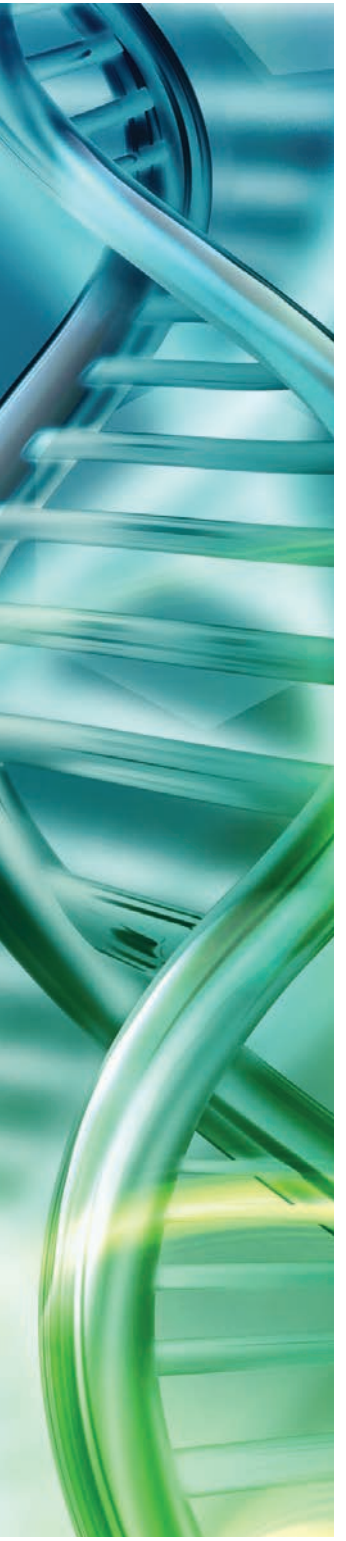
2019 SAGES

Symposium on Advances in
Genomics, Epidemiology & Statistics

Program & Abstract Booklet

Friday, June 7
9:00 a.m. - 6:00 p.m.

CENTER FOR
CLINICAL EPIDEMIOLOGY
AND BIOSTATISTICS



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 National Institute of Environmental Health Sciences, the National Institute on Aging, and the National Center for Advancing Translational Sciences. 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.

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 large population-based and case-control cohorts.

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.


Penn Medicine
Golden Helix
2018 SAGES
Symposium on Advances in Genomics, Epidemiology & Statistics
Friday, June 1, 2018
9:00am - 6:00pm
http://www.med.upenn.edu/sages/
Arthur H. Rubenstein Auditorium
Smilow Center for Translational Research (SCTR)
3400 Civic Center Blvd.

9:00-9:45am	**REGISTRATION AND BREAKFAST**
9:45-10:00am	**Welcome and Opening Remarks** **Marcella Devoto,** *CHOP and University of Pennsylvania*
10:00-11:30am	**SESSION 1** Moderator: Ingo Ruczinski, Johns Hopkins University
10:00-10:30am	**Pushing the boundaries of whole genome sequencing: from genotype to phenotype with a few extras in between** **Rasika Mathias,** *Johns Hopkins University*
10:30-11:00am	**Enhancing Electronic Health Record-derived data with data from secondary sources to address multifactorial problems in real life populations** **Blanca Himes,** *University of Pennsylvania*
11:00-11:30am	**Reframing tumor heterogeneity: gene mapping and precision treatments** **Nicola Camp,** *Huntsman Cancer Institute*
11:30am-1:00pm	**LUNCH**
12:00-1:00pm	**POSTER SESSION 1** (odd numbered posters)
1:00-2:30pm	**SESSION 2** Moderator: Joan Bailey-Wilson, National Human Genome Research Institute
1:00-1:15pm	**Interpretation of deep learning models in genomics: splicing codes as a case study** **Anupama Jha,** *University of Pennsylvania*
1:15-1:30pm	**Statistical methods for multi-condition genetic fine-mapping with applications to eQTL discovery in human tissues** **Gao Wang,** *University of Chicago*
1:30-2:00pm	**The genetic architecture of LDL cholesterol in the Amish** **Braxton Mitchell,** *University of Maryland*
2:00-2:30pm	**Using polygenic risk scores for breast cancer to inform screening: model fit, calibration, and utility** **Peter Kraft,** *Harvard University*
2:30-3:30pm	**COFFEE BREAK AND POSTER SESSION 2** (even numbered posters)
3:30-4:45pm	**SESSION 3** Moderator: Barbara Engelhardt, Princeton University
3:30-3:45pm	**POSTER AWARDS PRESENTATION**
3:45-4:15pm	**Scalable Bayesian multinomial logistic-normal models for the analysis of sequence count data** **Justin Silverman,** *Duke University*
4:15-4:45pm	**Genetic variation and regulation of the 3D genome** **Katie Pollard,** *Gladstone Institutes & UCSF*
4:45-6:00pm	**CONCLUSION AND COCKTAIL RECEPTION**

Poster Numbers & Titles

1	Transcriptomic analysis of coding genes and non-coding RNAs in grain-fed and grass-fed Angus cattle muscle tissue *Y Bai, JA Carrillo, Y He, Y Li, J Song*
2	mtDNA G4 Sequences Associate with Variants and Polymerase Stalling *T Butler, K Estep, J Sommers, R Maul, A Moore, S Bandinelli, L Ferrucci, D Schlessinger, J Ding, R Brosh Jr.*
3	Regulation of Janus kinase 2 by an inflammatory bowel disease causal noncoding SNP *C Cardinale, M March, X Lin, Y Liu, L Spruce, Z Wei, S Seeholzer, S Grant, H Hakonarson*
4	A novel locus identified in chromosome 14 of mouse modulates lens weight *J Cordero, R Williams, L Lu, C Simpson*
5	QTL Remapping of Murine Eye Weight Reveals Novel Candidate Genes for Ocular Growth *R Cordero, R Williams, L Lu , C Simpson*
6	Genome-wide cell-free DNA fragmentation in patients with cancer *S Cristiano, A Leal, J Phallen, J Fiksel, R Scharpf, V Velculescu*
7	Facilitating Analysis of Publicly Available ChIP-Seq Data for Integrative Studies *A Diwadkar, M Kan, BE Himes*
8	Deconvolution of Transcriptional Networks Identifies TCF4 as a Master Regulator in Schizophrenia *A Doostparast Torshizi, C Armoskus, H Zhang, MP Forrest, S Zhang, T Souaiaia, OV Evgrafov, JA Knowles, J Duan, K Wang*
9	Incorporating single-cell RNA-seq data to infer allele-specific expression *J Fan, R Xiao, M Li*
10	Enabling Precision Mitochondrial Medicine through Novel Integration, Visualization, and Complex Analytics of Clinical and Research Data *I George-Sankoh, L MacMullen, D Taylor, B Devkota, R Ganetzky, MJ Falk*
11	Ancestry Clustering and Classification Using an Autoencoder *S Gilhool, P Sleiman, H Hakonarson*
12	The association between African ancestry and telomere length across the African diaspora: evidence from the CAAPA study *K Iyer, M Taub, M Daya, S Chavan, K Barnes, T Beaty, R Mathias*
13	Airway Smooth Muscle-Specific Transcriptomic Signatures of Glucocorticoid Exposure *M Kan, C Koziol-White, M Shumyatcher, M Johnson, W Jester, RA Panettieri, BE Himes*
14	Genomic integrity of human induced pluripotent stem cells across nine studies in the NHLBI NextGen Program *K Kanchan, K Iyer, LR Yanek, MA Taub, C Malley, K Baldwin, L C Becker, U Broeckel, L Cheng, C Cowan, M D'Antonio, KA Frazer, I Carcamo-Orive, JW Knowles, T Quertermous, G Mostoslavsky, G Murphy, M Rabinovitch, DJ Rader, MH Steinberg, E Topol1, W Yang, CE Jaquish, I Ruczinski, RA Mathias*
15	Mediation analysis of alcohol use disorder and alcohol consumption reveals both shared and unique genetic architecture *RL Kember, RV Smith, M Vujkovik, H Zhou, AC Justice, J Gelernter, HR Kranzler*
16	Generalized Meta-Analysis for Combining Disparate Information Across Studies: Inference on Multiple Regression Models *P Kundu, N Chatterjee*
17	Integrative analysis of untranslated regions in human messenger RNAs uncovers G-quadruplexes as constrained regulatory features *D Lee, L Ghanem, Y Barash*

