,057).We confirmed the strongest T1D associations at HLA-DRB1 and HLA-DQB1 (P=6.10x10-175 and P=2.90x10-219, respectively), as well as the independent effect of HLA-B (P=1.67x10-14) and HLA-A (P=5.25x10-8) to T1D. We then performed the conditional analysis in the LADA and population-based controls cohort, observing significant association at HLA-DRB1 and HLA-DQB1 (P=5.93x10-22 and P=4.68x10-13, respectively), however we did not observe significant independent effects in HLA-B or HLA-A for LADA, highlighting a potential use for MHC Class I markers in differentiating T1D from LADA.

Abstract 20

Nonparametric Survival Analysis with Delayed Treatment Effect

K Nam[1,2,3], NC Henderson[4], D Feng[1,2,3]

1. BARDS.
2. Merck Research Labs.
3. Merck & Co., Inc.
4. Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

Clinical trials involving novel immuno-oncology therapies often exhibit survival profiles which violate the proportional hazards assumption. In such non-proportional hazards (NPH) settings, the treatment of interest may have a delayed effect where the survival curves in the two treatment arms may largely overlap or cross before the two curves separate. To flexibly model these scenarios, we describe a constrained non-parametric approach which allows the survival functions to have at most one crossing point without making any additional assumptions about how the survival curves in the two treatment arms are related. By utilizing constrained splines and a survival profile dependent step function to model the log cumulative hazard functions, our method provides interpretable measures of treatment while retaining the flexibility to account for survival profiles commonly seen for treatments with delayed effects.

Abstract 21

MAJIQ-HET robustly detects changes in RNA splicing between large heterogeneous sample groups

SS Norton[1], J Vaquero-Garcia[1], Y Barash[1,2]

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania.
2. Department of Computer and Information Sciences, School of Engineering and Applied Science, University of Pennsylvania.

Accurate quantification of differential splicing (DS) from RNA-Seq data is a challenging task which has received much attention in previous years. However, most of the works on DS have focused on comparing two conditions with few replicates, while DS analysis for large heterogeneous datasets such as GTEx remains a challenge.

In order to address this challenge, we developed MAJIQ-HET, which builds upon the MAJIQ framework (Vaquero et al, eLife 2016; Norton and Vaquero et al Bioinformatics 2017). Unlike many tools, the MAJIQ framework allows HET to accurately detect and quantify complex local splicing variations (LSVs), including de-novo junctions not included in the annotation. In contrast to the original MAJIQ, HET discards the assumption of a joint (hidden) inclusion level per group and replaces it with robust statistics, either parametric or nonparametric, which account for missing values common in splicing analysis. Coupled with efficient C/C++ implementation, the new algorithm is 20 times faster than MAJIQ 1.0 and is able to easily process thousands of samples. Using both real and synthetic data based on GTEx and TCGA, we demonstrate MAJIQ-HET's robust performance compared to current state of the art.

Abstract 22

Life History of Metastatic Breast Cancer Reveals Promising Therapeutic Targets

MR Paul[1,5,6], T Pan[1,5,6], D Pant[1,5,6], N Shih[2,5], Y Chen[1,5,6], LA Lee[1,5,6], A Solomon[1,5,6], D Lieberman[2], JJD Morrissette[2], D Soucier-Ernst[3,5], W Stavropoulos[4,5], KN Maxwell[3,5], C Clark[3,5], GK Belka[1,4,5], M Feldman[2,5], A DeMichele[3,5], LA Chodosh[1,3,5,6]

1. Departments of Cancer Biology.
2. Department of Pathology & Laboratory Medicine.
3. Department of Medicine.
4. Department of Radiology.
5. 2-PREVENT Translational Center of Excellence.
6. Abramson Family Cancer Research Institute at the Perelman School of Medicine at the University of Pennsylvania.

The majority of deaths from breast cancer result from distant metastatic disease, rather than primary tumors. Despite this, few genome-wide analyses have been performed comparing primary and metastatic tumors arising in the same patient to evaluate genomic and clonal evolution during tumor progression. To address this gap, we performed whole-exome, shallow whole-genome, and error-controlled ultra-deep sequencing to identify somatic coding mutations and aberrant copy-number in paired primary and metastatic tumors from 29 patients, as well as 38 additional metastases. We find that metastatic tumors are distinct from the originating primary tumors clinical biomarker, genomic, and cellular levels, and that metastases typically arise by linear evolution via monoclonal dissemination late in primary tumor development. Metastatic tumors exhibit increased genomic instability and additional oncogenic mutations, the majority of which were either acquired after dissemination or were present within only a rare subset of cells (<3%) within the primary tumor. Ten genes were found to be preferentially altered in metastases. Pathways preferentially mutated in metastases include focal adhesion, histone methyltransferases, cell-cycle regulation, WNT signaling, and mTOR signaling pathways. Quantitative assays confirmed that mTOR signaling is robustly activated in metastases compared to primary tumors of origin, and can be predicted by integrative genomic analyses. Collation of altered genes and pathways provides the genomic basis for the efficacy of mTORC1 and CDK4/6 inhibitors and alludes to a synergistic benefit of “tri-inhibition” of mTORC1/2, CDK4/6, and standard targets, ER or HER2 to treat patients with systemic cancer and to prevent resistance to targeted therapies.

