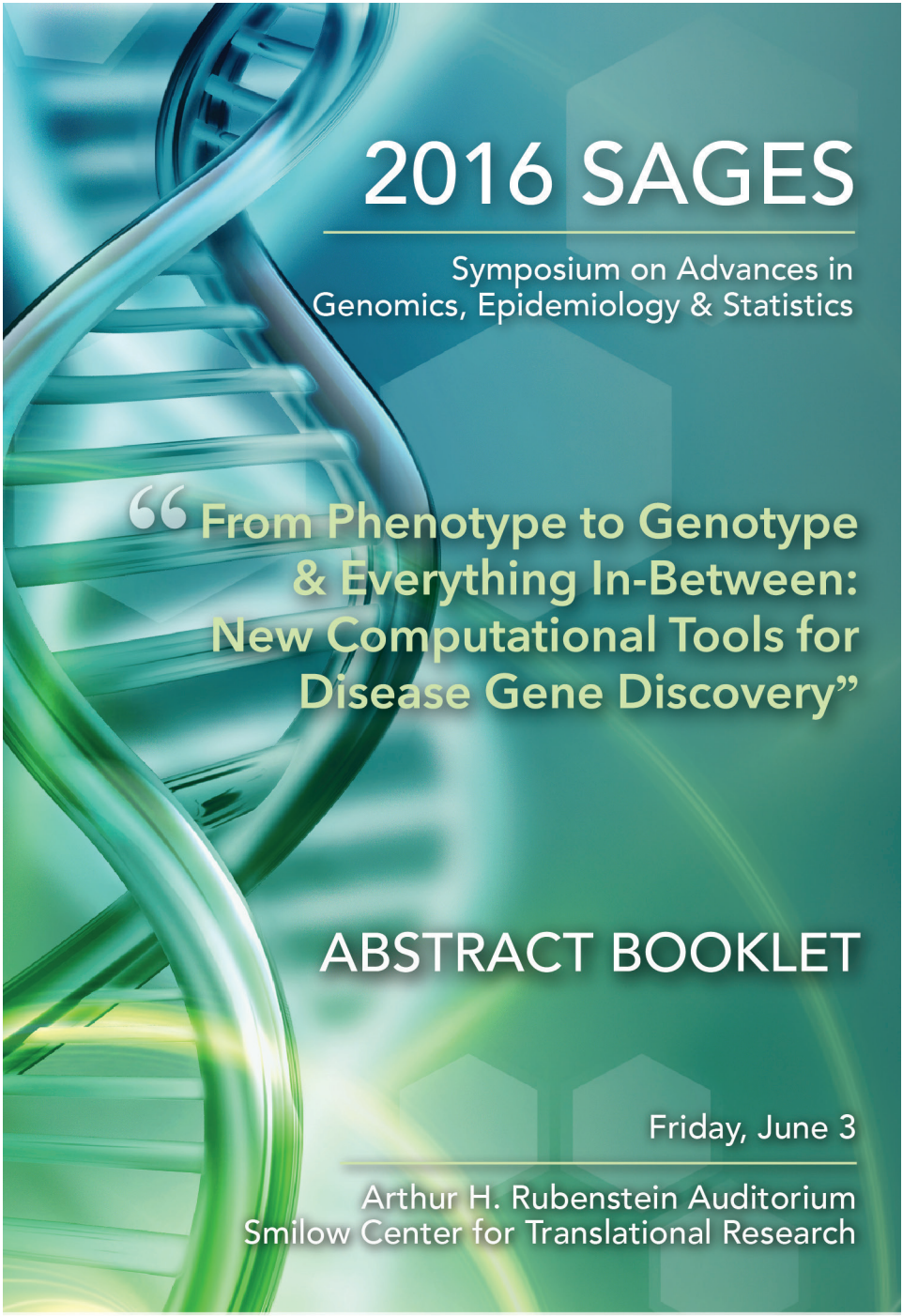

2016 SAGES

Symposium on Advances in Genomics, Epidemiology & Statistics

"From Phenotype to Genotype & Everything In-Between: New Computational Tools for Disease Gene Discovery"

ABSTRACT BOOKLET

Friday, June 3

Arthur H. Rubenstein Auditorium
Smilow Center for Translational Research

CGACT Center for Genetics and Complex Traits

The Children's Hospital of Philadelphia RESEARCH INSTITUTE

CCEB

NIH National Institutes of Health

Funding for this conference was made possible in part by grant R13 HG007809 from the National Human Genome Research Institute and National Institute on Deafness and Other Communication Disorders.