Poster Numbers & Titles

18	Joint Between-Sample Normalization and Differential Expression Detection through L0 Regularized Linear Regression *K Liu, L Shen, H Jiang*
19	Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data *Q Liu, L Fang, G Yu, D Wang, CL Xiao, K Wang*
20	Investigating the Genetic Architecture of Psychiatric Disorders and their Medical Comorbidity *AK Merikangas, RL Kember, K Ruparel, ME Calkins, RC Gur, RE Gur, L Almasy*
21	Defining regulatory variants for SLE susceptibility using an integrative post-GWAS functional genomic framework *J Molineros, L Loooger, C Sun, S Nath*
22	WhatsGNU: a tool for identifying proteomic novelty *AM Moustafa, PJ Planet*
23	Identifying SNP Associations in Under-Powered Whole-Genome Sequencing Association Studies Using eQTLs *JS Ngwa, LR Yanek, K Kammers, MA Taub, RB Scharpf, N Faraday, LC Becker, DM Becker, RA Mathias, I Ruczinski*
24	Comparison of ARIMA, neural networks and hybrid models: Renal failure forecasting in Gaza *S Safi*
25	Predicting Congenital Heart Defect risk from maternal SNPs *B Stear, D Hammond, D Taylor*
26	Modeling metabolic variation with single-cell expression data *Y Zhang, DM Taylor*
27	Genetic analysis of neuroblastoma in African American children *A Testori, Z Vaksman, S Diskin, J Maris, M Devoto*
28	A multi-ethnic genome-wide association study (GWAS) identifies eleven new loci associated with neuroblastoma *Z Vaksman, X Chang, G Lopez, A Modi, H Hakonarson, M Devoto, JM Maris, SJ Diskin*
29	Characterization of Genetic and Phenotypic Heterogeneity of Obstructive Sleep Apnea across Multiple United States Clinics *OJ Veatch, CR Bauer, DR Mazzotti, BT Keenan, JD Robishaw, K Bagai, BA Malow, AI Pack, SA Pendergrass*
30	Signatures in the myeloma transcriptome *RG Waller, MJ Madsen, J Gardner, D Sborov, NJ Camp*
31	Recovery of genetic heterogeneity for single-cell DNA sequencing *C Wu, NR Zhang*
32	Exploring the Genetic Architecture of Autism Spectrum Disorder without Intellectual Disability *J Zhang, A Ghorai, SC Taylor, LS Perez, HC Dow, BN Gehringer, ZL Griffiths, RL Kember, L Almasy, DJ Rader, ES Brodkin, M Bucan*
33	Phen2Gene: Rapid Phenotype Driven Gene Prioritization for Rare Diseases Using Human Phenotype Ontology Terms *M Zhao, L Fang, Y Chen, C Liu , G Lyon, C Weng, K Wang*
34	cTP-net: Prediction of surface protein abundance from single cell transcriptomes by deep neural networks *Z Zhou, C Ye, NR Zhang*


BENSTEIN, MBBCH AUDITORIUM
EXIT
Penn Medicine

Selected Abstracts & Poster Abstracts

Selected Abstract 1

Interpretation of deep learning models in genomics: splicing codes as a case study

A Jha[1], JK Aicher[2], D Singh[1], Y Barash[1,2]

1. Department of Computer and Information Science, School of Engineering, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

The success of deep learning models led to their fast adaptation for genomics tasks such as predicting DNA binding sites of proteins and RNA splicing outcomes. One major limitation of such models though, especially in application for biomedical tasks, is their black box nature, hindering interpretability. A recent promising method to address this limitation is Integrated Gradient (IG), which identifies features associated with prediction for a sample by a deep model. IG works by aggregating the gradients along the inputs that fall on the straight line between a baseline point and the sample of interest. In this work we address several limitations of IG. First, we define a procedure to identify features significantly associated with a specific prediction task such as differentially included exons in the brain. Then, we assess the effect of using different reference point definitions, and replacing the original single linear path used in IG with nonlinear variants. These variants include neighbors path in the original space (O-N-IG) and the hidden space (H-N-IG), and linear path in the hidden space (H-L-IG). Together, our proposed methods for selecting significant features, reference points, and paths for integrated gradients establish a framework to interpret deep learning models for genomic tasks.

Selected Abstract 2

Statistical methods for multi-condition genetic fine-mapping with applications to eQTL discovery in human tissues

G Wang[1], M Stephens[1,2]

1. Department of Human Genetics, University of Chicago.
2. Department of Statistics, University of Chicago.

In recent years, genetic association analysis in human genetics are typically performed for millions of variables (SNPs) over thousands of phenotypes, ranging from complex diseases to molecular phenotypes. Due to the presence of linkage disequilibrium (LD), localizing non-zero effect SNP from many other correlated ones is a hard problem. Particularly when SNPs have non-zero effect in multiple conditions, it is both interesting and challenging to understand whether it is truly pleiotropic effect, or artifact due to LD.

Here we introduce a new Bayesian regression model for genetic fine-mapping, which we call the Sum of Single Effects (SuSiE) model. We also introduce a corresponding new fitting procedure for SuSiE, called Iterative Bayesian Step-wise Selection (IBSS). IBSS algorithm computes a variational approximation to the posterior distribution under the SuSiE model, which is significantly faster than state-of-the-art fine-mapping methods that uses other approaches to approximate posterior. The model structure also yields, by design, independent confidence sets, each designed to capture one association signal, making the results easy to interpret, and ideal for guiding follow-up studies. When extended to performing multivariate analysis, we provide SuSiE a multivariate adaptive shrinkage prior which leverages sharing of genetic effects in data through a rigorous empirical Bayes framework. We applied multivariate SuSiE to cis-eQTL analysis in GTEx data where presence of pleiotropy and linkage is ambiguous. We show our multivariate fine-mapping elucidates genetic architecture of molecular phenotypes across human tissues, and has the potential to be applied to jointly analyze many GWAS traits.


nslational Research
Medicine
Identifying and inducing gene expression patterns underlying cell identity

Transcriptomic analysis of coding genes and non-coding RNAs in grain-fed and grass-fed Angus cattle muscle tissue

Y Bai[1], JA Carrillo[1], Y He[1], Y Li[1], J Song[1]

1. Department of Animal and Avian Sciences, University of Maryland

Beef represents a major dietary component and source of protein in many countries. Different feeding regimens could have influence on the growth and muscle development. Our preliminary studies demonstrated grass-fed cattle produce tender beef with lower total fat and a higher omega3/omega6 ratio than grain-fed ones. Meanwhile the rate of weight gain differed between groups. Although many candidate genes that showed differences in expression between grass-fed and grain-fed cattle. But the regulatory mechanism related to muscle development is needed to further study. In the present study, we used high-throughput RNA sequencing to compare expression profiles of coding and non-coding RNAs from muscle tissues in grain-fed and grass-fed Angus cattle. Expression profiling revealed that 22 lncRNAs, 249 mRNAs and 10 miRNAs had significantly different levels of expressing (FDR < 0.15). Relative to RNA levels in grass-fed, grain-fed had higher expression of 10 lncRNAs, 161 mRNAs and eight miRNAs, and lower expression of 12 lncRNAs, 88 mRNAs and two miRNAs. Functional analysis suggested that the differentially expressed transcripts are involved in biological processes such as mitochondrial respirasome, regulation of metabolic process, oxidation-reduction. Among the differentially expressed genes, 14 mitochondria-related nuclear genes were observed, including *ND1, APOPT1, COX6A2, CKMT2, ALDH6A1, CKMT2, MDH1* and so on. Notably, potential lncRNA-miRNA-mitochondria related nuclear genes interactions were examined. One lncRNA, three miRNAs and two mitochondria related nuclear genes formed three lncRNA lncRNA-miRNA-mitochondria related nuclear genes pathways, suggesting that regulatory pathways of coding and non-coding genes coordinately delineate the differences of muscle development and metabolites between grain-fed and grass-fed beef cattle. The findings of this study will provide deep insight into mechanisms that contribute for the difference beef quality.

Abstract 1

Abstract 2

mtDNA G4 Sequences Associate with Variants and Polymerase Stalling

T Butler[1], K Estep[2], J Sommers[2], R Maul[3], A Moore[4], S Bandinelli[6], L Ferrucci[7], D Schlessinger[8], J Ding[7], R Brosh Jr.[2]

1. Translational Gerontology Branch, National Institute on Aging, Baltimore, MD.
2. Laboratory of Molecular Gerontology, National Institute on Aging, Baltimore, MD.
3. Laboratory of Molecular Biology and Immunology, National Institute on Aging, Baltimore, MD.
4. Translational Gerontology Branch, National Institute on Aging, Baltimore, MD.
6. Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy.
7. Translational Gerontology Branch, National Institute on Aging, Baltimore, MD.
8. Laboratory of Genetics and Genomics, National Institute on Aging, NIH, Baltimore, MD.

Mutations in the mitochondrial (mt) genome are correlated with cancer and aging. G-quadruplexes (G4s), dynamic DNA structures that arise in guanine (G)-rich templates, potentially stall the mt replisome and thereby promote mutagenesis. We used computational analyses of genome sequence data from two Italian cohorts to demonstrate an association between G4s and mt variation. Using the software G4Hunter to predict G4-forming regions in mtDNA, we found statistically significant enrichment of mutations in stable G4 regions, with preferential enrichment of variants in G4 loops. In vitro biochemical data demonstrated that G4s potently block the mt replicative polymerase gamma (Pol γ). Addition of mt replisome-associated factors including TWINKLE helicase and mt single-strand binding protein were unable to stimulate pol γ synthesis through the G4 block; however, the G4-resolving helicase Pif1, known to partly reside in mt, allowed Pol γ to make fully extended product using the G4 template. We showed that mt primase-polymerase PrimPol further catalyzes error-prone nucleotide incorporation into G4 structures, suggesting its involvement in G4 bypass with accompanying increased risk of mutation. However, the high mutation frequency of PrimPol using the mt G4 template was reduced by the presence of Pif1. Altogether, the computational and biochemical approaches indicate that mt point mutations are enriched at stable G4 structures, consistent with replisome stalling at G-quadruplexes and reliance on error-prone DNA synthesis.