Abstract 23

DNA methylation changes in Alzheimer's disease across multiple brain regions implicate ANKRD30B

S Semick[*], R Bharadwaj[*], L Collado-Torres, R Tao, JH Shin, A Deep-Soboslay, J Weiss, D Weinberger, T Hyde, J Kleinman, A Jaffe[+], V Mattay[+]

1. Lieber Institute for Brain Development.
[*] Equal contribution
[+] Corresponding author

The etiology of Alzheimer's disease (AD), a complex neurodegenerative disease, is unknown but epigenetic factors may play a role. To better understand the potential role played by one such epigenetic mark that has been previously implicated in AD—DNA methylation (DNAm)—we surveyed the epigenome at 420,852 DNAm sites from unaffected controls (N=49) and AD patients (N=24) across four brain regions (190 control samples, 82 AD samples). We identified 1,494 sites with robust differential methylation across two or more brain regions collectively annotated to 855 genes (FDR<5%). Interestingly, differentially methylated sites were overrepresented in AD genetic risk loci (p=0.00457), and differentially methylated genes were enriched for biological processes related to cell-adhesion and calcium homeostasis (FDR<5%). We prioritized 179 genes with nearby differential methylation as having functional differences in at least one brain region by analyzing corresponding RNA-seq data; expression of these genes was coupled to nearby DNAm and these genes were differentially expressed (at p<0.05). This set of 179 functionally validated genes was enriched for protein-protein interactions (p= 0.000976) and included associations with previously reported (e.g. ANK1, DUSP22) as well as novel genes such as ANKRD30B. We also identified a large number of region-dependent AD changes (17,056, at FDR<5%). These results highlight particular DNAm changes in Alzheimer's disease that have direct biological correlates—implicating existing and novel genes with convergent evidence of dysregulation in AD.

Abstract 24

Bivariate GWAS scan identifies six novel loci associated with lipid levels and coronary artery disease

K Siewert[1], B Voight[2]

1. Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
2. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Plasma lipid levels are heritable and genetically associated with risk of coronary artery disease (CAD). However, genome-wide association studies (GWAS) routinely perform association studies of these traits independently of one another. Joint GWAS for two related phenotypes can lead to a higher-powered analysis to detect variants contributing to both traits. We performed a bivariate GWAS to discover novel loci associated with heart disease, using a CAD Meta-Analysis (122,733 cases and 424,528 controls), and lipid traits, using data from the Global Lipid Genetics Consortium (188,577 subjects). We identified six novel genome-wide significant loci, three which were associated with Triglycerides and CAD, two which were associated with LDL cholesterol and CAD, and one associated with Total Cholesterol and CAD. At several of our loci, the GWAS signals colocalize with expression level associations for one or more genes, indicating that these loci may be affecting disease risk through regulatory activity.

Abstract 25

Carpe D.I.E.M: A Data Integration Expectation Map of Multi-`Omics Data In Complex Disease Disparities

T Tate Hudson, C Williams-DeVane

Department of Biological and Biomedical Sciences, North Carolina Central University, Durham, NC, USA.

Advances in high throughput technologies and the availability of multi-`omics data present the opportunity for more holistic understandings of biological regulation in complex diseases and disparities. The complexity and disparate nature of various diseases requires the development of equally complex models with multiple layers of biological information. This however, requires the integration of biological, computational, and statistical domains. Currently, nonetheless, there exist major gaps in the availability and knowledge amongst the three domains. Typically, biologist experience problems with processing and analyzing biological data; therefore, seeking data scientist for more customized analysis. In contrast, some data scientists lack a thorough understanding of the regulation and complex interactions of various systems giving rise to varying complex phenotypes. This generally results in less comprehensive analysis and an overall narrow understanding of complex disease phenotypes. In this study, we present the Data Integration Expectation Map (D.I.E.M), where we explore the scientific value of integrating various `omic data combinations that can reveal mechanisms of biological regulation in disease disparities. Our goal is to convey the potential for integration of genomic, epigenomic, transcriptomic, proteomic, and metabolomic data for improving our understanding of the complexity and nature of disparity in complex disease traits. In doing so, this map will address the holes in the various domains necessary for integrated data analysis and interpretation. With that, we expect to gain a greater understanding of physiological processes contributing to disparities as well as the role each `omic interaction plays in screening, diagnosis, and prognosis of disease.

