

2017 SAGES

Symposium on Advances in
Genomics, Epidemiology & Statistics

ABSTRACT BOOKLET

Friday, June 9

9:00 a.m. - 6:00 p.m.

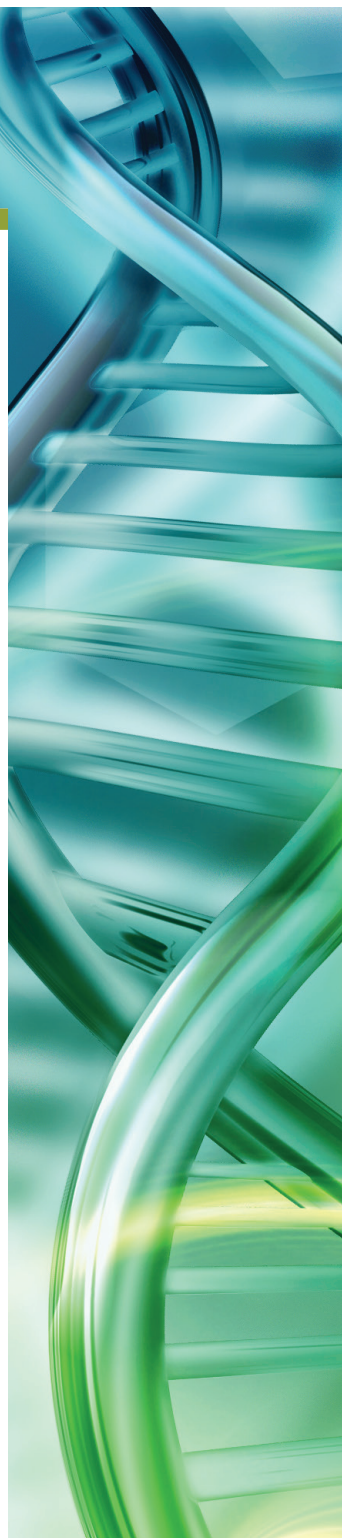
Arthur H. Rubenstein Auditorium
Smilow Center for Translational Research
3400 Civic Center Blvd.



CENTER FOR
CLINICAL EPIDEMIOLOGY
AND BIostatISTICS




CGACT [atgctaggatctatacatcagcactcgcgcga](https://www.cgact.org)
Center for Genetics and Complex Traits
[atgctaggatctatacatagtagctcgcgcagtcta](https://www.cgact.org)



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). Additional funding was provided by CHOP Genes, Genomes, and Pediatric Disease (GGPD) Research Affinity Group.

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.



Abstract 1

Genetic risk factors for PANDAS and Sydenham's chorea, two pediatric autoimmune diseases of the brain following Streptococcus infection

Iliir Agalliu¹, Michael V. Gonzalez², Frank D. Mentch², Dritan Agalliu³, Jennifer Frankovich⁴, Hakon Hakonarson² and Tyler Cutforth³

1. Dept. of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, 10461.
2. Center for Applied Genomics, Children's Hospital of Pennsylvania, Abramson Research Center, Philadelphia, PA, 19104.
3. Dept. of Neurology, Columbia University Medical Center, New York, NY, 10032.
4. PANS Clinic and Research Program & Dept. of Pediatric Allergy, Immunology and Rheumatology, Stanford University School of Medicine, Stanford, CA, 94305.

A recent meta-analysis of ten childhood autoimmune diseases, including type I diabetes and juvenile idiopathic arthritis, revealed a shared genetic susceptibility architecture that implicates signaling pathways for helper T cells, JAK-STAT, interferon and interleukin (Li et al., 2015); more than 70% of the identified loci are shared among three or more of these diseases. However, PANDAS/SC were not included in this analysis, due to their infectious trigger and complex diagnosis.

We are undertaking the first prospective and retrospective approach to assemble a cohort of PANDAS/SC patients and controls in order to perform a genome-wide association study (GWAS) and whole-exome sequencing, to discover genetic risk factors and pathways implicated in these diseases. The Children's Hospital of Philadelphia (CHOP) has identified 20 Sydenham's chorea and 28 PANDAS patients, plus several hundred potential cases whose clinical records strongly suggest a PANDAS diagnosis; medical records of these patients will be reviewed by experienced clinicians to validate the diagnosis. Added to these are ~170 PANDAS/PANS cases from Stanford who were recruited at their clinic, to which we will add patients from New York City (CUMC, AECOM). Age, sex- and ethnicity-matched controls will be drawn from the cohort of healthy control children in the CHOP database. We will compare cases and controls to identify genetic variants and biological pathways associated with PANDAS. Genes and pathways identified in other autoimmune pathologies, plus the Th17 cell and blood-brain barrier mechanisms implicated from animal models, should prove useful for a more focused as opposed to a hypothesis-free, genome-wide approach.

Integrative analysis identifies immune-related enhancers and lncRNAs perturbed by genetic variants associated with Alzheimer's disease

Alexandre Amlie-Wolf^{1,2}, Mitchell Tang¹, Jessica King¹, Beth Dombroski¹, Yi-Fan Chou¹, Elizabeth Mlynarski¹, Gerard D. Schellenberg¹, Li-San Wang^{1,2}

1. Dept. of Pathology & Laboratory Medicine, University of Pennsylvania.
2. Genomics & Computational Biology Graduate Group, University of Pennsylvania.

We developed the INFERNO (INFERRing the molecular mechanisms of NONcoding genetic variants) tool to analyze non-protein-coding genetic signals associated with late-onset Alzheimer's disease (AD). We defined sets of variants in LD with any locus-wide significant variant and overlapped these expanded variant sets with enhancers from 112 FANTOM5 tissue facets and 127 Roadmap tissues and cell types and quantified their effects on transcription factor binding sites (TFBSs), revealing enhancer dysregulation in all 19 tag regions from IGAP. Using a unified tissue categorization to harmonize data sources identified a significant enrichment of enhancer overlaps in the blood/immune and connective tissue categories. To identify the affected target genes, we performed co-localization analysis of the GWAS signals with GTEx eQTL data across 44 tissues. This identified strongly co-localized eQTL signals in 15 tag regions, 9 of which contained variants overlapping enhancers from the same tissue class as the eQTL signal. In 6 of these 9 tag regions, we prioritized individual variants that disrupted or created TFBSs, and in 5 of them, we prioritized variants with high probabilities of individually underlying the co-localization signals. Both approaches identified a strong signal in the *EPHA1* region targeting the *EPHA1-ASI* long noncoding RNA (lncRNA) which was validated by luciferase assay. We identified similarly affected lncRNAs in several tag regions, and expression correlation showed that they regulated several aspects of the immune response, which has been previously implicated in AD pathogenesis. These results demonstrate the power of the principled integration of functional genomics data to characterize noncoding genetic signals.

Abstract 2

Abstract 3

SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data

Rhonda Bacher¹, Li-Fang Chu², Ning Leng², Audrey P. Gasch³, James A. Thomson², Ron M. Stewart², Michael Newton^{1,4}, and Christina Kendzierski⁴

1. Dept. of Statistics, UW-Madison, Madison, WI.
2. Morgridge Institute for Research, Madison, WI.
3. Laboratory of Genetics, UW-Madison, Madison, WI.
4. Dept. of Biostatistics and Medical Informatics, UW-Madison, Madison, WI.