Abstract 3

Regulation of Janus kinase 2 by an inflammatory bowel disease causal noncoding SNP

C Cardinale[1], M March[1], X Lin[2], Y Liu[1], L Spruce[3], Z Wei[2], S Seeholzer[3], S Grant[1], H Hakonarson[1]

1. Center for Applied Genomics, Children's Hospital of Philadelphia.
2. Department of Computer Science, New Jersey Institute of Technology.
3. Proteomics Core Facility, Children's Hospital of Philadelphia.

Of the over 200 genetic loci that have been described as conferring a modest risk for developing inflammatory bowel disease, a subset have been pinpointed to an individual, noncoding single nucleotide polymorphism (SNP). In order to illustrate a model mechanism by which a trait-causing noncoding SNP can function, we selected rs1887428, located in the promoter region of the Janus kinase 2 (JAK2) gene for further study. Using affinity purification-mass spectrometry (AP-MS), we determined that the risk/G allele is bound preferentially by the transcription factor (TF) RBPJ, while the protective/C allele is bound preferentially by the homeobox TF CUX1. We constructed subclones of the Jurkat cell line by CRISPR/Cas9 genome editing, which contained risk or protective alleles of the SNP of interest, and analyzed them with transcriptome sequencing. Our results show that CUX1 and RBPJ modify expression of JAK2, and that while rs188748 makes a minor difference in the expression of JAK2, this effect is amplified to yield significant alterations (> 4-fold) in the expression of pathway member STAT5B. In the edited Jurkat subclones, as well as in a CpG methylation association study of human donors, the risk allele of this SNP is associated with increased DNA methylation of the JAK2 promoter. These results have implications for the interpretation of genome wide association study (GWAS) signals, SNP-to-gene assignment, and expression quantitative trait locus (eQTL) studies, because negligible cis effects on RNA expression may associated with significant trans effects and epigenetic modification.

Abstract 4

A novel locus identified in chromosome 14 of mouse modulates lens weight

J Cordero[1], R Williams[1], L Lu[1], C Simpson[1,2]

1. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN.
2. Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, TN.

Abnormalities of the size and shape of the lens will affect ocular refractive power and impair vision. In previous work, we mapped eye size of 700 mice to define quantitative trait loci that modulate eye size and retinal area, but we failed to detect any loci specifically controlling lens mass. Here, we exploited the power of new advanced mapping methods and much improved genotypes to remap our original lens data and to define a novel lens-specific locus. Lens weight was measured in 122 young adult cases from 26 BXD strains and parents-C57BL/6J and DBA/2J. Lens weight was corrected for variance associated with sex and age. In the original study, we used Haley-Knott mapping methods and about 300 markers. In the reanalysis, we exploit GEMMA software with leave-one-chromosome-out scheme as well as 7000 markers that are now integrated into the GeneNetwork2. We uncovered a locus that affects lens weight in the mouse but has no detectable effect on overall eye weight. It maps to chromosome 14 between 58.2-63.5 Mb (LOD 4.8). Of the 84 genes in the QTL region, we identified two candidate genes Fgf9 and Ctsb. We also detected a secondary locus at chromosome 5 (LOD 3.7) which aligns with a locus that controls overall eye weight in mouse and myopia in humans. The chromosome 5 locus therefore has a global influence on both eye and lens whereas the newly discovered locus on chromosome 14 is lens-specific. The discovery of a locus modulating lens weight may contribute to the understanding of genetics of lens development, coordination of ocular growth and disorders caused by structure abnormalities.

Abstract 5

QTL Remapping of Murine Eye Weight Reveals Novel Candidate Genes for Ocular Growth

R Cordero[1], R Williams[1], L Lu[1], C Simpson[1,2]

1. Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN.
2. Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, TN.

The global prevalence of myopia is growing and may affect almost half of the world's population by 2050. A pioneering study of mouse eye weight used interval mapping to locate quantitative trait loci (QTLs) that control normal variation in the architecture of the eye, lens, and retina in laboratory mice and found novel QTLs Eye1 and Eye2. In this study, we increased the sample size, resulting in a 4-fold increase in strains and 16-fold increase in cases. The 11,761 cases representing 112 BXDs and progenitors (C57BL/6J and DBA/2J), had an average age of 200 days (Mean eye weight = 23.9±0.2 mg). Eye weight measurements were corrected by multiple linear regression analysis to statistically control for covariance between eye weight and variables such as body weight, sex, and age. QTL mapping was conducted using 2017 BXD genotypes of the GeneNetwork module employing a linear mixed model (LMM) with leave-one-chromosome-out (LOCO) approach. Candidate genes were compared to known myopia genes in the CREAM consortium study (2018). We performed QTL remapping of adjusted eye weight using the new 2017 BXD Genotypes and found a significant locus on Chr 19 from 53Mb-58Mb with LOD peak of 4.9 at 56.3Mb. Among 68 genes in this locus, we found 3 strong candidate genes *Shoc2*, *Tcf7l2* and *Dclre1a* which are well expressed in the eye and associated with a very significant cis eQTL. A GWAS study by the CREAM consortium implicates TCF7L2 in myopia. Our findings illustrate the power of using enhanced bioinformatics tools and new mouse genotypes in mapping to improve localization of QTLs and identify promising genes.

Abstract 6

Genome-wide cell-free DNA fragmentation in patients with cancer

S Cristiano[1], A Leal[1], J Phallen[1], J Fiksel[1], R Scharpf[1], V Velculescu[1]

1. Johns Hopkins University

The high morbidity and mortality of cancer results from late diagnosis where therapeutic intervention is less effective, yet clinically proven biomarkers to broadly diagnose patients are not widely available. Analyses of cell-free DNA (cfDNA) in blood provide a noninvasive diagnostic avenue for patients with cancer. However, cfDNA analyses have largely focused on targeted sequencing of specific genes. Genome-wide analyses of cfDNA features may increase the resolution of changes in circulating tumor DNA compared to healthy cfDNA and promote more sensitive cancer detection. We developed an approach to analyze fragmentation profiles and cfDNA features across the genome and applied this method to analyze cfDNA from 236 patients with breast, colorectal, lung, ovarian, pancreatic, gastric, or bile duct cancers and 245 healthy individuals. Machine learning incorporating these features resulted in sensitivities of detection from 57% to >99% among seven cancer types at 98% specificity, as well as determine the tissue of origin to a limited number of sites. The results of these analyses highlight important properties of cfDNA and provide a facile approach for early detection of human cancer.

Abstract 7

Facilitating Analysis of Publicly Available ChIP-Seq Data for Integrative Studies

A Diwadkar[1], M Kan[1], BE Himes[1]

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

ChIP-Seq, a technique that allows for in-depth quantification of DNA sequences bound by transcription factors or histones, has been widely used to characterize genome-wide DNA-protein binding induced by specific exposures and conditions. Over 40,000 ChIP-Seq studies of various DNA-binding proteins are available in public repositories. Integrating results of multiple ChIP-Seq datasets offers a cost-effect avenue to identify robust DNA-protein binding sites and determine their cell-type specificity. We developed brocade, a computational pipeline for reproducible analysis of publicly available ChIP-Seq data, that facilitates creation of R markdown reports with information on public datasets downloaded, quality control, and differential binding comparisons. We used brocade to analyze ChIP-Seq datasets of glucocorticoid receptor (GR), a transcription factor that mediates transcriptional response to glucocorticoids, commonly used anti-inflammatory drugs. Specifically, we analyzed ChIP-Seq studies of airway smooth muscle, airway epithelial cells, A549 cells, childhood acute lymphoblastic leukemia cells, and lymphoblastoid cells to identify cell type-specific and global GR binding across the five cell types. Our results demonstrate the utility of the brocade pipeline and identify GR binding sites that may mediate tissue-specific glucocorticoid responses.

Abstract 8

Deconvolution of Transcriptional Networks Identifies TCF4 as a Master Regulator in Schizophrenia

A Doostparast Torshizi[1,2], C Armoskus[3,4], H Zhang[5], MP Forrest[6], S Zhang[5,7], T Souaiaia[3,4], OV Evgrafov[3,4], JA Knowles[3,4], J Duan[5,7], K Wang[1,2,4]

1. Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA.
2. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
3. College of Medicine, SUNY Downstate Medical Center, Brooklyn, NY.
4. Zilkhe Neurogenetic Institute, University of Southern California, Los Angeles, CA.
5. Center for Psychiatric Genetics, North Shore University Health System, Evanston, IL.
6. Department of Physiology, Northwestern University, Chicago, IL.
7. Department of Psychiatry and Behavioral Neurosciences, University of Chicago, Chicago, IL.