Abstract 26

Calculating Overall Biological Process Dysfunction Related to Autism Risk Genes Identifies Clinically-Meaningful Genetic Information

OJ Veatch[1], DR Mazzotti[1], JS Sutcliffe[2], RS Schultz[3], T Abel[4], B Tunc[3], SG Assouline[5], E Brodkin[6], JJ Michaelson[4], TK Nickl-Jockschat[4], ZE Warren[7], BA Malow[8], AI Pack[1]

1. Center for Sleep and Circadian Neurobiology, University of Pennsylvania.
2. Vanderbilt Genetics Institute, Vanderbilt University Medical Center.
3. Center for Autism Research, Children's Hospital of Philadelphia.
4. Department of Psychiatry, University of Iowa.
5. Belin-Blank Center for Gifted Education and Talent Development, University of Iowa.
6. Department of Psychiatry, University of Pennsylvania.
7. Department of Pediatrics, Vanderbilt University Medical Center.
8. Department of Neurology, Vanderbilt University Medical Center.

Autism is a neurodevelopmental condition which encompasses a wide-range of symptom severity. Expression of autism is influenced by inherited common, rare, and de novo genetic variation. It is unclear how genetic evidence is beneficial to understanding variable symptomatology. This is a necessary next step toward translating vast amounts of genetic data into clinically-useful information. De novo variants likely have stronger effects on gene function than inherited variants, but inherited variants may modify the effects of de novo variants. To determine if current evidence could be useful to informing treatment of symptoms in autism, we calculated the likelihood of overall dysfunction in biological processes that are enriched for autism candidate genes. Dysfunction in genes related to cognition distinguished a distinct subgroup of individuals with lower IQs, reduced adaptive behavior and increased social deficits compared to individuals with no evidence of dysfunction in cognition genes. Variation in the PTGS2, ABCA7 and SHANK3 genes was associated with assignment to this subgroup. Particularly, a stop-gain variant in the very important pharmacogene, PTGS2, was associated with having an IQ<70 (i.e., intellectual disability) and increased risk for irritable bowel syndrome. We expect that screening for deleterious variants in the subset of candidate genes involved in cognition may be useful to identifying genetic factors contributing to expression of phenotype differences in autism and help determine genes to prioritize for functional follow-up. This has implications in designing more comprehensive genetic testing panels and may provide the basis for more informed treatment in autism.

Abstract 27

Multiplexed in situ analysis of the human pancreas using imaging mass cytometry

YJ Wang, Daniel Traum, Jonathan Schug, Klaus Kaestner

Department of Genetics, Institute for Diabetes, Obesity and Metabolism,
The Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA.

Autoimmune destruction of the pancreatic beta cells is one of the hallmarks of type 1 diabetes. However, changes in the pancreatic tissue during the progression of the disease are not well characterized. Moreover, the cellular components of the immune system and their interactions with the primary pancreatic tissue have not been systematically investigated. Due to the complexity of type 1 diabetes and the rarity of tissue samples, a system that allows for in situ spatial mapping with multiplexed measurement at a single-cell resolution is highly desirable. The imaging mass cytometry technology enables simultaneous assessment of multiple antibody labeling on the same tissue section at 1 μm resolution. Using this system, we developed a panel with 33 antibodies which allows for combined measurement of cell type markers for the pancreatic, the immune system, stromal components, together with activities of signaling pathways, all directly from a single paraffin section. We established an image analysis pipeline that perform cell-level segmentation, cell-type quantification, and multidimensional analyses. We observed reduced islet size and structural integrity, as well as decreased stromal elements in the pancreatic tissue from type 1 diabetes patients. We also identified increased immune cell density, in particular B cells and CD8+T cells, in the peri-islet region in donors with recent onset of diabetes. We recognized macrophages as the most common immune cell type in all the pancreatic tissue analyzed.

In summary, imaging mass cytometry is a promising novel technology that will give us new insights on the pathophysiology of type 1 diabetes.