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies including an abundance of zeros, varying levels of technical bias across gene groups, and systematic variation in the effects of sequencing depth. The normalization methods traditionally used in bulk RNA-seq were not designed to accommodate these features and, consequently, applying them to the single-cell setting results in artifacts that bias downstream analyses. To address this, we developed SCnorm to enable efficient and accurate scRNA-seq normalization. Simulation and case study results suggest that the framework provides for increased accuracy in fold-change estimation as well as improvements in downstream inference.

The missing landscape of human genomic diversity in the Arabian Peninsula

Njlal Bakhsh¹, Latifa Jackson¹, Christopher Cross¹, Fatimah Jackson¹

1. Howard University, Washington, DC.

The Arabian Peninsula (AP) is the first site of human migration and habitation outside of Africa. As a major crossroad for human populations, the AP provides an opportunity to better understand early to modern changes in human demographic patterns through selections, admixture, gene flow, and migration. Dramatic climatic fluctuations have been recorded in the AP that contributed to contractions and expansions in water availability. These climatological perturbations are thought to have shaped genomic variations in this population. Recent reports indicate that a number of Arab nation-states have committed significant resources to genetically typing the national population, with the overall goal of determining the degree of genomic diversity in the AP. We sought to characterize currently typed genomic variation in Arabian populations to support the rationale for our proposed analyses of Saudi Arabian genomic diversity. Interestingly, in contrast to published claims, a comprehensive search of peer-reviewed reports on genomic analysis (N=20 papers) revealed no genomic data from four national genomic projects (Qatar, Saudi Arabia, Kuwait, and The United Arab Emirates). Our analysis demonstrates that while much fanfare and presumably resources have been devoted to defining the genomic landscape of the Arabian peoples, little actual data is available to either substantiate or support such an investment.

Abstract 4

Abstract 5

Regulation of Keratinocyte Gene Expression by the Skin Microbiome

Casey Bartow-McKenney¹, Jackie Meisel¹, Joseph Horwinski¹, Elizabeth Grice¹

1. Dept. of Dermatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

The skin microbiome represents a milieu of microorganismal communities adapted to their host in both composition and physiological potential, demonstrating a coevolutionary relationship that inhabits one of the largest human organs. However, cutaneous host-microbiome relationships of the skin remain poorly characterized, including the extent to which the microbiota modulate host genetic control. Our preliminary data from murine whole skin (dermis + epidermis) demonstrated that 2,820 genes were differentially expressed (DE) between germ-free (GF) and conventionally raised (CR) mice. Because epidermal keratinocytes are the predominant cells in contact with skin microbiota, here we explore the epidermal-specific host response to microbial colonization by utilizing RNA-seq on keratinocytes isolated from GF, CR, and conventionalized (CV) mice, a model for acute colonization. Our preliminary analyses found the expression profiles of GF and CR mice to be most distinct, with over 6,000 DE genes, thus recapitulating our previous findings. Additionally, genes that were DE between all three colonization states exhibit a pattern whereby average expression in CV mice falls significantly between the average expression levels of GF and CR mice, demonstrating a microbial-induced shift in CV keratinocyte transcription after colonization. Moreover, CR mice were enriched in gene ontology attributes such as cellular adhesion and development of the skin barrier and epidermis, indicating a more physiological response. A further understanding of this host-microbiome relationship may aid in identifying etiological causes and putative amelioration of skin disorders such as psoriasis and atopic dermatitis, which have been previously associated with shifts in the composition of the skin microbiome.

Linear Discriminant Analysis Predicts Extension in Patients with Juvenile Idiopathic Arthritis

AC Brescia¹, MMSimonds², SM McCahan², HT Bunnell², KE Sullivan³, CD Rose¹

1. Nemours/AI DuPont Hospital for Children, Wilmington DE.
2. Nemours Biomedical Research, Wilmington, DE.
3. Immunology, Children's Hospital of Philadelphia, University of Pennsylvania.

Purpose: Our goal is to develop predictive synovial biomarkers to identify which children with oligoarticular juvenile idiopathic arthritis (JIA) will have persistent course (PR) (<4 involved joints) vs extended course (E) (>5 affected joints after 6 months of disease), before the two courses can be distinguished clinically.

Methods: Synovial fluid from arthrocenteses was used to establish primary cultures. RNA from cultured passage 3-6 fibroblast-like synoviocytes (FLS) were isolated, amplified and hybridized to Affymetrix Human GeneChips. Global gene expression of FLS from 12 PR and 11 E samples were obtained. Data was filtered for log₂ expression >4 in all samples of either E or PR, then for |1.5|-fold change. LIMMA revealed 83 probesets with differential expression between E vs PR FLS (7% FDR).

Results: Hierarchical clustering revealed like samples cluster together. Of these 83 probesets, 9 had secreted proteins. Linear discriminant analysis modeling on these 9 revealed 6 genes (KLHL13, MAMLD1, ANKRD44, CD14, HSPBAP1, and MBP) which could correctly predict group, E or PR, 100% of the time using leave-one-out cross validation. Expression of secreted proteins was confirmed in synovial fluid by ELISA.

Conclusion: FLS from JIA patients who remained persistent vs those who will extend have different transcriptomes. E samples preceded extension in the majority of patients, highlighting that there are detectable differences in the transcriptome of the FLS early in course. The differentially expressed genes, especially for secreted proteins, provide a starting point for development of biomarkers to distinguish between PR and E JIA using aspirated synovial fluid.

Abstract 6

Abstract 7

Cloudy With A Chance of Reproducibility: Interactive Quality Control for Genetic Studies

Brian S. Cole, PhD¹ and Jason H. Moore, PhD¹

1. Institute for Biomedical Informatics, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania.

Genome-wide Association Studies (GWAS) have revealed associations between human genetic variants and hundreds of traits, including susceptibility to complex diseases. GWAS relies critically on rigorous Quality Control (QC) procedures to eliminate known sources of false positive results and ensure that downstream statistical assumptions are met. However, GWAS QC pipelines are difficult to reproduce due to a lack of any mechanism to ensure consistency of a QC pipeline across different study sites. We developed CLINK, an interactive and self-contained GWAS QC virtual appliance. CLINK combines public data, open source software, and cloud computing services to provide an entirely cloud-based interactive workflow for GWAS QC which includes all steps needed including exploratory and high-resolution graphics and results reporting. CLINK provides reproducibility by combining a notebook-based, interactive workflow with cloud imaging and templating technology, enabling complete reproducibility across time and space. Importantly, CLINK uses stateless firewalls, encryption, and secure socket protocols for security. CLINK does not require any command-line knowledge, exposing itself as a notebook which can be safely copied and cloned without retaining sensitive data. By incorporating cloud automation technology, we demonstrate that CLINK can also be used from wireless mobile devices such as cell phones and tablets. We demonstrate the utility of CLINK by performing fully cloud-based QC and analysis of a large case-control study of primary open-angle glaucoma.