Tissue-specific reverse engineering of transcriptional networks has uncovered master regulators (MRs) of cellular networks in various cancers, yet the application of this method to neuropsychiatric disorders is largely unexplored. Here, using RNA-Seq data on postmortem dorsolateral prefrontal cortex (DLPFC) from schizophrenia (SCZ) patients and control subjects, we deconvolved the transcriptional network to identify MRs that mediate expression of target genes. Together with an independent RNA-Seq data on cultured primary neuronal cells derived from olfactory neuroepithelium, we identified TCF4, as one of the top candidate MRs that may be potentially dysregulated in SCZ. We validated the dysregulated TCF4-related transcriptional network through examining the transcription factor binding footprints inferred from human induced pluripotent stem cell (hiPSC)-derived neuronal ATAC-Seq data, as well as direct binding sites obtained from ChIP-seq data in SH-SY5Y cells. The predicted TCF4 transcriptional targets were enriched for genes showing transcriptomic changes upon knockdown of TCF4 in hiPSC-derived neural progenitor cells (NPC) and glutamatergic neurons (Glut_N), based on observations from three separate cell lines. The altered TCF4 gene network perturbations in NPC, as compared to that in Glut_N, was more similar to the expression differences in the TCF4 gene network observed in the DLPFC of individuals with SCZ. Moreover, TCF4-associated gene expression changes in NPC were more enriched than Glut_N for pathways involved in neuronal activity, genome-wide significant SCZ risk genes, and SCZ-associated de novo mutations. Our results suggest that TCF4 may potentially serve as a MR of a gene network that confers susceptibility to SCZ.

Abstract 9

Incorporating single-cell RNA-seq data to infer allele-specific expression

J Fan[1], R Xiao[1], M Li[1]

1. University of Pennsylvania.

Allele-specific expression (ASE) can be quantified by the relative expression of two alleles in a diploid individual, and such expression imbalance may explain phenotypic variation and disease pathophysiology. Existing methods detect ASE using easily obtainable bulk RNA-seq data, a data type that averages out possible heterogeneity in a mixture of different cell types. Since ASE may vary across different cell types, with the recent advance in single-cell RNA sequencing (scRNA-seq) technologies, characterizing ASE at the cell type resolution may help reveal more about the gene regulation. However, scRNA-seq data is costly to generate and noisy with excessive zeros due to transcriptional bursting. Therefore, it is desirable to incorporate information obtained from scRNA-seq data together with bulk data to infer cell type specific ASE. By employing cell type deconvolution and simultaneously modeling of multi-individual information, we are able to detect cell type specific ASE. Extensive simulations indicate that our method performs consistently well under a variety of scenarios.

Abstract 10

Enabling Precision Mitochondrial Medicine through Novel Integration, Visualization, and Complex Analytics of Clinical and Research Data

I George-Sankoh[1, 3], L MacMullen[1] , D Taylor[2,3], B Devkota[3], R Ganetzky[1,2], MJ Falk[1,2]

1. Mitochondrial Medicine Frontier Program, Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia PA.
2. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
3. Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA.

Precision Mitochondrial Medicine will require novel informatics approaches that support intuitive mining of traditional clinical evaluations integrated with objective outcome assessments at both individual patient and cohort levels. To this end, we have developed and launched a novel data management approach that readily collates, integrates, validates, and allows direct clinician visualization and complex analytic capabilities of all clinical and research data generated within The Children's Hospital of Philadelphia (CHOP).

We sought to develop a robust data integration tool that efficiently extracts and unifies updated medical, genomic, clinical and research data collected in all potential domains within a single database server. A key concern was maintaining data integrity without duplication or loss, regular streaming updates, selective accessibility to identified vs deidentified data, and connectivity between the electronic medical record (EPIC) and research databases (REDCap, Excel, OnCore, etc).

Our data integration solution adopts a data warehouse built in Alteryx. Alteryx serves as a data staging warehouse that pulls from all desired data sources to enable sophisticated analytics for supervised and unsupervised models, allowing novel algorithms to be developed by our in-house data integration bioinformatics team that support custom analyses of high dimensionality data. These integrated data are then directly exported to a commercial resource, Tableau, which is hosted in-house in a virtual machine (VM) readily accessible via the Web with password protection by clinicians, scientists, and researchers.

Our ultimate goal is to enable prioritization of precision medical care and personalized clinical trial outcome measures that leverage direct clinician mining. This will be aided by machine learning approaches to predict mitochondrial diagnosis, prognoses, biomarkers, and therapeutic response based on complex arrays of molecular, biochemical and clinical outcomes.

Ancestry Clustering and Classification Using an Autoencode

S Gilhool[1], P Sleiman[1,2], H Hakonarson[1,2]

1. Center for Applied Genomics, Children's Hospital of Philadelphia.
2. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania.

It is well known that allele frequencies vary between populations of different ancestry. Therefore, it is useful to have an effective means of grouping subjects by ancestry. This is often done by performing a principal component analysis (PCA), and manually separating groups based on visual inspection. We present a complementary approach to cluster individuals based on ancestry through the use of an autoencoder. An autoencoder is a type of deep neural network operates by reconstructing its input data after passing it through a low-dimensional space that serves as a bottleneck. The fact that the data can be reconstructed implies that its more compact representation at the bottleneck is a meaningful way to encode the input data. We used the autoencoder to determine low-dimensional representations of SNP array data for thousands of patients in the CHOP biobank. The results from the autoencoder exhibit clear clustering by ancestry. We further used the autoencoder representations to classify samples by ancestry.

Abstract 11

Abstract 12

The association between African ancestry and telomere length across the African diaspora: evidence from the CAAPA study

K Iyer[1], M Taub[2], M Daya[3], S Chavan[3], K Barnes[3], T Beaty[1], R Mathias[1,4] on behalf of CAAPA Consortium

1. Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.
2. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD.
3. Department of Medicine, University of Colorado, Denver, CO.
4. Department of Medicine, Johns Hopkins University, Baltimore, MD.

Over the past two decades, telomere length (TL) has emerged as a marker of biological aging. This biomarker is highly heritable, reflects gender dysmorphism and is also influenced by race/ethnicity where some studies reported that people of African ancestry harbor longer TL than those of European ancestry, other studies observed the opposite trend and yet others found no differences. However, to our knowledge, association between the admixture components and TL has not been explored.

Objective: Determine the association between the estimated ancestry proportions and TL by leveraging existing whole genome sequence (WGS) data from a multi-centered case-control study - Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA).

Methods: CAAPA recruited N=726 African-admixed individuals from North, Central, South America, Caribbean Islands and Yoruba-speaking individuals from Ibadan, Nigeria. Global estimates of ancestry were obtained using ADMIXTURE. TL on each individual was estimated using a bioinformatic approach–TelSeq. Linear regression models were implemented to assess the nature of relationship between TL and African ancestry (%YRI) across the African diaspora after adjusting for Native American ancestry (%NAT), age, sex and asthma case status.

Results: Increased %YRI and %NAT admixture were significantly associated with increased TL after adjusting for age and sex. Additional adjustments for asthma case status did not change the association.

Conclusion & Future work: Our preliminary analysis suggest that admixture proportions influence TL in people of African descent. As CAAPA genomes were collected across 16 different sites, we are also examining the possibility of technical sources of variation that may have been inevitably introduced in our TL data and could confound our associations.

Airway Smooth Muscle-Specific Transcriptomic Signatures of Glucocorticoid Exposure

M Kan[1], C Koziol-White[2], M Shumyatcher[1], M Johnson[2], W Jester[2], RA Panettieri Jr[2], BE Himes[1]

1. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA
2. Rutgers Institute for Translational Medicine and Science, Rutgers University, State University of New Jersey, New Brunswick, NJ.

Glucocorticoids, commonly used asthma controller medications, decrease symptoms in most patients, but some remain symptomatic despite high dose treatment. The physiological basis underlying glucocorticoid response, especially among asthma patients with severe, refractory disease is not fully understood. We sought to identify differences between fatal asthma and non-asthma donor-derived airway smooth muscle (ASM) cell transcriptomic response to glucocorticoid exposure, and to compare ASM-specific changes to those of other cell types. In cells derived from 9 fatal asthma and 8 non-asthma donors, RNA-Seq was used to measure ASM transcriptome changes after exposure to budesonide (100nM 24hr) or control vehicle (DMSO). Differential expression results were obtained for this dataset, as well as 13 publicly available glucocorticoid response transcriptomic datasets corresponding to 7 cell types. Specific genes were differentially expressed in response to glucocorticoid exposure: 7,835 and 6,957 in non-asthma and fatal asthma donor-derived ASM cells, respectively (adjusted p-value <0.05). Transcriptomic changes in response to glucocorticoid exposure were similar in fatal asthma and non-asthma donor-derived ASM, with enriched ontological pathways that included cytokine- and chemokine-related categories. Comparison of glucocorticoid-induced changes of the non-asthma ASM transcriptome to that of 6 other cell types showed that ASM has a distinct glucocorticoid response signature that is also present in fatal asthma donor-derived ASM cells.