ABSTRACT 1

A large de novo germline sequencing study of bilateral RB patients finds no higher oncogenic potential of nonsense mutations at CpG sites in RB1 gene

*Varun Aggarwala[1], Arupa Ganguly[2] and Benjamin F. Voight[2,3]

1. Genomics and Computational Biology Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA 19104
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, PA 19104
3. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA 19104

Retinoblastoma (RB) is a childhood cancer of the developing retina, caused by mutations in the RB1, a tumor suppressor gene. Clinical sequencing studies of the RB1 gene have found an enrichment of (Arg->Stop) nonsense mutations at CpG sites, in both bilateral and unilateral form of RB, and have suggested a preferential oncogenic role of CpG sites in causing RB. Moreover, the role of other type of missense and splice mutations in causing RB is not clear, although they are present in several cases. In this study we identified the largest set of 268 germline de novo mutations from bilateral RB patients, and identified oncogenic potential of a mutation class by measuring the enrichment of mutation spectrum beyond an expected model of mutation rate. As expected, we find a higher oncogenic potential of nonsense and essential splice mutations in RB1 gene, but surprisingly found no enrichment of Arg->Stop (CGA->TGA) variants within the class of nonsense mutations, suggesting no preferential role of CpG sites in causing RB. Moreover, we find an enrichment of missense mutations at Exon 20 (a pocket binding domain region), and splice mutations at donor regions of Exon 6, 12 and 24, suggesting previously unidentified oncogenic potential at these sites.

*Eligible for poster award

ABSTRACT 2

Identified structural variants associated with multiple phenotypes of COPDGene African American Study Cohort

*Ferdouse Begum[1], Ingo Ruczinski[2], Shengchao Li[3], Margaret M. Parker[1], Jacqueline Hetmanski[1], Terri H. Beaty[1], Edwin Silverman[4], James Crapo[5], and COPDGene Investigators

1. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Maryland, USA
2. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Maryland, USA
3. Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA
4. Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
5. Department of Medicine, National Jewish Health, Denver, USA

Chronic obstructive pulmonary disease (COPD) is the third leading cause of mortality in USA. Though COPD has a well-recognized environmental risk factor (i.e. cigarette smoking), recent well-powered genome-wide association studies (GWAS) have identified multiple genomic regions where single nucleotide polymorphic (SNP) markers are strongly and consistently associated with risk to COPD. To thoroughly investigate the genetic architecture underlying COPD and related phenotypes, it is also important to explore the role of structural variants, since they can alter gene expression and have been shown to be causal for some diseases.

We delineated CNVs using PennCNV on 9,076 COPDGene participants using genome-wide marker data generated using Illumina's Omni-Express array. COPDGene subjects are comprised of one-third African-American and two-thirds non-Hispanic white adult smokers, with or without COPD. After employing rigorous quality control procedures to reduce the false positive CNV calls, we tested for association between CNV components (defined as disjoint intervals of copy number regions within race) and several COPD-related phenotypes.

We detected hemizygous deletions achieving genome-wide significance on chromosome 5q35.2, near the gene *FAM153B*, in tests of association with total lung capacity assessed by chest CT among African-Americans. We also detected hemizygous deletions on chromosome 3p26.1 associated with two smoking behavior related phenotypes.

*Eligible for poster award

ABSTRACT 3

Synovial fluid proteins differentiate patients with oligoarticular juvenile idiopathic arthritis who are destined to extend from those who will remains persistent in course

AC Brescia[1], MM Simonds[2], KE Sullivan[3], CD Rose[1]

1. Nemours AI duPont Hospital for Children, Wilmington, DE
2. Nemours Biomedical Research, Wilmington, DE
3. The Children's Hospital of Philadelphia, Philadelphia, PA

Background: Children with oligoarticular juvenile idiopathic arthritis (oligo-JIA) who have an extended course (>5 joints after 6 months) have a worse prognosis than those who have a persistent course (<4 or fewer joints). Our goal is to identify protein biomarkers which will discriminate between patients whose involvement will remain persistently oligo-JIA and those who are destined to extend.

Methods: As part of a separate ongoing IRB approved protocol, remnant synovial fluid was obtained from patients undergoing medically indicated arthrocenteses. Using our clinical database, JIA samples were separated into two groups: oligo-JIA with persistent course (PR), oligo-JIA with extended course (E). All samples were from steroid-naïve joints and samples from E were obtained prior to time of extension. Using synovial fluid samples from 4 PR and 4 E patients, RayBiotech Membrane-based Antibody Arrays were completed for expression of 89 cytokines, chemokines, and other inflammatory proteins and intensity measured using ImageJ. Key significant proteins were confirmed using R&D Systems Quantikine Colorimetric Sandwich ELISAs.

Results: 20 proteins had significantly different expression levels. To confirm these results, ELISAs were performed. ELISA confirmed the antibody array results.

Conclusions: Samples of synovial fluid from differing courses of JIA, PR and E, taken before the ultimate course is clinically apparent, have significantly different protein levels of important chemokines and cytokines. These differences can be exploited to predict which patients may extend, allowing earlier therapeutic intervention to prevent long term disability.

ABSTRACT 4

Genome Wide Association Study Identifies EFEMP1 as a New Candidate Biliary Atresia Susceptibility Gene

*Ying Chen, Christopher M. Grochowski, Melissa Gilbert, Ramakrishnan Rajagopalan, Kathleen M. Loomes, Nancy B. Spinner, Marcella Devoto.