Defiant: (DMRs: Easy, Fast, Identification and ANnotation) Identifies Differentially Methylated Regions from Iron-Deficient Rat Hippocampus

David E. Condon¹, Phu V. Tran¹, Yu-Chin Lien¹, Jonathan Schug¹, Michael K. Georgieff², Rebecca A. Simmons¹, Kyoung-Jae Won¹

1. Perelman School of Medicine, University of Pennsylvania, Philadelphia PA.

Background: Identification of differentially methylated regions (DMRs) is the initial step towards the study of DNA methylation-mediated gene regulation. Previous approaches to call DMRs suffer from false prediction, use extreme resources, and/or require tedious preprocessing.

Results: We developed a new approach called “Defiant” to identify DMRs. Employing Weighted Welch Expansion (WWE), Defiant showed superior performance to other predictors in the series of benchmarking tests on artificial and real data. Defiant was subsequently used to investigate DNA methylation changes in iron-deficient rat hippocampus. Defiant identified DMRs close to genes associated with neuronal development and plasticity, which were not identified by its competitor. Importantly, Defiant runs between 5 to 479 times faster than currently available software packages. Accepting diverse input formats for DNA methylation data, Defiant does not require additional preprocessing. Defiant works for whole-genome bisulfite sequencing (WGBS), reduced-representation bisulfite sequencing (RRBS), Tet-assisted bisulfite sequencing (TAB-seq), and HpaII tiny fragment enrichment by ligation-mediated PCR-tag (HELP) assays.

Abstract 8

Abstract 9

A Bayesian Allele Specific Expression Model for Large Scale Genetic Expression studies using a Sparse Overdispersed Poisson Generalized Linear Models

Genna Gliner¹, YoSon Park², Brigitte Lamarche³, Christopher Brown², Barbara E Engelhardt^{3,4}

1. Operations Research and Financial Engineering Dept., Princeton University, Princeton, NJ, 08540, USA.
2. Dept. of Genetics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, 19104, USA.
3. Computer Science Dept., Princeton University, Princeton, NJ, 08540, USA.
4. Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, 08540, USA.

Studying the relationship between genetic regulatory elements and phenotypic expression through allele specific expression (ASE) has proven to be a powerful method to study the genetic origins of phenotypic variation. We present a Bayesian model for allele specific expression that applies an Overdispersed Poisson Linear Model (OPLM) to recover genetic variants using haplotypes from phased genotype data and mapped read counts from RNA-seq data. Current ASE detection methods use the relative abundance of expression between two alleles at a genetic variant in heterozygous individuals to search for statistically significant associations at genetic variants, throwing away relevant data across tissues, samples, and genes. The OPLM includes data from all individuals across all sites while explicitly modeling relatedness between individuals through the overdispersion term. This allows flexibility in the model to apply sparsity inducing priors on the coefficients of the predictors that explicitly include phasing accuracy and linkage disequilibrium across the haplotype sites. We demonstrate our model on the Genotype Tissue Expression (GTEx) consortium dataset to identify ASE loci across tissues and show how our model improves upon existing ASE detection methods.

Bayesian Hierarchical Modeling of Genic Sub-Region Intolerance

Tristan J. Hayeck¹, Nicholas Stong¹, David Goldstein¹, Andrew Allen^{1,2}

1. Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA.
2. Dept. of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA.

Understanding the local distribution of intolerance within genes will be key in correctly deciphering relationships between genetic variation and disease. Frequently, pathogenic mutations can cluster in particular sub-regions of a disease gene, while other benign mutations may occur in other parts of the same gene. Previous methods have had success describing regional intolerance using marginal measures of departure from the amount of common function variance given the total genic variation. We expand on this concept, developing a hierarchical Bayesian model framework for analysis of genomic variation, jointly modeling genes and their sub-regions (domain or exon). The Bayesian hierarchical framework allows for direct inference on the joint posterior distribution of gene level and sub-region level effects and allows for identification of genes with high levels of variation in regional intolerance. Using standing human variation from the Exome Aggregation Consortium (ExAC), we demonstrate classification of intolerant sub-regions in key genes such as MECP2, VHL, and SCN1A. We further assess our ability to identify regions of the genome harboring pathogenic disease mutations through comparison of our regional intolerance scores and the presence of disease mutations from ClinVar and HgMD mutation databases. Initial results demonstrate significant association between sub-region intolerance scores and the presence of pathogenic variants in ClinVar and HgMD. These results suggest our framework can provide improved classification of intolerant regions and improve diagnostic interpretation of both known and novel variation.

Abstract 10

Abstract 11

Violence and Allostatic load in African American young adults

Latifa Jackson^{1,2}, Max Shestov³, Forough Saadatmand², Joseph Wright²

1. National Human Genome Center, College of Medicine, Howard University, Washington DC.
2. Pediatrics, College of Medicine, Howard University, Washington DC.
3. Genomics and Computational Biology, Perelman College of Medicine University of Pennsylvania, Philadelphia, PA, USA.

Allostatic load is the cumulative wear and tear experienced by the immune system in response to chronic environmental stressors. Many studies have observed increased allostatic load in African American populations in comparison to their European American counterparts. This difference in allostatic load has been attributed to the inherent stressors associated with socioeconomic status and race. A key environmental stressor in the lives of young African Americans is the occurrence of violence. We wanted to understand the effect of violence and mental health perturbations on allostatic load. Understanding how the experience of violence contributes to stress biomarkers is a critical step in parsing and ultimately reducing its effect on African American young adults. We study a cohort of 557 young African American adults aged 18-25 years old (females N= 274, males N=283) from the Washington DC area. Study participants were surveyed with respect to environmental determinants both in their childhood and as adults. We found that both perceived mental health and violence were correlated to elevated stress biomarkers. When Epstein Barr Virus viral capsid antigen IgM was compared to violence features characterized in the BADU dataset, we found that internalization of environmental stressors were most strongly correlated to elevated allostatic load markers. This work suggests that internalization of experienced violence may be as important as the actual violence experience.

Integrative Deep Models for Alternative Splicing

Anupama Jha¹, Matthew R Gazzara^{1,2,3}, Yoseph Barash^{1,2,*}

1. Department of Computer and Information Science, School of Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
3. Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

Advancements in sequencing technologies have highlighted the role of alternative splicing (AS) in increasing transcriptome complexity. This role of AS, combined with the relation of aberrant splicing to malignant states, motivated two streams of research, experimental and computational. The first involves a myriad of techniques such as RNA-Seq and CLIP-Seq to identify splicing regulators and their putative targets. The second involves probabilistic models, also known as splicing codes, which infer regulatory mechanisms and predict splicing outcome directly from genomic sequence. To date, these models have utilized only expression data. In this work we address two related challenges: Can we improve on previous models for AS outcome prediction and can we integrate additional sources of data to improve predictions for AS regulatory factors. We perform a detailed comparison of two previous modeling approaches, Bayesian and Deep Neural networks, dissecting the confounding effects of datasets and target functions. We then develop a new target function for AS prediction and show that it significantly improves model accuracy. Next, we develop a modeling framework to incorporate CLIP-Seq, knockdown and over-expression experiments, which are inherently noisy and suffer from missing values. Using several datasets involving key splice factors in mouse brain, muscle and heart we demonstrate both the prediction improvements and biological insights offered by our new models. Overall, the framework we propose offers a scalable integrative solution to improve splicing code modeling as vast amounts of relevant genomic data become available.