Abstract 13

Abstract 14

Genomic integrity of human induced pluripotent stem cells across nine studies in the NHLBI NextGen Program

K Kanchan[1], K Iyer[1], LR Yanek[1], MA Taub[2], C Malley[1], K Baldwin[3], LC Becker[1], U Broeckel[4], L Cheng[1], C Cowan[5], M D'Antonio[6], KA Frazer[6], I Carcamo-Orive[7], JW Knowles[7], T Quertermous[7], G Mostoslavsky[8], G Murphy[8], M Rabinovitch[9], DJ Rader[10], MH Steinberg[11], E Topol[12], W Yang[13], CE Jaquish[14], I Ruczinski[2], RA Mathias[1]

1. Dept. of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD.
2. Dept. of Biostatistics, Bloomberg School of Public health, Johns Hopkins University, Baltimore, MD.
3. Dept. of Molecular and Cellular Neuroscience, Dorris Neuroscience Center, The Scripps Research Institute, La Jolla, CA, USA.
4. Dept. of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA.
5. Div. of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA.
6. Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA.
7. Stanford University School of Medicine, Stanford Center for InheritedCardiovascular Disease, Stanford University, Stanford, CA.
8. The Center for Regenerative Medicine, Boston Medical Center, Boston University School of Medicine, Boston, MA, USA.
9. Dept. of Pediatrics, Stanford University, Stanford, CA, USA
10. Dept. of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
11. Dept. of Medicine, Section of Hematology-Oncology, Boston University School of Medicine, Boston, MA.
12. Dept. of Molecular Medicine, The Scripps Research Institute, La Jolla, CA.
13. Penn Center for Pulmonary Biology and Institute for Regenerative Medicine, University of Pennsylvania, Philadelphia, PA.
14. National Heart, Lung, and Blood Institute, NIH, Bethesda, MD, USA.

Prior studies suggest that human induced pluripotent stem cell (hiPSC) lines suffer from genomic instabilities (e.g. karyotypic abnormalities, chromosomal aberrations). Within the NHLBI sponsored NextGen program, hiPSC lines were generated from a range of sources (PBMCs, fibroblasts and lung epithelium) at nine sites to study the complex genetics of metabolic, cardiovascular and respiratory diseases. The objective of this work is to assess the genomic integrity of the NextGen hiPSCs. We examined the integrity of hiPSC lines by comparing to their matched donor DNA from blood/PBMCs leveraging 1.3 million genetic variants from the Illumina Multiethnic Genotyping (MEGA) array. Two levels of genomic integrity were investigated using: (1) genotype discordance between 1,060 parent donor DNA and hiPSC lines; and (2) structural variability of 506 hiPSC lines using copy number variants (CNVs). We detected low rates of genotype discordance (median = 0.002%) between the donor and hiPSC lines; and observed that 149 hiPSC lines acquired 258 CNVs relative to the donor DNA. The cumulative impact per hiPSC line was small; 85% showing less than 2 Mb of cumulative CNV coverage. Furthermore, we identified six recurrent regions of CNVs on chromosomes 1, 2, 3, 16 and 20 that overlapped with cancer associated genes. In general, hiPSCs acquired deletions included tumor suppressor genes whereas duplications included oncogenes. Overall, a low level of genomic instability was observed in the NextGen generated hiPSC lines as compared to prior reports. However, given the observation of structural instability in regions with known cancer associated genes, our results substantiate the importance of systematic evaluation of genetic variations in hiPSCs before using them as disease/research models.our results substantiate the importance of systematic evaluation of genetic variations in hiPSCs before using them as disease/research models.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.

Abstract 15

Mediation analysis of alcohol use disorder and alcohol consumption reveals both shared and unique genetic architecture

RL Kember[1], RV Smith[2], M Vujkovik[1], H Zhou[3], AC Justice[3], J Gelernter[3], HR Kranzler on behalf of the VA Million Veteran Program[1]

1. University of Pennsylvania
2. University of Louisville
3. Yale School of Medicine

Recent GWAS of alcohol use have uncovered key differences in the underlying genetic architectures of alcohol consumption and alcohol use disorder (AUD), with the latter correlated with psychiatric disorders. We used the large sample size and longitudinal data available in the Million Veteran Program to evaluate the mediating effect of alcohol consumption (measured by the AUDIT-C) on the genetic risk of AUD. To examine direct and indirect SNP effects on phenotype, we extracted all European-ancestry patients with genotype data and at least one measure of AUDIT-C. For the subset with AUD, we retained only patients with an AUDIT-C measure prior to diagnosis. If a patient had multiple AUDIT-C measures, the maximum score was analyzed. We conducted three GWAS on this dataset: AUD case (n=18,049)/control (n=115,317), maximum AUDIT-C (n=133,366), and AUD with maximum AUDIT-C as a covariate. The well-established protective variant in ADH1B, rs1229984, was significant in all three GWASs, showing both a direct effect on AUD and an indirect effect on AUD through alcohol consumption. Other SNPs with direct effects on AUD were identified in genes ZBTB16 (p=3.2x10-8), DRD2 (p=1.5x10-7), and ADAMTS1 (p=3.3x10-6), among others. SNPs with direct effects on AUD tended to be in or near genes which are brain expressed and/or previously associated with psychiatric phenotypes. SNPs which were mediated by alcohol consumption were in the alcohol dehydrogenases (ADH1C: p=5.9x10-3; ADH6: p=6.0x10-3) or associated with peripheral phenotypes such as cholesterol level. Delineating the genetic architectures of AUD and AUDIT-C measures could differentiate loci which contribute to AUD but not alcohol consumption and thereby lead to a better understanding of disease pathways that are amenable to treatment interventions.

Abstract 16

Generalized Meta-Analysis for Combining Disparate Information Across Studies : Inference on Multiple Regression Models

P Kundu[1], N Chatterjee[1]

1. Johns Hopkins University School of Public Health.

Data integration is a process of fusing information from multiple data sources, giving a unified way to draw inference on real world problems. The objective here is to describe recent methods for inferring parameters associated with rich models using disparate information available from multiple studies. Meta-analysis, because of both logistical convenience and statistical efficiency, is widely popular for synthesizing information on common parameters of interest across multiple studies. We will describe a generalization of the meta-analysis that allows estimation of parameters associated with a multiple regression model through meta-analysis of studies which may individually have information only on partial sets of the regressors. An application of the method will be illustrated through a real data example involving the development of a breast cancer risk prediction model using disparate risk factor information from multiple studies. Further, we will show how the proposed framework can be used for the efficient analysis of data from two-phase epidemiologic designs that create disparate covariate information across the phases by design.

Abstract 17

Integrative analysis of untranslated regions in human messenger RNAs uncovers G-quadruplexes as constrained regulatory features

D Lee[1], L Ghanem[2], Y Barash[3]

1. Graduate Group in Computational Biology and Genomics, University of Pennsylvania.
2. Children's Hospital of Philadelphia.
3. Department of Genetics, University of Pennsylvania.

Identifying regulatory elements in the noncoding genome is a fundamental challenge in biology. G-quadruplex (G4) sequences are abundant in untranslated regions (UTRs) of human messenger RNAs, but their functional importance remains unclear. By integrating multiple sources of genetic and genomic data, we show that putative G-quadruplex forming sequences (pG4) in 5' and 3' UTRs are selectively constrained, and enriched for cis-eQTLs and RNA-binding protein (RBP) interactions. Using over 15,000 whole-genome sequences, we uncover patterns of selection at single-nucleotide resolution in UTR pG4s supporting their capacity for secondary-structure formation. In parallel, we identify new proteins with evidence for preferential binding at pG4s from ENCODE annotations, and delineate putative regulatory networks composed of shared binding targets. Finally, by mapping variants in the NIH GWAS Catalogue and ClinVar, we find enrichment for disease-associated variation in 3'UTR pG4s. At a GWAS pG4-variant associated with hypertension in HSPB7, we uncover robust allelic imbalance in GTEx RNA-seq across multiple tissues, suggesting that changes in gene expression associated with pG4 disruption underlie the observed phenotypic association. Taken together, our results establish UTR G-quadruplexes as important cis-regulatory features, and point to a putative link between disruption within UTR pG4 and susceptibility to human disease.

Abstract 18

Joint Between-Sample Normalization and Differential Expression Detection through L0 Regularized Linear Regression

K Liu[1], L Shen[1], H Jiang[2]

1. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA.
2. Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.

A fundamental task for RNA-seq data analysis is to determine whether the RNA-seq read counts for a gene or exon are significantly different across experimental conditions. Since the RNA-seq measurements are relative in nature, between-sample normalization of counts is an essential step in differential expression (DE) analysis. In most existing methods the normalization step is independent of DE analysis, which is not well justified since ideally normalization should be based on non-DE genes only. Recently, Jiang and Zhan proposed a robust statistical model for joint between-sample normalization and DE analysis from log-transformed RNA-seq data. Sample-specific normalization factors are modeled as unknown parameters in the gene-wise linear models, and the L0 penalty is introduced to induce sparsity in the regression coefficients. In their model, the experimental conditions are assumed to be categorical (e.g., 0 for control and 1 for case). In this work, Jiang and Zhan's model is generalized to accommodate continuous/numerical experimental conditions, and a linear regression model is used to detect genes for which the expression level is significantly predicted by any experimental conditions or a particular experimental condition. Furthermore, an efficient algorithm is developed to solve for the global solution of the resultant high-dimensional, non-convex and non-differentiable penalized least squares regression problem. Extensive simulation studies and a real RNA-seq data example show that when a large proportion (e.g., >30%) of genes are differentially expressed in an asymmetric manner, the proposed method outperforms existing methods and the performance gain is more substantial as the sample size increases.