The Children's Hospital of Philadelphia and the University of Pennsylvania, Philadelphia, PA

Biliary atresia (BA) is a rare liver disease presenting within the first months of life. It is characterized by obliteration of the extrahepatic biliary tree in a progressive, necroinflammatory manner, leading to cholestasis, fibrosis, cirrhosis, and chronic liver damage, and accounts for 50% of pediatric liver transplantations. The etiology of BA is not well understood, although environmental, inflammatory and genetic risk factors have been proposed. In this study, we performed a GWAS in 450 European-American non-syndromic BA patients collected through the Childhood Liver Disease Research Network (ChiLDReN) and 1981 controls from the Age-Related Eye Disease Study (AREDS). Adjusted logistic regression was carried out to test SNP association with BA. The most significant SNP was rs10865291 on 2p16 ($p = 2.7x10-7$; OR = 1.6), located in the fifth intron of *EFEMP1*. *EFEMP1* encodes the EGF-containing fibulin-like extracellular matrix protein Fibulin-3, which has been implicated in tissue regeneration and organogenesis. ddPCR performed on cDNA from BA liver specimens collected at time of liver transplant, from control liver, and from cholestatic disease liver (n = 5 each) revealed that *EFEMP1* expression was significantly increased in BA livers and cholestatic disease livers. Immunohistochemistry showed that Fibulin-3 was specifically expressed in tubular structures around portal tracts in BA livers. Follow-up co-staining with CK-19 and alpha smooth muscle actin will be performed to determine cell type-specific expression. Ongoing studies include replication of the GWAS in additional BA cohorts. The identification of this new candidate susceptibility gene will facilitate characterizing the genetic basis of BA.

*Eligible for poster award

ABSTRACT 5

Identifying functionally relevant Mental Health hotspots and polymorphisms in African Americans

*Christopher Cross[1], Latifa Jackson[2], and Max Shestov[3]

1. National Human Genome Center, College of Medicine, Howard University
2. Anatomy, College of Medicine, Howard University
3. Biomedical Engineering, Sciences and Health Systems, Drexel University

BACKGROUND: African Americans are 20 percent more likely to report serious psychological distress than European Americans. Recent work by our group suggests that genes sitting at the intersection of addiction and mental health form genomic hotspots that are functionally cohesive with drugable targets. This lead us to question how the mental health genes underlying schizophrenia, bipolar disorder, post-traumatic stress disorder, and depression form functional hotspots in the genome, whether they have unique allele frequencies in African ancestry populations, and whether those hotspots had drugable targets.

METHODS: Mental health associated gene sets were obtained from NCBI Gene and were projected onto gene ontology categories and cellular pathways to draw a bioinformatics portrayal of mental health disorders. Single nucleotide polymorphism datasets (HapMap, Human Genome Diversity Panel) were used identify sequence variants that were significantly different in African Americans from other populations. Finally we annotated the drug binding sites lying in each hotspot to predict the effects of these drugs on human populations.

RESULTS: Mapping addiction genes onto human genome resulted in eight gene clusters, with at least 20 mental health genes (Range: 21-46 genes/ hotspot) in a 4Mb distance along DNA. Hotspot genes were involved in neurological transmission, responses to organic substances, and cell-cell signaling. Analysis of hotspot drug binding sites show that mental health has a treatment overlaps with the functional annotations

CONCLUSION: This approach identified 10 key putative variants for mental health in individuals of African Ancestry, found functionally cohesive hotspots. Drugable targets annotation suggest that treatment could be better targeted (in progress).

*Eligible for poster award

ABSTRACT 6

Exploring Tissue-Specific Effects of Rare Non-Coding Variants

*Farhan Damani[1], Yungil Kim[1], Xin Li[3], Joe Davis[4], Emily Tsang[3], Colby Chiang[5], Zachary Zappala[3,4], Ben Strober[2], The GTEx Consortium, Ira Hall[5], Stephen B. Montgomery[3,4], Alexis Battle[1]

1. Department of Computer Science, Johns Hopkins University, United States
2. Department of Biomedical Engineering, Johns Hopkins University School of Medicine, United States
3. Department of Pathology, Stanford University School of Medicine, United States
4. Department of Genetics, Stanford University School of Medicine, United States
5. The Genome Institute, Washington University School of Medicine, United States

Large-scale whole genome and transcriptome sequencing provide us with a unique opportunity to understand the role of rare non-coding variants in the gene regulatory process. Given the excess of rare variants from recent human population growth and the fact that over 98% of the human genome is made of noncoding DNA, understanding the role of these potentially high impact variants will be critical to our understanding of individual disease risk. We seek to model tissue-specific cis-regulatory effects of rare variants by integrating whole genome sequencing, genomic annotations, and gene expression data across 48 human tissues from the GTEx Project. These methods capitalize on a growing body of rich epigenetic data to quantity the deleteriousness of a variant, using factors such as a conservation score, transcription factor binding sites, and location of the variant. We have developed a hierarchical Bayes latent variable model, which attempts to capture both shared and tissue-specific effects from rare variants in a transfer-learning framework. We learn our model using expectation-maximization and apply shrinkage estimation techniques in the maximization step in order to update our hierarchical parameters. After learning our model, we cluster our genomic annotation parameters and find that rare variants have similar effects in tissues with similar functionality. Our probabilistic model also offers great potential for isolating specific rare variants that play functional roles in any given tissue.