Abstract 13

eQTL analysis of megakaryocytes derived from induced pluripotent stem cells

K. Kammers¹, M.A. Taub², I. Ruczinski², J. Martin³, L.R. Yanek³, A. Frazee², Y. Gao⁴, D. Hoyle⁴, N. Faraday³, D.M. Becker³, L. Cheng⁴, Z.Z. Wang⁴, J.T. Leek², L.C. Becker³, R.A. Mathias³

1. Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Johns Hopkins University School of Medicine, Baltimore, MD
2. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.
3. The GeneSTAR Research Program, Johns Hopkins School of Medicine, Baltimore, MD.
4. Division of Hematology and Institute for Cell Engineering, Johns Hopkins School of Medicine, Baltimore, MD.

Understanding the biology of platelet aggregation is important to prevent inappropriate vascular thrombosis. GWAS studies have identified common variants associated with platelet aggregation, but because they are intronic or intergenic, their biological link to platelet function is unclear. To explore potential function of genetic variants in this context, we have produced pluripotent stem cells (iPSCs) from mono-nuclear cells in subjects from our parent platelet aggregation studies, and then derived megakaryocytes (MKs), the precursor cells for anucleate platelets, from the iPSCs to determine patterns of gene expression in the MKs related to specific genetic variants.

For 198 MK cell lines we generated genotype data on the Illumina 1M GWAS array with 1,003,451 SNPs, and RNA-seq data from extracted non-ribosomal RNA. Including iPSC-derived MKs on 108 European American (EA) and 90 African American (AA) subjects, we identify a total of $N=7,190$ and $N=525$ cis-eQTLs in the two racial groups ($q<0.05$), respectively. Given unequal capture of common variation on GWAS arrays between African and European ancestry subjects as represented here, we rely on a distance-based replication scheme to look for eQTL replication between the two groups (replication at a distance of 5kb). Of the 7,190 cis-eQTLs discovered in the EAs, 328 were replicated in AAs ($q<0.05$), and of the 525 cis-eQTLs discovered in the AAs, 261 were replicated in the EAs ($q<0.05$). A high number of the detected cis-eQTLs (38%) are unique to MKs compared to 44 other tissue types that are reported in the latest version of the GTEx Project.

Genetic and Phenotypic Heterogeneity of Mood Disorders in a Large Multigenerational Pedigree

Rachel L. Kember¹, Liping Hou², Xiao Ji¹, Lars H. Andersen³, Lisa N. Estrella³, Francis J. McMahon², Christopher D. Brown¹, Maja Bucan¹

1. Dept. of Genetics, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, USA.
2. Human Genetics Branch, National Institute of Mental Health Intramural Research Program, National Institutes of Health, Bethesda, MD, USA.
3. Lancaster General Health/Penn Medicine, University of Pennsylvania Health System, Lancaster, PA, USA.

Bipolar disorder (BD) is a mental disorder characterized by alternating periods of depression and mania. Individuals with BD have higher levels of early mortality than the general population, and a substantial proportion of this may be linked to increased risk for co-morbid diseases. Medical co-morbidity may be a consequence of either chromosomal proximity of a BD-risk gene to a gene underlying non-psychiatric phenotype, or the pleiotropic effect of a risk-allele. In order to identify the molecular events that underlie BD and related medical co-morbidities, we imputed whole genome sequence (WGS) for an extended multigenerational Old Order Amish pedigree segregating BD and related disorders. We mapped disease causing variants at known Mendelian loci and performed genomic profiling using polygenic risk scores (PRS) for several complex diseases, allowing us to genetically 'diagnose' all family members. To explore the contribution of disease genes to BD we performed gene-based and variant-based association tests for BD, and found that Mendelian disease genes are enriched in the top results from both tests. We identified disease causing variants in cardiovascular genes which are found at a higher frequency in individuals with BD, including a variant in APOB associated with hypercholesterolemia. Finally, we examined PRS for a number of traits and found that higher PRS for lipid traits and diabetes were associated with BD in this pedigree. Taken together, our results indicate that medical co-morbidity between complex diseases and Mendelian disorders arises as a combination of chromosomal proximity of disease causing variants and pleiotropy of disease genes.

Modified Random Forest for Trio Data with Alternative Splitting Criterion to allow for Missing Genotypes

Qing Li¹, Emily Holzinger¹, Joan E. Bailey-Wilson¹

1. Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Baltimore, MD

Random forests (RF) is an ensemble method, which analyzes data and summarizes results using a large number of classification or regression trees. It has proven useful method in detecting complex gene-gene interactions associated with a trait in case-control samples. The case-parent trio design is an efficient family-based study design which is robust to population stratification. Previously, we proposed a modification of the RF algorithm for trio data analysis with appropriate sampling approach. For ease of implementation, our method utilizes the rpart package to conduct classification tree (CART) analysis. In this work, we implemented alternative splitting criterion within CART to account for the relationship across bootstrapped samples. At each node of the CART approach, only samples with genotype data are included in the analysis, eliminating the need to impute all missing genotypes. To account for linkage disequilibrium among markers, we employed a binning method to divide the nearby or biologically relevant markers into sets, and only one marker is selected from this set to be used in the CART analysis. Using simulated data, our method proved to have increased power to detect associated sets of markers compared to our original approach, including markers involved in gene-gene interaction. Various ways to detect gene-gene interaction effects besides testing only those genes with significant marginal effects were explored.



Rare Copy Number Variants in Over 100,000 Subjects Reveal Novel Disease Associations

Rose Yun Li^{1,2,17#}, Joseph T. Glessner^{1,4,15#}, Bradley P. Coe¹², Jin Li¹, Xiao Chang¹, Charly Kao¹, Anna Cederquist¹, Cecilia Kim¹, Munir Khan¹, Frank Mentch¹, Maria Garris¹, Debra Abrams¹, F. George Otieno¹, Patrick M.A. Sleiman^{1,2,8}, Evan E. Eichler¹³, Hakon Hakonarson^{1,2,8}

1. The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA.
2. Dept. of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
8. Division of Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA;
12. Dept. of Genome Sciences, University of Washington School of Medicine, Seattle, Washington;
13. Dept. of Genome Sciences, Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, Washington.
14. Psychiatric and Neurodevelopmental Genetics Unit, Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA.
15. Dept. of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA.
17. Medical Scientist Training Program, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA.

Copy number variations (CNVs) significantly impact human genomic diversity affecting gene function that may contribute to both common and rare diseases. We report the 'CNV landscape' across the genome of 100,028 unrelated individuals of European ancestry, based on genome-wide SNP and CGH arrays. We observed over 1.7×10^6 CNVs, averaging 650kb CNV-burden/individual and mapping to 11,314-deletion, 5,625-duplication, and 2,746-homozygous-deletion (hd) CNV Regions (CNVRs). Most CNVRs are rare (frequency <0.01) and recurrent ($>98.5\%$), although 25% of hdCNVRs are private. Over 58% of CNVRs overlapped at least one gene, and were enriched ($P < 1 \times 10^{-4}$) for OMIM morbid genes (enrichment ratio 'ER'=2.94), GWAS loci (ER=1.52) and non-coding RNAs (ER=1.44). CNVR-bearing loci were strongly enriched (35.3%) for recombination hotspots (ER >10 ; $P < 1 \times 10^{-4}$), underscoring a potential impact of natural selection in driving human genome diversity and CNV distribution. In contrast, rates of overlap with microsatellites and segmental duplications were the same as observed at random. The CNVRs identified were associated with 4 major disease categories including autoimmune/inflammatory (n=11,489), oncologic (n=9,105), congenital heart/metabolic (n=2,581), and neuro-developmental (n=33,841) diseases. Of the 131 significant disease-associated CNVRs (DA-CNVRs) identified, 18% overlapped GWAS hits (ER=2.40 and 1.20, $P < 1 \times 10^{-4}$ including those associated with autoimmunity (*ITGB8*; $P < 6.87 \times 10^{-19}$), neurodevelopmental defect (*HCN2*; $P < 6.53 \times 10^{-47}$), as well as multiple DA-hdCNVRs such as *ERRB4*, associated with multiple sclerosis. Our work encompassing the genome wide CNV landscape of $>100K$ individuals underscores the value of large-scale, genome-wide CNV analysis and the need to consider common and rare CNVs in understanding the genetic contribution to complex disease.