Abstract 19

Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data

Q Liu[1], L Fang[1], G Yu[2,3], D Wang[3], CL Xiao[2], K Wang[1,4]

1. Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA USA.
2. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China.
3. Grandomics Biosciences, Beijing 102200, China.
4. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

DNA base modifications, such as C5-methylcytosine (5mC) and N6-methyldeoxyadenosine (6mA), are important types of epigenetic regulations. Short-read bisulfite sequencing and long-read PacBio sequencing have inherent limitations to detect DNA modifications. Here, using raw electric signals of Oxford Nanopore long-read sequencing data, we design DeepMod, a bidirectional recurrent neural network (RNN) with long short-term memory (LSTM) to detect DNA modifications. We sequence a human genome HX1 and a Chlamydomonas reinhardtii genome using Nanopore sequencing, and then evaluate DeepMod on three types of genomes (Escherichia coli, Chlamydomonas reinhardtii and human genomes). For 5mC detection, DeepMod achieves average precision up to 0.99 for both synthetically introduced and naturally occurring modifications. For 6mA detection, DeepMod achieves ~0.9 average precision on Escherichia coli data, and have improved performance than existing methods on Chlamydomonas reinhardtii data. In conclusion, DeepMod performs well for genome-scale detection of DNA modifications and will facilitate epigenetic analysis on diverse species.

Abstract 20

Investigating the Genetic Architecture of Psychiatric Disorders and their Medical Comorbidity

AK Merikangas[1], RL Kember[2,3], K Ruparel[4], ME Calkins[4], RC Gur[4], RE Gur[4], L Almasy[1,2]

1. Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA.
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
3. Crescenz VA Medical Center, Philadelphia, PA.
4. Department of Psychiatry, Neuropsychiatry Section, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Background: Previous work has demonstrated pervasive comorbidity between medical and psychiatric disorders in both adults and youth. Here we examine these patterns of comorbidity across development and evaluate the extent to which they may result from common underlying genetic factors.

Method: The sample includes 5175 European-ancestry youth (ages 8-21 years; 51.7% female) from the University of Pennsylvania Neurodevelopmental Cohort study sampled from pediatric clinics at the Children's Hospital of Philadelphia.
Medical conditions were derived from interview data and medical record information. Psychiatric disorders were assessed with a structured diagnostic interview. Polygenic Risk Scores (PRS) were calculated using PRSice2 software package and publicly available GWAS. Sex and age-adjusted logistic regression models were used to evaluate the associations between PRS for medical and psychiatric disorders.

Results: Specific associations emerged between: ear nose and throat disorders with psychosis symptoms; central nervous system disorders with Attention Deficit Hyperactivity Disorder (ADHD), and behavior disorders and general psychopathology; whereas developmental disorders were associated with a broad range of psychiatric disorders. The overall medical disorder severity rating was associated with all of the predicted anxiety, ADHD, behavior disorders, mood disorders, and overall psychopathology. The ADHD PRS was associated with ADHD and behavior disorders, the MDD PRS was associated with mood disorders, but no other PRS showed associations with the disorders examined here.

Discussion: These findings demonstrate strong overlap between medical and psychiatric conditions in youth, but common genetic risk factors do not appear to underlie this comorbidity. Other mechanisms, including environmental factors, may influence these associations.

Abstract 21

Defining regulatory variants for SLE susceptibility using an integrative post-GWAS functional genomic framework

J Molineros[1], L Loooger[2], C Sun[1], S Nath[1]

1. Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA.
2. Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA, USA.

Multiple association studies have uncovered numerous genetic associations for systemic lupus erythematosus (SLE), an autoimmune disease with a strong genetic component. However the identity of the true causal variants linking their target gene(s) are largely unknown. Correct identification of susceptibility genes, relevant tissues and involved cell types is crucial for identifying pathogenic mechanisms and ultimately for designing therapeutics.

We performed a comprehensive analysis using reported SLE loci, immune cell-specific gene expression and 3D chromatin interaction datasets to define predisposing variants and their target genes. We collated 308 reported SLE variants and 12,070 correlated ($r^2>0.7$) variants to define 122 independent signals. We identified 60 deleterious coding SNPs (23 loci; CADD>12.37), and further 2,308 eQTLs correlated with 190 genes in monocytes, neutrophils, CD4+T, CD8+T, and B cells. We identified eQTLs in gene promoters for 60 genes, and in enhancers (from promoter-enhancer interactions) for 173 genes. Overall, target genes were overrepresented in 169 immune related pathways including Interferon gamma signaling ($P=7.09\times10^{-15}$). Additionally, 14 genes from 12 independent signals that exhibited aberrant isoform ratios, including SYNGR1 where rs61616683 risk allele increased 7-fold expression of non-sense mediated decay transcript.

By leveraging immune cell expression, epigenetic information, and 3D interaction data, we identified functional SNPs and SLE genes at 77 independent SLE signals. The 90.3% of functional SNPs collocated with enhancers, followed by promoters, and deleterious SNPs. Together, we identified 190 target susceptibility SLE genes from 69 independent signals expressed in immune cell types that were overrepresented in immune pathways.

Abstract 22

WhatsGNU: a tool for identifying proteomic novelty

AM Moustafa[1], PJ Planet[1, 2, 3]

1. Division of Pediatric Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA, USA.
2. Department of Pediatrics, Perelman College of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
3. Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA.

For many bacterial species, thousands of whole genome sequences are now available in public databases. To understand diversity and novelty in such enormous collections we need computationally scalable tools that can quickly contextualize genomes based on their similarities to known variation and identify features of each genome that make them unique. Here we present a tool based on exact match proteomic compression that, in seconds, delivers a whole proteome report classifying any new genome to the strain level, and provides a detailed report of protein variants that may have novel functional differences. The technique utilizes the natural variation in public databases to rank protein sequences based on the number of observed exact protein matches (the Gene Novelty Unit (GNU) score) in all known genomes. The GNU score affords a convenient way to measure known protein diversity, describe protein conservation, identify the closest match genomes, and, most importantly, assay for novelty. We use this technique to characterize the panproteome of Mycobacterium tuberculosis, Pseudomonas aeruginosa and Staphylococcus aureus, and we provide a tool, WhatsGNU, that can be used to quickly create whole protein reports and classify genomes. We suggest that this technique could be extended to most bacterial species for which large numbers of genomes are currently available. WhatsGNU is available from https://github.com/ahmedmagds/WhatsGNU.

Abstract 23

Identifying SNP Associations in Under-Powered Whole-Genome Sequencing Association Studies Using eQTLs

JS Ngwa[1], LR Yanek[2], K Kammers[3], MA Taub[1], RB Scharpf[3], N Faraday[4], LC Becker[2], DM Becker[2], RA Mathias[2], I Ruczinski[1]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.
2. Department of Medicine, Johns Hopkins School of Medicine.
3. Department of Oncology, Johns Hopkins School of Medicine.
4. Department of Anesthesiology and Critical Care Medicine, Johns Hopkins School of Medicine.

GWAS studies have successfully identified thousands of SNPs associated with complex traits, however identifying the functional elements through which these genetic variants exert their effects remains a critical challenge. Recently, there is increasing evidence that SNPs associated with complex traits are more likely to be expression quantitative trait loci (eQTLs). Thus, incorporating eQTL information can potentially improve power in highlighting causal genes. Our goal was to investigate the potential to detect novel risk loci among eQTLs only. Our data comprised of nine platelet aggregation traits from the GeneSTAR study, with whole genome sequencing (WGS) family data in European Americans (EA) and African Americans (AA). RNA-seq data were generated from extracted non-ribosomal RNA from 185 megakaryocyte (MKs) cell lines and 290 platelet samples. This included iPSC-derived MKs on 84 AA and 101 EA subjects as well as platelets on 110 AA and 180 EA subjects. We fit a linear mixed model for genetic association on each population, including covariates, ancestry information captured in principal components and random effects from genetic relatedness. We conducted fixed effects meta-analysis using summary statistics from EA and AA populations. Since eQTLs typically exhibit very strong patterns of linkage disequilibrium, we performed permutation analysis using 1000 permutations for all nine traits to obtain family-wise error rates for eQTL SNPs, substantially lowering the genome-wide significance threshold compared to the standard Bonferroni procedure. Our analyses confirmed known risk loci such as *PEAR1*, *ADRA2A* and *ARHGEF3*. A number of novel genetic loci associated with platelet aggregation were also identified: *ADAM22*, *APIP*, *ARAP2*, *BANF2*, *C6orf195*, *CBLN2*, *CEP68*, *CTNNA1*, *GPR98*, *GTF2IRD1*, *HIVEP2*, *IMPG2*, *LOC642236*, *MACROD2*, *NT5C1B-RDH14*, *PI4KAP1*, *RAB1A*, *RPTOR*, *SENP7*, *SLC1A4* and *TMEM120B*.