*Eligible for poster award

ABSTRACT 7

Causal Effects of Blood Pressure Components on Brain Atrophy and White Matter Integrity: A Mendelian Randomization Analysis Based On Individual Participant Data

*Jie Ding, Osorio Meirelles, Lenore J Launer

Laboratory of Epidemiology and Population Sciences, National Institute on Aging, National Institutes of Health, Bethesda, MD

Background: Elevated blood pressure (BP), especially beginning at young and middle age, has been implicated in the pathogenesis of brain tissue loss, presumed to be a global indicator of atrophy, and white matter lesions (WML), an indicator of ischemic and myelin damage. Both structural brain measures are important substrates for cognitive impairment and dementia in older people. As yet, the causality has not been established for any specific BP component (i.e. systolic [SBP], diastolic [DBP], mean arterial [MAP] or pulse pressure [PP]) as assessed in middle age. The aim of this research project is to utilize multiple independent single nucleotide polymorphisms (SNPs) as instrumental variables in the Mendelian randomization approach to investigate the causal role of elevated BP in the development of brain tissue atrophy and WML.

Methods: We will construct unweighted and weighted genetic risk scores based on independent SNPs with established associations with SBP, DBP, PP and MAP. We will then determine whether the use of genetic risk score as an instrumental variable would provide support for a causal association between individual BP components and neuroimaging endophenotypes including whole brain volume, total gray matter volume and white matter microstructural integrity by estimating fractional anisotropy and mean diffusivity. Finally, we will conduct Mendelian randomization meta-analyses in population-based prospective studies participating in the Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium.

Significance: Establishing a firm causal link between elevated BP in mid-life and subtle brain injury would identify the earliest BP-related changes in brain health and further emphasize the importance of early monitoring and management of hypertension as a preventive strategy against cognitive decline late in life, in addition to the recognized benefits of treatment on cardiovascular and cerebrovascular events.

*Eligible for poster award

ABSTRACT 8

PheWAS study using research participants' self -reported data provides insight intoTh17/IL-17 pathway

M.G. Ehm[1], J.L. Aponte[3], M.N. Chiano[2], L.M. Yerges-Armstrong[1], J.N. Barker[5], S.F. Cook, A. Gupta[1], D.A. Hinds[4], L. Li[3], T. Johnson[2], M.A. Simpson[5], C. Tian[4], L.C. McCarthy[2], D. Rajpal[1], D.M. Waterworth[1]

1. Target Sciences, GlaxoSmithKline, King of Prussia, PA;
2. Target Sciences, GlaxoSmithKline, Stevenage, UK;
3. PAREXEL International, Research Triangle Park, NC;
4. 23andMe, Inc., Mountain View, CA

A PheWAS study of variants within genes involved with the Th17/Il17 pathway was performed using patient reported phenotypes and genetic data from 23andMe. Results replicated known associations with several autoimmune phenotypes including Irritable Bowel Syndrome (IBS), Crohn's Disease, and psoriasis, illustrating that participants' self-reported outcomes can be a surrogate for clinically assessed conditions. Novel associations with an FDR significance level of 0.05 included association of *IL23R* Arg381Gln variant with a decrease in dandruff frequency, *RORC* regulatory variant (rs4845604) with protection from allergy, *TRAF3IP2* Asp10Asn variant, with risk of male pattern baldness and *TYK2* Ile683Ser variant with increased weight, height and BMI as well as risk of tonsillectomy, strep throat and teenage acne. This study enabled a rapid assessment of dermatological and immuno-inflammation phenotypes in large numbers and the study of allergy and infection phenotypes, previously unstudied, which has shed light on conditions being considered for drugs targeting this pathway.

ABSTRACT 9

Prioritize Risk Genetic Variants in Regulatory DNA Sequences Using Disease-relevant Gene Regulatory Networks

Long Gao[3],*, Yasin Uzun[1],*, Xiaoke Ma[4], Jiahui Wang[3], Bing He[1], Kai Tan[1,2]

1. Division of Oncology and Center for Childhood Cancer Research, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
2. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
3. Department of Biomedical Engineering, University of Iowa, Iowa City, IA, 52242, USA
4. School of Computer Science and Technology, Xidian University, Xi'an, Shanxi, China