Abstract 16

Abstract 17

A comparison of methods for identification of genetic variants related to age-of-onset of Cystic fibrosis related diabetes

Hua Ling¹, Peng Zhang¹, Elizabeth W. Pugh¹, Melis Atalar², Scott M. Blackman³

1. Center for Inherited Disease Research, Johns Hopkins University, USA.
2. The McKusick-Nathan Institute of Genetic Medicine, Johns Hopkins University, USA.
3. Division of Pediatric Endocrinology, Johns Hopkins University, USA.

Cystic fibrosis (CF) is a monogenic disease that affects more than 80,000 people worldwide and causes life-limiting lung disease and pancreatic dysfunction. Diabetes (CF-related diabetes or CFRD) is the most common extrapulmonary complication of CF and affects >40% people with CF by adulthood with a broad range of age of onset. This variation in CFRD risk has been shown to be heritable, and Blackman et al. (2013) identified five loci associated with CFRD onset, analyzed as a survival trait while excluding related individuals in a total of 3,059 samples. To better account for relatedness in survival analyses, we investigate different strategies using a family subset of the above data, the CF Twin and Sibling study, which includes 396 samples from 288 small families (siblings and half siblings) genotyped on the Illumina 610-Quad. Our preliminary analyses show using mixed-effect Cox models, either with family-specific random intercept or with correlated random intercept using a kinship coefficient matrix, yield slightly better control of inflation of type 1 error compared to Cox proportional hazard model including related individuals ($\lambda = 1.00$ and 1.096 for family-specific and correlated random intercept respectively vs 1.12 for Cox proportional hazard model with related individuals), but not compared to maximally unrelated subset analysis ($\lambda = 1.03$). Analyses of PC-adjusted Martingale residuals in linear mixed model with relatedness as random effects yield similar results ($\lambda = 1.096$). Further analyses will be performed to better understand the difference in results between models. Supported by CF Foundation.

Risks of familial breast cancer associated with known and proposed breast cancer susceptibility genes

Kara N. Maxwell¹, Thomas Paul Slavin^{2,3}, Jenna Lilyquist^{4,5}, Joseph Vijai⁶, Susan L. Neuhausen³, Steve N. Hart⁴, Vignesh Ravichandran⁴, Tinu Thomas⁶, Ann Maria⁶, Kasmintan Schrader⁷, Raymond Moore⁴, Chunling Hu⁴, Bradley Wubbenhorst⁸, Brandon M. Wenz⁸, Kurt D'Andrea⁸, Susan M. Domchek¹, Mark E. Robson⁹, Paolo Peterlongo¹⁰, Paolo Radice¹⁰, James M. Ford¹¹, Judy E. Garber¹², Csilla Szabo¹³, Kenneth Offit^{6,9}, Fergus J. Couch⁴, Jeffrey N. Weitzel^{2,3}, Katherine L. Nathanson⁸

1. Dept. of Medicine, Div. of Hematology-Oncology, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA.
2. Dept. of Medical Oncology, Div. of Clinical Cancer Genetics, City of Hope, Duarte, CA.
3. Dept. of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA.
4. Dept. of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN
5. Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN.
6. Clinical Genetics Research Lab, Dept. of Medicine & Dept. of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY.
7. Hereditary Cancer Program, British Columbia Cancer Agency, Vancouver, BC.
8. Dept. of Medicine, Division of Translational Medicine and Genetics, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, PA.
9. Clinical Genetics Service, Dept. of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY.
10. IFOM, the FIRC Institute of Molecular Oncology, and Dept. of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.
11. Div. of Oncology, Stanford University School of Medicine, Stanford, CA.
12. Center for Cancer Genetics and Prevention, Dana Farber Cancer Institute, Boston, MA.
13. National Institutes of Health, Bethesda, MD.

A better understanding of gene-specific risks for development of breast cancer will lead to improved screening, prevention, and therapeutic strategies for individuals identified to carry germline mutations. We performed targeted massively-parallel sequencing to identify mutations and large genomic rearrangements in 26 known or proposed breast cancer susceptibility genes in 2134 *BRCA*-negative women with familial breast cancer (FBC). A case-control analysis was performed comparing the frequency of internally classified mutations identified in FBC cases to that in non-Finnish European controls from the Exome Aggregation Consortium (ExAC) excluding samples from The Cancer Genome Atlas. Including large genomic rearrangements, mutations were identified in 8.2% of FBC cases compared to 6.2% of ExAC controls, including mutations in high-penetrance genes (0.6% in cases vs. 0.1% in controls), moderate-penetrance genes (3.7% vs 1.7%), and seven cases with two mutations (0.3%). The remainder of FBC cases and ExAC controls had mutations in proposed breast cancer genes (1.6% of cases vs 2.4% of controls), Lynch syndrome genes (0.5% vs. 0.5%) or were heterozygous *MUTYH* carriers (1.5% vs. 1.5%). Case-control analysis demonstrated significant associations with FBC for *ATM*, *PALB2*, and *TP53* mutations (OR>3.0, p<10⁻⁴), *BARD1* mutations (OR=3.2, p=0.012), and *CHEK2* truncating mutations (OR=1.6, p=0.041). Our results therefore demonstrate that only approximately 4% of *BRCA1/2* negative FBC patients have mutations in genes definitively associated with breast cancer at this time. Large case-control studies are needed to fully evaluate the breast cancer risks associated with moderate penetrance and proposed breast cancer susceptibility genes.

Abstract 19

Determining and inducing gene expression patterns underlying cell identity

Ian A. Mellis^{1,2}, Wenli Yang^{3,4}, Parisha P. Shah⁵, Rajan Jain^{3,4,5,6}, Arjun Raj^{1,2,7}

1. Department of Bioengineering, University of Pennsylvania, Philadelphia PA.
2. Genomics and Computational Biology Group.
3. Institute for Regenerative Medicine.
4. Department of Medicine, University of Pennsylvania, Philadelphia PA.
5. Department of Cell and Developmental Biology, University of Pennsylvania, Philadelphia PA.
6. Penn Cardiovascular Institute.
7. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA.