Abstract 28

Bulk Tissue Gene Expression Deconvolution Using Single Cell RNA-seq Data

X Wang[1], M Li[2], N Zhang[3]

1. University of Pennsylvania.
2. University of Pennsylvania.
3. University of Pennsylvania.

The identification of cell types within bulk tissue and estimation of their proportions play an important role in the study of disease. Many clinical and population gene expression studies using bulk RNA-sequencing, where the average gene expression across different cell types in the tissue is measured. Deconvolution efforts on bulk gene expression data is difficult in terms of the inferences of cell-type specific proportion and gene expression, especially when the underlying cell types are unknown. Recent advances in single-cell RNA sequencing (scRNA-seq) have led to many studies that profile the transcriptomes within individual cells, allowing for the identification of new cell types and the better characterization of existing cell types. For example, the newly initiated Human Cell Atlas project aims to create comprehensive reference maps of all human cells as a basis for understanding human health. However, scRNA-seq is not well suited to characterizing the proportion of each cell type in a tissue, because the cell dissociation and isolation steps of the procedure are biased in favor of certain cell types. In this study, we explore how such scRNA-seq reference data can be paired with bulk RNA-seq to improve bulk tissue deconvolution. Findings from this analysis will allow us to compare cell type compositions between normal and diseased individuals and examine cell type-specific effect in human diseases.

Abstract 29

Identifying Tissue-Specific Functional Interaction Modules: An Amygdala Imaging Genetic Study

X Yao[1], K Liu[1], J Yan[2], K Nho[2], S Risacher[2], C Greene[3], J Moore[1], A Saykin[2], L Shen[1]

1. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia.
2. Department of Radiology and Imaging Sciences, School of Medicine, Indiana University, Indianapolis.
3. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine , University of Pennsylvania, Philadelphia.

Network-based genome-wide association studies (GWAS) aim to identify functional modules from biological networks that are enriched by top GWAS findings. A majority of module identification studies employ tissue-free networks that lack phenotypic specificity. We employ a novel network-based GWAS (NetWAS) approach to identify modules from a tissue-specific functional network, and demonstrate it in an amygdala imaging genetic studies.

Participants include 989 ADNI subjects. GWAS of amygdala FDG-PET measures were performed to obtain 20,168 gene-level p-values. Amygdala-specific functional network was downloaded from GIANT (http://giant.princeton.edu/). Three NetWAS methods were implemented to reprioritize GWAS results: a previously proposed SVM-based approach that employ significant/nonsignificant status; two regression-based approaches, ridge regression (Ridge) and support vector regression (SVR), which utilize continuous p-values as responses. Genes were reprioritized according to predictions (in Ridge or SVR) or distance from hyperplane (in SVM). AUC of reprioritizations were assessed using documented AD genes as gold standard positives. Link clustering was employed on top reprioritizations to detect modules. Top GWAS findings were used to assess the enrichment of candidate modules, and identify significant ones. Functional annotation was performed on the identified modules. All NetWAS approaches yielded much denser interactions among top findings, and obtained higher AUCs than GWAS. Regression methods outperformed SVM, suggesting that continuous significance measures provide valuable information than binary status. Among top 50 Ridge findings, five modules were identified and enriched by top 50 GWAS findings. These modules were functionally annotated by neurodegenerative diseases, cognition, learning and memory.

Abstract 30

Generalized Integration Model for Improved Statistical Inference by Leveraging External Summary Data

H Zhang[1], L Deng[1], M Schiffman[1], J Qin[2], K Yu[1]

1. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA.
2. National Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, Maryland 20892, USA.

Meta-analysis has become a powerful tool for enhanced inference by gathering evidence from multiple sources. It pools summary-level data from different studies to improve estimating efficiency with the assumption that all participating studies are analyzed under the same statistical model. It is challenging to integrate external summary data calculated from different models with a newly conducted internal study in which individual-level data is collected. We develop a novel statistical inference framework based on the generalized integration model, which effectively synthesizes internal and external information according to their variations for multivariate analysis. The new framework is versatile to incorporate various types of summary data from multiple sources. We establish asymptotic properties for the proposed procedure, and prove that the new estimate is theoretically more efficient than the internal data based maximum likelihood estimate, and the recently developed constraint maximum likelihood estimate that incorporates the outside information. We illustrate an application of our method by evaluating cervical cancer risk using data from a large cervical screening program.

2018 SAGES Organizing Committee

Marcella Devoto, Chair
University of Pennsylvania,
Children's Hospital of Philadelphia

Joan Bailey-Wilson
National Human Genome Research Institute

Barbara Engelhardt
Princeton University

Iuliana Ionita-Laza
Columbia University

Hongzhe Li
University of Pennsylvania

Nandita Mitra
University of Pennsylvania

Adam Naj
University of Pennsylvania

Ingo Ruczinski
Johns Hopkins University

Notes

Notes

Notes