* Equal contribution

Technology development has transformed the field of human genetics by revealing thousands of common and rare DNA sequence variants. Interpretation of these variants in both healthy individuals and those with complex diseases remains a daunting task. The vast majority (over 90%) of associated variants have been found to localize outside of known protein-coding sequences, thus impeding the direct interpretation of their functional effects. Here we present Annotation of Regulatory Variants using Integrated Networks (ARVIN), a general computational framework for identifying causal noncoding variants that affect a specific trait against the background of ubiquitous genetic variation. In addition to various genomic and epigenomic annotations, we develop and characterize multiple network based features to identify relevant variants. We applied our method to 233 regulatory mutations annotated in Human Gene Mutation Database (HGMD) for 20 types of diseases, and for various classification tasks our comprehensive model is competitive or even outperforms models using genomic and epigenomic features alone. Moreover, our network based approach provides functional context by linking identified variants to related signal transduction pathways. Thus, our framework presents an effective strategy to prioritize causal genetic variants.

ABSTRACT 10

Assessing the role of genetic variation in amplified musculoskeletal pain syndromes (AMPS) in pediatric populations

*Michael V. Gonzalez[1,2]., Sabrina A. Gmuca[2], Jenna Tress[2], Frank D. Mentch[1], Rosetta Chiavacci[1], Patrick Sleiman[1], Gordon A. Barr[3,4], David D. Sherry[2,4], and Hakon Hakonarson[1,3]

1. Center for Applied Genomics, The Children's Hospital of Philadelphia, Abramson Research Center, Philadelphia, PA, 19104
2. Amplified Pain Program, Division of Rheumatology, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104
3. Department of Anesthesiology and Critical Care Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104
4. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104.

Amplified musculoskeletal pain syndromes (AMPS) comprise a range of disorders that can present with a variety of signs and symptoms related to increasing pain over time, allodynia, and disproportional dysfunction. AMPS are most prevalent in adolescent females of European Caucasian (EC) descent (approx. 75%) and often significantly interfere with daily function and quality of life. Additionally, pediatric amplified musculoskeletal disorders differ significantly from those occurring in adults. The etiology and pathophysiological mechanisms of AMPS are currently unknown. Familial aggregation and candidate gene studies suggest a strong genetic component in the development of AMPS, but many of the genetic susceptibility factors remain unidentified. Genome-wide association studies (GWAS) facilitate impartial queries of common genetic variation, allowing identification of novel genetic contributions to the development of complex disorders. To our knowledge, this is the first report of a genome-wide investigation of AMPS in a large pediatric cohort. To this end, 359 AMPS cases and 4151 controls were selected for genotyping utilizing single nucleotide polymorphism (SNP) arrays. The discovery cohort was comprised of females of EC descent and demonstrated nominal association between AMPS incidence and a series of novel SNPs at the *SLC6A5* locus (top SNP rs1617769, $P= 4.8x10^{-7}$ and rs1792974, $P= 5.1x10^{-6}$). Interestingly, the *SLC6A5* gene encodes the GlyT2 receptor which has important roles in regulation of glycinergic neurotransmission. Replication and validation of these results in an independent cohort of 629 AMPS cases is currently underway. Additionally, the GWAS identified other genomic regions that offer enticing candidates for consideration in future studies.

*Eligible for poster award

ABSTRACT 11

r2VIM: Variable selection method for identifying interaction effects

*Emily Holzinger[1], Silke Szymcak[2], James Malley[3], Abhijit Dasgupta[4], Qing Li[1], Joan Bailey-Wilson[1]

1. Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA.
2. Institute of Medical Informatics and Statistics, University of Kiel, Kiel, Germany.
3. Center for Information Technology, National Institutes of Health, Bethesda, Maryland, USA.
4. Clinical Trials and Outcomes Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland, USA.

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, such as interactions between variables with small main effects. These types of effects likely contribute to the heritability of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no standardized method of selecting variables so that false positives are reduced while retaining adequate power. To this end, we have developed a novel variable selection method called r2VIM. This method incorporates recurrency and variance estimation to assist in optimal threshold selection. For this study, we specifically address how this method performs in simulated data with close to completely epistatic effects (i.e. no marginal effects).

Our initial findings indicate that the optimal selection threshold can often identify both interacting loci while reducing the number of false positives in the selected variables. However, the optimal threshold is highly dependent on the underlying simulated genetic model, which is unknown in biological data. To address this, we also test a permutation procedure to generate null VIM distributions based on the actual genotype data to guide threshold selection. We permute the phenotype and re-run r2VIM to get a new estimate of the null variance. This is then used to choose a selection threshold for the non-permuted analysis. We tested the permutation method on a subset of the simulated data used in the initial analysis. The results suggest that the permutation procedure can guide optimal threshold selection in data with strong interaction effects in a manner that retains locus detection power and a low false positive selection rate.