Within an individual, cells of different cell types behave differently from one another despite containing identical genetic material. Developmental cues and other extrinsic factors influence these differential phenotypes both by adding new material and by directing cells to change themselves into a stable state we call “cell type”. Given these specific cues, cell-intrinsic activity, most broadly regulated through gene expression, is essential for the entrenchment and maintenance of a stable cell type. One important characteristic of cell types is stability to perturbations, such as inflammation, circadian rhythm, and other sources of non-cell type-specific signaling. Indeed, to our frustration 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 hope to understand and engineer cell-intrinsic patterns of gene regulation underlying the maintenance of cell identity for two predefined cell types of interest: fibroblasts and cardiomyocytes. In the first stage of this project we are doing transcriptome-wide expression profiling with RNAseq on panels of drug-perturbed fibroblasts and cardiomyocytes to identify gene expression patterns that are most consistently found in each cell type, despite perturbations. Then, in combination with genome-wide chromatin accessibility profiling with ATAC-seq and transcription factor binding site prediction, we will identify transcription factors that are likely directly regulating these gene expression patterns of interest. Lastly, we will develop a more efficient fibroblast to cardiomyocyte transdifferentiation protocol using inhibitors of fibroblast-specific factors and overexpression of cardiomyocyte-specific factors.

Urinary Epidermal Growth Factor and Monocyte Chemoattractant Protein-1 as Biomarkers of Renal Involvement in ANCA-Associated Vasculitis

Catherine E. Najem¹, Wenjun Ju², Haley M. Gore², Viji Nair², David Cuthbertson³, Rennie L. Rhee¹, Laura Mariani², Simon Carette⁴, Gary S. Hoffman⁵, Nader A. Khalidi⁶, Curry L. Koenig⁷, Carol A. Langford⁵, Carol McAlear¹, Paul A. Monach⁸, Larry W. Moreland⁹, Christian Pagnoux⁴, Philip Seo¹⁰, Ulrich Specks¹¹, Antoine G. Sreih¹, Steven R. Ytterberg¹¹, Jeffrey Krischer³, Matthias Kretzler², and Peter A. Merkel¹, for the Vasculitis Clinical Research Consortium and the North American Nephrotic Syndrome Network

1. University of Pennsylvania, Philadelphia, PA.
2. University of Michigan, Ann Arbor, MI, USA.
3. University of South Florida, Tampa, FL.
4. Mount Sinai Hospital, Toronto, Ontario.
5. Cleveland Clinic, Cleveland, OH.
6. McMaster University, Hamilton, Ontario.
7. University of Utah, Salt Lake City, UT.
8. Boston University, Boston, MA.
9. University of Pittsburgh, Pittsburgh, PA.
10. Johns Hopkins University, Baltimore, MD.
11. Mayo Clinic, Rochester, MN.

Introduction: Urinary EGF (uEGF) and MCP-1 (uMCP-1) predict active and chronic renal disease in primary glomerulopathies. This study examined uEGF and uMCP-1 as biomarkers of renal disease in ANCA-associated vasculitis (AAV). **Methods:** Data from patients with AAV enrolled in a prospective multicenter cohort were included. uEGF and uMCP-1 were measured at baseline, an active renal disease visit, 1-2 visits prior to and after the active renal disease visit, and at 1 year. Chronic kidney disease (CKD) was defined as eGFR < 60 mL/min/1.73 m² for >3 months and disease activity was measured by BVAS/WG.

Results: 112 patients had active renal disease. Mean uEGF/Cr at index was lower than the mean uEGF/Cr at pre-index (p<0.01) but was not lower than the mean uEGF/Cr at post-index (p=0.33). Patients at pre index and post index have lower uEGF/Cr (p=0.49 and p=0.06). Mean uMCP-1/Cr at index was higher than the mean uMCP-1/Cr at pre-index (p= 0.01) and was higher than the post index (p< 0.01). Patients at pre index and post index have lower uMCP-1/Cr (p=0.04 and p<0.01 respectively). Patients with higher baseline uEGF/Cr have a 38% less risk of developing CKD [HR=0.62, 95% CI (0.43, 0.88), p=0.01]). Patients with higher baseline uMCP-1/Cr have a 14% higher risk of developing CKD [HR=1.14, (0.88, 1.48), p=0.33]).

Conclusions: In AAV, lower uEGF at baseline predict progression to CKD independent of urine albumin/creatinine and eGFR. Urinary MCP-1 correlate with renal disease activity in AAV. uEGF is a useful biomarker in AAV similar to in lupus and nephrotic syndrome.

Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates

Scott Norton^{1,2}, Jordi Vaquero-Garcia^{2,3}, Yoseph Barash^{2,3}

1. Biomedical Graduate Studies, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States.
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States.
3. Department of Computer and Information Sciences, School of Engineering, University of Pennsylvania, Philadelphia, United States.

A key component in many RNA-Seq based studies is contrasting multiple replicates from different experimental conditions. In this setup replicates play a key role as they allow to capture underlying biological variability inherent to the compared conditions, as well as experimental variability. However, what constitutes a “bad” replicate is not necessarily well defined. Consequently, researchers might discard valuable data or downstream analysis may be hampered by failed experiments.

Here we develop a probability model to weigh a given RNA-Seq sample as a representative of an experimental condition when performing alternative splicing analysis. We demonstrate that this model detects outlier samples which are consistently and significantly different compared to other samples from the same condition. Moreover, we show that instead of discarding such samples the proposed weighting scheme can be used to downweight samples and specific splicing variations suspected as outliers, gaining statistical power. These weights can then be used for differential splicing (DS) analysis, where the resulting algorithm offers a generalization of the MAJIQ algorithm. Using both synthetic and real-life data we perform an extensive evaluation of the improved MAJIQ algorithm in different scenarios involving perturbed samples, mislabeled samples, no-signal groups, and different levels of coverage, showing it compares favorably to other tools. Overall, this work offers an outlier detection algorithm that can be combined with any splicing pipeline, a generalized and improved version of MAJIQ for differential splicing detection, and an evaluation pipeline researchers can use to evaluate which algorithm may work best for their needs.

The genomic landscape of matched primary and metastatic breast cancer tumors

Matt R. Paul¹, Tien-chi Pan¹, Dhruv Pant¹, Natalie Shih², Yan Chen¹, George Belka¹, David Lieberman², Jennifer J. D. Morrissette², Danielle Soucier^{3,4}, Michael Feldman², Angela DeMichele^{3,4,5}, Lewis A. Chodosh^{1,2,6}

1. Department of Cancer Biology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
2. Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
3. Department of Medicine, Division of Hematology-Oncology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
4. Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
5. Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
6. Department of Medicine, Division of Endocrinology, Diabetes and Metabolism at the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.

The majority of deaths from breast cancer are due to distant metastatic disease. However, metastatic tumors have been largely ignored when selecting samples for comprehensive genomic analyses as well as to inform individual treatment for the disease. This is mainly a result of the relative difficulty to biopsy smaller, less accessible metastatic tumors, but also due to the widely held view of genomic determinism; that is, the majority of actionable mutations can be discovered in the primary tumor. Using WES and sWGS to determine coding mutations and aberrant copy-number in 17 primary-metastatic tumor pairs, we found that the majority of pairs show significantly large measures of discordance at the genomic level. Furthermore, we found that in a most cases, metastatic tumors follow a branched pattern of evolution, suggesting that metastatic colonization occurred much earlier in time before primary tumor diagnosis and characterization. Selection is made evident based on the decreased clonal heterogeneity of metastatic tumors, as well as the relative enrichment of coding and copy-number alterations in several different cellular pathways such as the PI3K signaling cascade. In conclusion, the small degree of concordance between primary and metastatic tumors due to evolutionary distance, as well as the presence of activating and targetable mutations specifically in metastatic tumors suggest that metastatic tumors should be comprehensively characterized when informing treatment for patients and to discover new paradigms underlying breast cancer tumor survival.