Abstract 24

Comparison of ARIMA, neural networks and hybrid models: Renal failure forecasting in Gaza

Samir Safi[1]

1. The Islamic University of Gaza.

For time series, the problem that statisticians often face is how to choose the appropriate time series model for forecasting its future values. Using individual linear or nonlinear model is insufficient in modeling and forecasting the time series, because these models most likely contain both linear and nonlinear patterns. In this study, we demonstrate how a hybrid forecasting model that combines ARIMA with artificial neural network (ANN) can be used for forecasting number of renal failure in Gaza. The proposed model considers the linear and nonlinear patterns in the real data of renal failure cases simultaneously so that it brightens up the time series better. The empirical results on the renal failure in Gaza show that the forecasting performances by the hybrid model are superior to those of ARIMA model and ANNs models based on different measures of forecasting accuracy.

Predicting Congenital Heart Defect risk from maternal SNPs

B Stear[1], D Hammond[2], D Taylor[2]

1. Drexel University.
2. Children's Hospital of Philadelphia.

Congenital heart defects (CHD) are the leading cause of infant mortality affecting 1% of all newborns. Despite advancements in pediatric cardiovascular research and medical care, early detection and prenatal screening remain the most effective ways to increase CHD survivability. Studies estimate that between 28-58% of pregnancies are terminated when severe CHD is found. For these reasons we applied machine learning to predict if a mother is at risk of having a child with CHD. The algorithms we included in the analysis are logistic regression (LR), random forest (RF) and support vector machine (SVM). We trained and tested the models on a large single nucleotide polymorphism (SNP) count matrix which was obtained from whole exome sequencing data. The models were evaluated on different subsets of the data matrix which was filtered based on varying minimum allele frequency (MAF) thresholds. To evaluate the performance of the models we report the classification accuracy and the area under the curve (AUC) of the receiver operator characteristic. The models achieved accuracies and AUCs on the different subsets of data between 75% and 80%.

Abstract 26

Modeling metabolic variation with single-cell expression data

Y Zhang[1,2], DM Taylor[1,3]

1. Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA.
2. Department of Genetics, Rutgers University, Piscataway, NJ, USA.
3. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Cellular metabolism is known to play critical roles in developmental regulation and progression. However, metabolic profiles at the single cell are difficult to obtain with current technologies. We sought to explore metabolic changes at different stages of development by using single-cell RNA-seq expression data as informative for cell metabolic state. We present metabolic models based on published single-cell expression data from different developmental mouse tissues at five distinct time points (9.5-13.5d). We separate single-cell profiles by their main mapped developmental trajectory, sub trajectory, development day, and cell type, and perform analyses using the Cost Optimization Reaction Dependency Assessment (CORDA) method, generating genome-scale metabolic models. By optimizing these models for key metabolic reaction products, we are able to detect previous empirically observed trends in metabolic regulation. We present results and comparisons of expression and metabolism at different cell types and stages across mouse tissues and developmental times.

Genetic analysis of neuroblastoma in African American children

A Testori[1,2], Z Vaksman[2], S Diskin[2,3], J Maris[2,3], M Devoto[2,3]

1. Dept. of Molecular Medicine and Medical Biotechnologies, University of Naples Federico II, Naples, Italy.
2. Children's Hospital of Philadelphia, Philadelphia, PA, USA.
3. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Neuroblastoma (NB), a pediatric cancer with a high degree of clinical heterogeneity, is rarer in African-American (AA) than European-American (EA) children. AA children with NB, however, more frequently develop the high-risk form of NB and have lower survival. We have identified several loci associated to NB by GWAS in EA children, but few of them replicated in AAs. Our expanded AA cohort of 674 cases and 3113 controls allowed us to investigate the genetic susceptibility of NB in this population. Following genotyping and imputation, we confirmed one NB susceptibility gene (BARD1), which reached genome-wide significance in the high-risk AA cases. A polygenic score including all SNPs with $p < 1.55x10\text{-}6$ in the EA GWAS, showed significant association ($p = 2.7x10\text{-}11$), and explained ~3% of NB risk variance in AAs. However, significance of the polygenic score dropped with inclusion of additional SNPs, suggesting limited sharing of NB genetic risk factors between EAs and AAs, or a genetic architecture of NB with limited contribution from common SNPs. We also used admixture mapping to test whether NB risk variants are located in genomic regions showing different proportions of African and European ancestry in AA cases versus controls. We detected a signal on chromosome 6 near the HACE1-LIN28B risk locus, where cases show increased European ancestry. A region upstream of TP53, another known NB risk gene, showed increased African ancestry in high-risk cases. Variants at this locus may help explain susceptibility to the high-risk form of NB that disproportionately affects AA children with NB.

Abstract 27

Abstract 28

A multi-ethnic genome-wide association study (GWAS) identifies eleven new loci associated with neuroblastoma

Z Vaksman[1], X Chang[1], G Lopez[1], A Modi[1], H Hakonarson[2-4], M Devoto[2,5,6], JM Maris[1], SJ Diskin[1-3,7,8]

1. Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA.
2. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
3. Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA.
4. Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, PA.
5. University of Rome "La Sapienza", Rome, Italy.
6. Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
7. Genomics and Computational Biology, Biomedical Graduate Studies, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.
8. Abramson Family Cancer Research Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA.

Neuroblastoma (NBL) is the most commonly diagnosed pediatric extracranial solid tumor and accounts for 12% of childhood cancer deaths. Over the past decade, GWAS have identified over a dozen NBL susceptibility loci, but have necessarily focused on patients of European ancestry. Here, we report results from a large multi-ethnic GWAS encompassing 6,393 NBL cases and 39,569 controls: European Americans (3,788 cases, 22,052 controls), African American (794 cases, 5,854 controls) and Hispanic individuals (1,118 cases and 1,707 controls). We first identified eleven new loci associated with NBL in the European cohort (prange: 1.6x10-9 to 3.0x10-7). Considering these novel loci together with those previously reported, twelve regions replicated in African Americans (8 loci; prange: 1.6x10-6 to 0.04) and/or Hispanics (11 loci; prange: 1.2x10-8 to 0.03). Phenotypic assessment of the loci was done using methylation quantitative trait loci (meQTL) and expression quantitative trait loci (eQTL) analysis (GTeX). A strong meQTL and eQTL was observed at 1p36.22 (rs2480775, pme=6.42x10-32; pe=4.90x10-6 with PEX14), 16q12.1 (rs12930079, pme=4.22x10-135, pe=1.42x10-79 with *HEATR3*) and 14q11.2 (rs2273626, pme=1.68x10-19, pe=5.40x10-6 with HOUS4) among others. The most robustly replicated region was at the *BARD1* locus on chromosome 2q35 (rs34358404; pmeta=6.33x10-47). Here, we observed significant meQTL (pme=1.88x10-16) and eQTL (pe=6.60x10-30) for BARD1, and the rs34358404 risk allele was associated with an increased number of double strand breaks (p=1.33x10-4) and chromosome 17 translocation events (p=1.02x10-7) in primary tumors. Taken together, these results substantially extend the number of NBL susceptibility loci known, provide the first assessment of genetic risk in Hispanic patients, and implicate homologous recombination deficiencies in NBL tumorigenesis.

Characterization of Genetic and Phenotypic Heterogeneity of Obstructive Sleep Apnea across Multiple United States Clinics

Abstract 29

OJ Veatch[1, 2], CR Bauer[3], DR Mazzotti[1], BT Keenan[1], JD Robishaw[4], K Bagai[2], BA Malow[2], AI Pack[1], SA Pendergrass[3]

1. Center for Sleep and Circadian Neurobiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.
2. Sleep Disorders Division, Department of Neurology, Vanderbilt University Medical Center, Nashville, TN.
3. Geisinger, Rockville, MD.
4. Department of Biomedical Science, Charles E. Schmidt College of Medicine, Florida Atlantic University, Boca Raton, FL.

Obstructive sleep apnea (OSA) is a heterogeneous disorder defined by variable expressivity of core symptoms, risk factors and comorbidities. OSA is linked to multiple negative health outcomes, making identification of effective approaches for treatment an important area of research. Many genes implicated in OSA overlap with those associated with risk factors and comorbidities. Understanding the shared genetic mechanisms contributing to expression of symptoms and comorbidities in OSA may inform personalized treatment.

We leveraged electronic health records (EHRs) and biorepositories available at Geisinger (n=318,236), Vanderbilt University Medical Center (n=36,986), and the eMERGE Network (n=83,717) to conduct one of the largest OSA studies to date. We examined associations between 51 candidate variants (SNPs) identified through systematic literature review and EHR-derived phenotypic algorithms used to define OSA. We conducted phenome-wide association studies to identify genetic relationships between OSA and comorbidities.