* Eligible for poster award

ABSTRACT 12

Gender Differences in the Experience of Violence, Discrimination, and Stress Hormone in African Americans: Implications for Public Health

*Latifa Jackson[1], Maksim Shestov[2], Muneer Abbas[3], Frough Saadatmand[4]

1. National Human Genome Center, Howard University
2. Biomedical Engineering, Sciences and Health Systems, Drexel University
3. Microbiology, College of Medicine, Howard University
4. Pediatrics, College of Medicine, Howard University

Violence is more prevalent in African American communities than in other American communities. This has impacts not only on criminal justice interventions, but also on the physical and mental health of these communities including their risk for acquiring life threatening diseases. While many studies have focused on the effects of violence on African American males, we sought to understand the relative gender effects that violence has on African American females. This is of particular interest given the wide array of scientific literature suggesting maternal health has inter-generational health effects. Introducing gender biases associated with exposure to violence, depression, and immune function is an important step in understanding how young women perceive and internalize societal violence directed toward and around them. We study a cohort of 557 young African American adults aged 18-25 years old from the Washington DC area. We use sociological, epidemiological, mental health, computational biology and quantitative genetics approaches to build a predictive portrait of the effects of violence on African American health. This study demonstrates that African American males and females experience different constellations of societal violence, that African American women report greater perception of racial and gender bias, and that cortisol, an indicator of stress response is correlated to perceived discrimination. This work contributes current understandings of how violence contributes to negative health outcomes, and lays the foundation for a predictive model for sociological, health and behavior risk that young African Americans encounter.

*Eligible for poster award

ABSTRACT 13

Increased burden of deleterious variants in essential genes in autism spectrum disorder

*Xiao Ji[1, 2], Rachel L. Kember[2], Christopher D. Brown[2] and Maja Bucan[2, 3]

1. Genomics and Computational Biology Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104
2. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104
3. Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Autism spectrum disorder (ASD) is a heterogeneous, highly heritable neurodevelopmental syndrome characterized by impaired social interaction, communication and repetitive behavior. It is estimated that hundreds of genes contribute to ASD. We asked if genes with a strong effect on survival and fitness contribute to ASD risk and ASD-associated traits. Human orthologs of genes with an essential role in pre- and postnatal development in the mouse ("essential genes") are enriched for human disease genes and under strong purifying selection relative to human orthologs of mouse genes with known non-lethal phenotype ("non-essential genes"). This intolerance to deleterious mutations, coupled with commonly observed haploinsufficiency of essential genes, suggests a possible cumulative effect of a large number of small-effect variants in essential genes on a disease phenotype such as ASD.

With a comprehensive catalog of 3,915 mammalian essential genes, we provide evidence for a stronger contribution of essential genes to ASD risk, compared to non-essential genes. By examining the exonic variations from 1,781 ASD families, we demonstrate a higher burden of damaging mutations in essential genes in ASD probands compared to their non-ASD siblings. The analysis of essential genes in the developing brain identified clusters of co-expressed EGs implicated in ASD. Finally, we suggest a high priority list of 29 EGs with potential ASD risk as targets for future functional and behavioral studies. Overall, we show that large-scale studies of gene function in model organisms provide a powerful approach for prioritization of genes and pathogenic variants identified by sequencing studies of human disease.

*Eligible for poster award

ABSTRACT 14

Accounting for technical batch effects in single-cell RNA sequencing analysis

*Cheng Jia,[2] Mingyao Li,[2] Nancy R. Zhang[1]

1. Statistics Department, The Wharton School,
2. Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA

In single-cell RNA sequencing (scRNA-seq), cell-specific technical variations can often lead to batch effects, thus confounding downstream analyses, causing biased inferences, inflated false positive rates and reduced power. In order to account for these cell-to-cell technical differences, we propose a statistical framework, TASC (Toolkit for Analysis of Single Cell RNA-seq). TASC uses an empirical Bayes approach to reliably model the cell-specific rela-tionships of dropout rates and amplification efficiency to the true expression of a gene using external RNA spike-ins. Subsequently, TASC incorporates the estimated batch effects into a hierarchical mixture model to perform analyses such as quantifying gene expression, esti-mating biological variance and detecting differentially expressed genes. TASC is capable of adjusting for additional covariates to further eliminate confounding originated from sources such as cell properties and demographics. In addition, it is programmed to be highly efficient computationally, taking advantage of multi-threaded parallelization. In our simulation stud-ies based on real scRNA-seq datasets, TASC displays superior sensitivity and specificity in detection of differential expression. Moreover, it is robust in the presence of batch effects in comparison to competing methods. In real data analyses, TASC discovers the largest num-ber of differentially expressed genes overall, and its performance is consistent across various sample sizes.