Abstract 23

Identification of epistatic interactions between the human RNA demethylases FTO and ALKBH5 with gene set enrichment analysis informed by differential methylation

Elizabeth R. Piette¹ and Jason H. Moore¹

1. Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.; piette@upenn.edu (corresponding); jhmoore@upenn.edu

The Genetic Analysis Workshop presents an opportunity to collaboratively evaluate methodology relevant to current issues in genetic epidemiology. The GAW20 data combine real clinical trial data with fictitious epigenetic drug response endpoints. Considering the evidence suggesting that networks of interactions between many genes underlie complex phenotypes, we utilize differential methylation status to identify a relevant gene set for enrichment analysis and use this to infer potential biological function underlying drug response. We highlight the pertinence of considering the potential for widespread epistatic interactions in the absence of main effects, and present evidence of epistasis between SNPs on the two RNA demethylases FTO and ALKBH5.



Change in ancestry related assortative mating in the United States and implications for genetic studies

Ronnie Sebro MD, PhD¹; Gina M Peloso, PhD²; Josee Dupuis, PhD^{2,3}; Neil J Risch^{4,5,6}

1. University of Pennsylvania, Department of Radiology, Philadelphia, PA.
2. Department of Biostatistics Boston University School of Public Health, Boston.
3. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham.
4. Institute for Human Genetics University of California, San Francisco CA.
5. Department of Epidemiology and Biostatistics University of California, San Francisco, CA.
6. Division of Research Kaiser Permanente, Oakland, CA.

Genetic similarity of spouses can reflect factors influencing mate choice, such as physical/behavioral characteristics, and patterns of social endogamy. Spouse correlations for both genetic ancestry and measured traits may impact genotype distributions (Hardy Weinberg and linkage equilibrium), and therefore genetic association studies. Here we evaluate white spouse-pairs from the Framingham Heart Study (FHS) original and offspring cohorts (N=124 and 755, respectively) to explore spousal genetic similarity and its consequences. Two principal components (PCs) of the genome-wide association (GWA) data were identified, with the first (PC1) delineating clines of Northern/Western to Southern European ancestry and the second (PC2) delineating clines of Ashkenazi Jewish ancestry. In the original (older) cohort, there was a striking positive correlation between the spouses in PC1 ($r=0.73$, $P=3 \times 10^{-22}$) and also for PC2 ($r=0.80$, $P=7 \times 10^{-29}$). In the offspring cohort, the spouse correlations were lower but still highly significant for PC1 ($r=0.38$, $P=7 \times 10^{-28}$) and for PC2 ($r=0.45$, $P=2 \times 10^{-39}$). We observed significant Hardy-Weinberg disequilibrium for single nucleotide polymorphisms (SNPs) loading heavily on PC1 and PC2 across 3 generations, and also significant linkage disequilibrium between unlinked SNPs; both decreased with time, consistent with reduced ancestral endogamy over generations and congruent with theoretical calculations. Ignoring ancestry, estimates of spouse kinship have a mean significantly greater than 0, and more so in the earlier generations. Adjusting kinship estimates for genetic ancestry through the use of PCs led to a mean spouse kinship not different from 0, demonstrating that spouse genetic similarity could be fully attributed to ancestral assortative mating. These findings also have significance for studies of heritability that are based on distantly related individuals (kinship less than 0.05), as we also demonstrate the poor correlation of kinship estimates in that range when ancestry is or is not taken into account.

Genes & Geography: A comprehensive study of geographical effects on addiction and immunity in populations

Maksim Shestov¹ and Latifa F. Jackson²

1. Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA.
2. National Human Genome Center, College of Medicine, Howard University, Washington, DC.

Finding the genetic markers that influence infectious and chronic disease phenotypes has been an area of significant biological study. Understanding complex disease related traits like addiction has been hampered by the lack of functional insights in to the human genome. We hypothesized that environmental factors such as geographical location, relative pathogen burden and infection rates will identify allele frequency differences in for immune and addiction gene hotspots between populations that are consistent with natural selection within human populations living predominantly in tropical environments.

To test whether there are correlative relationships between ecoregions, disease and population allele frequencies, we use genes contained in addiction and immunity curated by NCBI. Immune associated genes were identified from NCBI gene lists using immune related search terms. These terms were added to 587 genes previously identified as being involved in opiate, dopamine, and GABA reception addiction. These genes were then projected onto the genome to identify cluster regions of genetic importance for immunity and addiction. Clusters were defined as regions of the genome with more than 15 genes within a 1.5Mb linear genomic window. When addiction and immunity gene lists were combined, we found that they created three hotspots located on chromosomes 11, 17, and 19. Human polymorphism data was surveyed from the 1054 individuals comprising 51 populations of the Human Genome Diversity Panel, 1148 individuals comprising the 11 sample populations of the HapMap Project and the 1092 individuals representing the 1000 Genomes dataset. Our analyses demonstrate that when human populations are grouped into tropical versus non-tropical living groups, significant differences in allele frequencies at the hotspot located on chromosome 11 for 5 polymorphisms were found.

DNA methylation markers associated with injection drug use status and HIV infection among chronic injection drug users in the ALIVE study

Chang Shu¹, Kelly M. Bakulski, Kelly S. Benke², Andrew E. Jaffe³, Shaocheng Wang¹, Sarven Sabuncian¹, Shruti Mehta¹, Gregory D. Kirk¹, Brion S. Maher¹

1. Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.
2. University of Michigan, Ann Arbor, MI, USA 3Lieber Institute for Brain Development, Baltimore, MD, USA.

Background: Injection drug use (IDU) is associated with biological modification of the genome, i.e., epigenetics marks such as DNA methylation and histone modification. However, the biological mechanisms on how substance use and abstinence affects epigenetic outcomes remain largely unknown.

Objective: To conduct the epigenome-wide association analyses among injection drug users to identify whether any injection drug use during the past six months is associated with genome-wide blood epigenetic markers.

Methods: In the AIDS Linked to the Intravenous Experience (ALIVE) study, blood was obtained from 288 current IDUs, resampled after cessation and then again after relapse (total samples = 774). Blood DNA methylation marks were measured using the Illumina Infinium MethylationEPIC BeadChip. Standard procedures in the minfi R package were used to preprocess raw Illumina image files into noob processed methylation beta values. Differences in DNA methylation at individual probes by current injection status were tested using generalized linear regression including age, gender, race and cell heterogeneity as covariates.

Results: DNA methylation at individual loci (cg10801015, $p=2.47 \times 10^{-11}$; cg11415166, $p=5.15 \times 10^{-10}$; cg03426703, $p=2.62 \times 10^{-9}$; cg14977491, $p=9.94 \times 10^{-6}$) is significantly associated with current injection drug use status after correction for multiple testing. Those CpG sites were near the PDP1, NARFL, DVL2 and PFN2 genes correspondingly, and their related pathways are metabolism and some signaling pathways.

Conclusion: In a preliminary study, we performed a genome-wide scan of methylation changes in a longitudinal study of injection drug users and identified genomic locations exhibiting significant changes in peripheral DNA methylation associated with injection drug use status.