Eight candidate SNPs were associated with OSA, seven with sleep architecture and respiratory indices, and 16 with other EHR-derived diseases (e.g., cardiovascular disease, diabetes, anemia). The most prevalent comorbidity was essential hypertension (EH). Patients with EH had 4-fold higher odds of having OSA compared to those without. Notably, 74% of patients with both conditions had their first EH code at least 5 years before their first OSA code.

Results characterize robust OSA-associated genomic variants, and convergent mechanisms influencing risk for multiple disorders in the same individual. This could provide the basis for detecting undiagnosed OSA in the EHR and informing more personalized treatments of OSA and related comorbidities.

Abstract 30

Signatures in the myeloma transcriptome

RG Waller[1], MJ Madsen[1], J Gardner[1], D Sborov[1], NJ Camp[1]

1. University of Utah School of Medicine.

Clinical management and research of multiple myeloma (MM) is impeded by tumor heterogeneity. A standard approach to deconstruct tumor heterogeneity using RNA sequence data is hierarchical clustering to determine mutually exclusive categorical subtypes. However, categorical subtypes are unidimensional and may fail to capture potential important variation. An alternate approach is to establish tumor phenotypes that capture the total transcriptome variation as quantitative expression dimensions using principal component analysis (PCA). The dimensions are orthogonal – each dimension is an independent tumor characteristic, a linear combination of the representative genes. For each patient, a transcriptome signature, based on the dimension values in their tumor can be determined. Associations between individual dimensions or multi-dimension signatures and clinical outcomes can be explored. We hypothesize a quantitative framework for MM tumors will uncover biologically relevant components of tumors and may also reflect specific molecular liabilities and therapeutic vulnerabilities. RNA sequencing on treatment-naïve, CD138 sorted, tumor cells was obtained on 768 patients in the Clinical Outcomes on MM Genetic Profiles Assessment study. SALMON gene-based expression counts were normalized for gene length, library size, and RNA composition. Multi-stage PCA was performed on the normalized counts to derive orthogonal, quantitative tumor dimensions. Future work will associate the quantitative dimensions with demographic, clinical, and genetic (germline and somatic) characteristics using penalized linear regression modeling. In sum, we present a new quantitative framework to represent expression variability in MM tumors that provides more flexibility for statistical modeling with clinical relevant endpoints and ultimately the potential to improve precision cancer care.

Abstract 31

Recovery of genetic heterogeneity for single-cell DNA sequencing

C Wu[1], NR Zhang[1]

1. University of Pennsylvania, Philadelphia, PA.

Genetic heterogeneity in tumor cells indicates the molecular mechanisms underlying tumor evolutionary dynamics. Intratumor heterogeneity is commonly studied at the clonal level where a “clone” is usually defined by DNA mutations such as single nucleotide variations (SNVs). Recently developed single-cell DNA-sequencing (scDNA-seq) techniques enable SNV detection at the single-cell level. However, the sequencing data are low-coverage and noisy, and computational methods are required to address these issues. Currently, mutation detection methods developed for scDNA-seq data are mostly ineffective or impractical. In this study, we are developing computational methods to denoise scDNA-seq data using an unsupervised neural network called autoencoder. In our method, we borrow the raw allele frequency of alternative alleles across all detected SNV sites in all cells to impute the missing genotypes and denoise all positions. The results show that the autoencoder outperforms two existing methods in terms of SNV signal estimation. We found that the autoencoder was able to recover SNV signals in varying degrees under different simulated scenarios. When the sequencing coverage is low, the autoencoder performs better with larger numbers of cells and loci as predictors, and they are robust to inclusion of false positive SNV loci in the data. Furthermore, the SNV signals estimated by the proposed method allows the recovery of underlying clonal structures. Our results reveals the potential of low dimensional latent space methods in SNV signal recovery for scDNA-seq data, which would facilitate more understanding of intratumor heterogeneity and the underlying molecular mechanisms.

Abstract 32

Exploring the Genetic Architecture of Autism Spectrum Disorder without Intellectual Disability

J Zhang[1], A Ghorai[2], SC Taylor[3], LS Perez[4], HC Dow[4], BN Gehringer[4], ZL Griffiths[4], RL Kember[2], L Almasy[*2,5], DJ Rader[*2], ES Brodkin[*6], M Bucan[*2]

1. Graduate Group in Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA.
2. Department of Genetics, University of Pennsylvania, Philadelphia, PA.
3. Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA.
4. University of Pennsylvania, Philadelphia, PA.
5. Children's Hospital of Philadelphia, Philadelphia, PA.
6. Department of Psychiatry, University of Pennsylvania, Philadelphia, PA.

* Project leaders

Autism Spectrum Disorder (ASD) is a group of heterogeneous disorders including individuals with and without intellectual disability (ID). Genetic factors represent over 50% of ASD risk. The genetic architecture of ASD is complex, and involves common and rare, inherited and de novo single nucleotide variations (SNVs) and copy number variations (CNVs). The Autism Spectrum Program of Excellence (ASPE) at Penn is focusing on the genetic roots of ASD without ID. Prior efforts to understand the genetic etiology of ASD without ID have been limited. ASPE is recruiting individuals with ASD without ID (probands) and their families, collecting deep phenotype and whole genome sequence (WGS) on all family members. Currently, 94 probands and 146 relatives have been recruited. 57 probands and 35 relatives have been sequenced. Standard quality control procedures were performed on ASPE WGS data. Major global ancestry groups were identified using principal component analysis. We then performed the relatedness check and analysis of runs of homozygosity. To evaluate the contribution of common genetic variants to the risk of ASD, we calculated ASD polygenic risk score (PRS). Compared to relatives, the mean ASD PRS is elevated in ASPE probands. To evaluate the contribution of rare genetic variants to the risk of ASD, CNVs, and SNVs have been called and annotated. We detected inherited and de novo CNVs and pathogenic and likely pathogenic SNVs in known ASD genes (e.g. NRXN1, SHANK3) in ASPE participants. By focusing on ASD without ID as a more homogeneous group, ASPE complements existing ASD genetic studies and provides insights into the genetic architecture of ASD without ID.

Abstract 33

Phen2Gene: Rapid Phenotype Driven Gene Prioritization for Rare Diseases Using Human Phenotype Ontology Terms

M Zhao[1], L Fang[1], Y Chen[1], C Liu[2] , G Lyon[3] , C Weng[2] , K Wang[1,4]

1. Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia.
2. Department of Biomedical Informatics, Columbia University Medical Center.
3. Institute for Basic Research in Developmental Disabilities (IBR).
4. Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine.

Incorporation of clinical phenotypes can greatly facilitate interpretation of whole genome or exome sequencing data to prioritize candidate genes for patients with Mendelian diseases. Human Phenotype Ontology (HPO) terms are increasingly used in diagnostic settings to characterize the clinical phenotypes of patients, so methods that leverage HPO terms can improve diagnostic sequence integration. Although existing HPO annotations map each phenotype term to multiple candidate genes, they do not provide a means to weight/rank the genes or combine multiple phenotype terms for each patient. To address these challenges, here we first compile a knowledgebase (HPO2Gene-KB) that associates each HPO term with a list of prioritized genes, each with a weight indicating the strength of the relationships. We next present Phen2Gene, a rapid computational approach that searches and prioritizes candidate genes from phenotype information. Given a patient with suspected Mendelian diseases, Phen2Gene takes a list of HPO terms as input, pre-processes the HPO terms to handle nested relationships and different levels of specificity, and calculates ranked gene scores for the patient. Compared to existing tools such as Phenolyzer, Phen2Gene has comparable performance but generates results instantly with much faster speed. Thus, Phen2Gene can serve as a real-time phenotype driven gene prioritization tool to aid clinical diagnosis in rare diseases.

Abstract 34

cTP-net: Prediction of surface protein abundance from single cell transcriptomes by deep neural networks

Z Zhou[1], C Ye[2], NR Zhang [3]

1. Graduate Group in Genomics and Computational Biology. University of Pennsylvania.
2. School of Medicine, Tsinghua University.
3. Department of Statistics, The Wharton School, University of Pennsylvania.

Recent developments of CITE-seq and REAP-seq technologies allow quantification of single cell transcriptome and surface proteins abundance in the same cell, which are both important features to characterize cell states and label cell population, especially in immune cells. However, many single cell studies, including Human Cell Atlas project, quantify the transcriptome only, and do not have cell-matched measurements of protein markers of interest. Here we propose a framework, single cell transcriptome to protein prediction with neural network (cTP-net), which harness a multi branched deep neural network to accurately predict single-cell surface protein relative abundances from scRNA-seq data. We benchmark cTP-net's prediction on a diverse testbed of immune cell populations, and show that it achieves correlation > 0.9 for the 10 proteins examined in cross-validation. We also found cTP-net to have good generalization power, retaining high accuracy in tissues, cell types and technologies that differ from, but related to, the training data. We generate predictions of 12 cell surface proteins for 270,000 cord blood mononuclear cells (CBMC) that were deposited in the Human Cell Atlas portal and illustrate that multimodal data analysis by cTP-net can achieve a more detailed characterization of cellular phenotypes than transcriptome measurement alone.

Notes

Notes

Notes

Notes

Notes

2019 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

Adam Naj
University of Pennsylvania

Ingo Ruczinski
Johns Hopkins University