*Eligible for poster award

ABSTRACT 15

Assessing intra-tumor heterogeneity and tracking longitudinal and spatial clonal evolution by next-generation sequencing

*Yuchao Jiang[1,2], Yu Qiu[3,4,5,6], Andy J Minn[3,4,5,6], Nancy R Zhang[2]

1. Genomics and Computational Biology Graduate Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
2. Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
3. Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
4. Department of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
5. Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
6. Institute of Immunology, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA

Cancer is driven by genetic and epigenetic alterations that follows Darwinian evolution. Recently, there have been increasing efforts to sequence the tumor from the same patient at multiple time points and/or from multiple spatially separated resections. We propose a method, Canopy, for reconstructing subclonal phylogeny using both copy number alterations (CNAs) and single nucleotide alterations (SNAs) from sample(s) derived from a single patient. Canopy provides a mathematical framework that enumerates all possible CNA-SNA phases and gives confidence assessments of all possible phylogenetic configurations. On a whole-exome study of a transplantable metastasis model derived from human breast cancer cell line MDA-MB-231, Canopy successfully deconvolutes the mixed cell sublines, using the single cell sublines as ground truth, and identifies DNA signatures that can be prognostic of distant metastasis. On a whole-genome sequencing dataset of the primary tumor and relapse genome of a leukemia patient, Canopy predicts phylogenetic histories in concordance with existing knowledge. On a whole-genome sequencing dataset of the breast cancer tumor and its subsequent metastatic xenograft, Canopy's inferred clonal phylogeny is concordant with genomic markers of major clonal genotype and is confirmed by single-cell sequencing. Finally, through simulations, we explore the effects of various parameters on deconvolution accuracy, and evaluate performance with comparison against existing methods. Collectively, Canopy provides a rigorous foundation for statistical inference on repeated sequencing experiments from evolving populations delineated temporally and spatially.

* Eligible for poster award

ABSTRACT 16

Integrity of induced pluripotent stem cell (iPSC) derived megakaryocytes as assessed by genetic and transcriptomic analysis

*Kai Kammers[1], Margaret Taub[1], Ingo Ruczinski[1], Joshua Martin[2], Lisa R Yanek[2], Alyssa Frazee[1], Yongxing Gao[3], Dixie Hoyle[3], Nauder Faraday[2], Diane M Becker[2], Linzhao Cheng[3], Zack Z. Wang[3], Jeff Leek[1], Lewis C Becker[2], Rasika A Mathias[2]

1. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD
2. The GeneSTAR Program, Johns Hopkins School of Medicine, Baltimore, MD
3. Division of Hematology and Institute for Cell Engineering, Johns Hopkins School of Medicine, Baltimore, MD

Aggregation of platelets in the blood may initiate arterial occlusions causing heart attacks or strokes. To understand the biology of platelet aggregation, we examined genetic and transcriptomic data from megakaryocytes (MKs), the precursor cells for anucleate platelets, that are derived from induced pluripotent stem cells (iPSCs).

To this end it is essential that the iPSC-derived MKs retain their genomic integrity during production or expansion. This was examined using three alternative measures of integrity of the MK cell lines: (1) mutation rates comparing parent cell DNA to iPSC cell DNA and onward to the differentiated MK DNA; (2) structural integrity using copy number variation (CNV) on the same; and (3) transcriptomic signatures of the derived MK cells. Another goal is to detect patterns of transcript expression in the MKs related to specific genetic variants. We are currently performing extensive eQTL analysis to categorize 'functional' relevance of the GWAS-identified determinants of platelet aggregation leveraging the genotype and RNA-seq data.

To date, our data contains MKs from 209 people with informative genotypes. Given a high genetic and transcriptomic integrity of the iPSC-derived MKs, we found several hundred cis-eQTLs in European Americans and African Americans and see a high replication between the two groups. The majority of cis-eQTLs are unique to MKs compared to other tissue types that are reported in GTEx Portal.

* Eligible for poster award

ABSTRACT 17

Epidemiology and Clinical studies related to Elongated Styloid Process (ESP) causing head and neck pain

*Khan, HM, Fraser, AD, Mupparapu, M

Penn Dental Medicine, Division of Oral & Maxillofacial Radiology, Philadelphia, PA

Background: For dental clinicians, it is not uncommon to encounter a patient with Elongated Styloid Process (ESP). Four percent of the population has been found to have ESP. Of this population, 4-10% are estimated to have symptomatic elongation syndrome uniltarally.[1,2] Females have a 3:1 predilection compared to males, and 30-50 years old is the common age group affected.[2] Appropriately diagnosing cases through imaging can provide patients with necessary treatment and pain relief.

Materials & Methods: Screened radiographs and electronic records of patients who visited Penn Dental Medicine's (PDM) Oral Diagnosis clinic from January 2015 to December 2015 after an IRB approval. Patients who reported either pain or discomfort in the head and neck area were further imaged. Although there were several styloid calcifications, only 4 patients fit the criteria for ESP.

Results: All four patients had panoramic radiographs displaying ESP. Only one patient had a CBCT due to extended pain and a prior biopsy.

Discussion/Conclusions: The examination of four patients exhibited different degrees of ESP. The patient who presented with neck pain that necessitated a CBCT, was ultimately diagnosed with Eagle Syndrome. Among the other studies, patients were only evaluated based on panoramic radiographs. CBCT volumes were not obtained in these patients because they either reported no pain, or the source of pain was not from their styloid elongation. Although the prevalence of ESP is rare, individual cases are crucial to study as patients can potentially present with severe pain, and therefore the implications in dentistry are monumental.