QRank: A novel quantile regression tool for eQTL discovery

Xiaoyu Song^{1,†}, Gen Li², Zhenwei Zhou², Xianling Wang², Iuliana Ionita-Laza², and Ying Wei²

1. Heilbrunn Department of Population & Family Health, Columbia University, New York, NY 10032, USA.
2. Department of Biostatistics, Columbia University, New York, NY 10032, USA.

Motivation: Over the past decade, there has been a remarkable improvement in our understanding of the role of genetic variation in complex human diseases, especially via genome-wide association studies. However, the underlying molecular mechanisms are still poorly characterized, impeding the development of therapeutic interventions. Identifying genetic variants that influence the expression level of a gene, i.e. expression quantitative trait loci (eQTLs), can help us understand how genetic variants influence traits at the molecular level. While most eQTL studies focus on identifying mean effects on gene expression using linear regression, evidence suggests that genetic variation can impact the entire distribution of the expression level. Motivated by the potential higher order associations, several studies investigated variance eQTLs.

Results: In this paper, we develop a Quantile Rank-score based test (QRank), which provides an easy way to identify eQTLs that are associated with the conditional quantile functions of gene expression. We have applied the proposed QRank to the Genotype-Tissue Expression project, an international tissue bank for studying the relationship between genetic variation and gene expression in human tissues, and found that the proposed QRank complements the existing methods, and identifies new eQTLs with heterogeneous effects across different quantile levels. Notably, we show that the eQTLs identified by QRank but missed by linear regression are associated with greater enrichment in genome-wide significant SNPs from the GWAS catalog, and are also more likely to be tissue specific than eQTLs identified by linear regression.

Availability: An R package is available on R CRAN at:
<https://cran.r-project.org/web/packages/QRank>
Contact: xs2148@cumc.columbia.edu

Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention

Cristian Tomasetti^{1,2}, Lu Li², Bert Vogelstein³

1. Division of Biostatistics and Bioinformatics, Dept. of Oncology, Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, 550 N Broadway, Baltimore, MD 21205, USA.
2. Dept. of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA.
3. Ludwig Center & Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, 1650 Orleans St, Baltimore, MD 21205, USA.

The role of environmental factors (E) and heredity (H) in cancer causation has been confirmed in multiple studies. We recently hypothesized that mutations occurring randomly as a result of the normal processes associated with cellular replication (R) are also an important contributor to cancer and can explain why cancers occur much more commonly in some tissues than others.

To determine what fractions of cancer-causing mutations result from E, H, or R, we developed a novel approach based on the integration of genome-wide sequencing and epidemiological data. When normalized for the incidence of each of 32 cancer types, the application of this methodology yields that 29% of the mutations in cancers occurring in the UK are attributable to E, 5% of the mutations are attributable to H, and therefore 66% are presumably due to R. The estimate for the proportion that may be attributed to R varies considerably: it is less than 40% in cancers such as those of the lung, esophagus, and skin and 80% or more in cancers such as those of the prostate, brain, and breast.

Given the mathematical relationship between cancer etiology and cancer preventability, the proportion of mutations caused by environmental factors is always less than the proportion of cancers preventable by avoidance of these factors. Thus, our estimate that a maximum of 29% of the mutations in these cancers are due to E is perfectly compatible with the estimate that 42% of these cancers are preventable by avoiding known risk factors.

Abstract 29

Assessing the Geospatial Distribution of Asthma Exacerbations in Philadelphia Using Electronic Health Record (EHR)-Derived Data

Sherrie Xie¹, Rebecca Greenblatt¹, Michael Z. Levy¹, Andrea Apter¹, Michelle Ross¹, Blanca E. Himes¹

1. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA.
Email: xiex@vet.upenn.edu

Asthma exacerbations, episodes of worsening asthma symptoms requiring the use of systemic corticosteroids to prevent serious outcomes, are a major cause of morbidity and health care costs in the US. To effectively make progress toward understanding how demographic and environmental factors influence asthma exacerbations, Electronic Health Record (EHR)-derived data is valuable as a longitudinal repository of events affecting diverse populations. We obtained de-identified patient-level data for adult asthma encounters within the University of Pennsylvania Health System (UPHS) occurring between 2011 and 2014. Variables extracted included codified demographic information, geocodes corresponding to place of residence, ICD-9 encounter codes, prescribed medications, and patient addresses. We restricted our analyses to asthma patients who had at least one primary ICD-9 code for asthma and a prescription for albuterol. Cases were defined as patients who had at least one asthma exacerbation, defined as an encounter with a primary ICD-9 code for asthma and a new prescription for oral corticosteroids, while controls had no exacerbations. Spatial analysis was performed using generalized additive models with the *MapGAM* R package. From 2,748 cases and 5,464 controls, we determined through multivariate analysis that *black race/ethnicity*, older ages, *grade ≥ 3 obesity*, current or previous smoking history, and *Medicare or Medicaid* financial class vs. *Private Insurance* were independent predictors of asthma exacerbations. GAM analysis found that asthma exacerbations were associated with geographic location. The global test statistic against the null hypothesis that exacerbation odds did not depend on location was highly significant ($p < 0.001$) suggesting that even after adjusting for covariates (i.e. race, age, BMI, smoking status, financial class), asthma exacerbations were highly spatially correlated. Local GAM tests identified hot spots with significantly increased exacerbation rates, including a region in Southwest Philadelphia ($p < 0.01$). Our results suggest that EHR data is helpful to understand the geospatial distribution of asthma exacerbations in Philadelphia.

Supported in part by NHLBI R00 HL105663 (BEH), R01 HL133433 (BEH), T32 AI070077-08 (SX), and NIEHS P30 ES013508.

Testing for genetic association in case-control studies incorporating multivariate disease characteristics

Haoyu Zhang, BS¹; Thomas U. Ahearn, Ph.D.²;
Montserrat Garcia-Closas, M.D.²; Nilanjan Chatterjee, Ph.D.^{1,3}

1. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University.
2. National Cancer Institute Division Cancer Epidemiology and Genetics.
3. Department of Oncology, School of Medicine, Johns Hopkins University.

As sample size for genome-wide association studies continues to rise, there is unprecedented opportunity for obtaining new insights to genetic architecture of complex diseases. Many diseases like breast cancer are intrinsically heterogeneous consisting of subtypes that could be defined by various pathologic and molecular disease characteristics. We propose a two-stage modeling framework for modeling genetic associations in GWAS of cancers utilizing multivariate tumor characteristics. The framework can be used to test for overall genetic association, global etiologic heterogeneity and individual etiologic heterogeneity in terms of tumor characteristics. We propose efficient methods for handling missing tumor characteristics so that all cases, irrespective of whether they have complete tumor characteristics data or not, can efficiently contribute to the analysis. Preliminary applications will be illustrated based on analysis of a large GWAS ($N_{\text{case}}=106,571$ and $N_{\text{control}}=95,762$) of breast cancer incorporating ER, PR and HER2 status, three clinically relevant tumor characteristics. The methods will first be demonstrated based on known susceptibility-associated SNPs which reported by previous large-scale GWAS studies and ongoing unpublished studies. We also applied our methods on whole genome analysis with around 21 million imputed SNPs.

Abstract 30