References:
1. Rechtweg JS, Wax MK. Eagle's syndrome: a review. Am J Otolaryngol. 1998;19:316–321.
2. Zeckler S-R, Betancur AG, Yaniv G. The eagle is landing: Eagle syndrome — an important differential diagnosis. Br J Gen Pract. 2012;62(602):501-502

* Eligible for poster award

ABSTRACT 18

Training a Type 2 Diabetes Specific Functional Sequence Predictor

*Kimberly M. Lorenz[1,2,3] and Benjamin F. Voight[1,2,3]

1. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia PA
2. Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia PA
3. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia PA

Type 2 diabetes (T2D) is a complex polygenic disorder characterized by an inability to maintain glucose homeostasis. Genome wide association studies have identified >100 independent loci contributing to the disease, with evidence for hundreds more. However, much of this variation lies in noncoding regions of the genome, making causal variants and sequences difficult to identify and study. We hypothesize that the regulatory features underlying causal variants are disease specific and identifiable from data—the regulatory architecture that influences T2D susceptibility is distinct compared to other diseases. However, models that underlie contemporary functional variant predictors, such as CADD, GWAVA, and FATHMM-MKL, are disease-agnostic to maintain generalizability. To better characterize and identify T2D variants, we have trained a disease specific functional variant predictor using known disease loci and publicly available genomic and epigenomic annotations from the ENCODE and Epigenome Roadmap consortia. Our current model uses a binomial based penalized logistic regression (elastic net) for feature selection. We are better able to classify T2D loci compared to current predictors, with an area under the curve (AUC) of 0.84 compared to their AUCs of 0.76-0.79. Our model retains biologic intuition, as we observe that selected features skew towards T2D-relevant tissues. When we include additional features from T2D-specific tissue types outside of ENCODE or Roadmap, our AUC increases to 0.87. We intend to improve upon this foundation by identifying and including further T2D specific annotations, and will test additional methods of feature selection.

* Eligible for poster award

ABSTRACT 19

Theoretical constraints imposed by the transcriptome on genotype-phenotype maps

*Ian A. Mellis[1] and Arjun Raj[1,2]

1. Genomics & Computational Biology, Perelman School of Medicine, University of Pennsylvania
2. Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania

A central goal of systems biology is identifying the mechanisms by which genotype influences phenotype. We know from decades of molecular genetics experimentation that information passing from the genome into the physiology of an organism essentially always first takes the form of gene expression or regulation thereof. Additionally, the field of quantitative genetics has accumulated significant evidence in support of pleiotropic effects being common in several metazoans, although there is still debate over the shapes of the distributions of these effects. In a population, there is a limited number of genetic variants in which phenotypic information can be encoded, an even smaller number of potential transcripts with which to transmit that information, and a diversity of traits to reliably pass on. Therefore, we are faced with a problem: how do genetic systems encode a large amount of heritable phenotypic variability in a wide array of traits (1) with massive cross-talk across the genetic and biochemical networks underlying individual phenotypes and (2) while subject to necessarily transferring information through the transcriptome?

We are exploring the idea that the observed networks of weak genetic effects encode a genotype to phenotype map encompassing the diverse phenotypic variation observable in a population compressively and isometrically. The compressive property of this mapping corresponds to the limited number of genes that serve as an intermediate between the larger numbers of possible variants and the myriad phenotypic variations in organisms. The isometric property is needed to account for additivity of many quantitative traits, i.e., narrow-sense heritability.

* Eligible for poster award

ABSTRACT 20

Linkage Analyses Reveals Significant Association for Myopia

*Anthony M. Musolf[1], Claire L. Simpson[1], Federico Murgia[2], Laura Portas[2], Joan E. Bailey-Wilson[1], Dwight Stambolian[3]

1. Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA
2. Institute of Population Genetics, CNR, Li Punti, Sassari, Italy
3. Department of Ophthalmology, University of Pennsylvania, Philadelphia, PA, USA

Myopia is a condition where overgrowth of the eye causes light to focus in front of the retina, leading to blurring of distant images. It is one of the most common causes of reduced vision worldwide, affecting 1 in 4 Americans.

We have genotype data from extended pedigrees with multiple individuals affected with myopia. These families come from three discrete populations: African-Americans, Ashkenazi Jews and Amish. Two-point and multi-point linkage analyses were performed on each family.

Familywise two-point LOD scores of greater than 2 were observed at chromosomes 3, 4, 6, 10, 13, 14 and 22 in individual Ashkenazi Jewish families. Cumulative LOD scores of greater than three were observed on chromosomes 1, 6, 8, and 16; with the marker on 16 having a cumulative LOD score of 7.3.

We also observed familywise two-point LOD scores greater than 2 in individual Pennsylvania Amish on chromosomes 2, 3, 9, 15, and 16. Cumulative LOD scores of greater than three were observed on chromosomes 4 and 8.

Although no two-point LOD scores higher than 2 were observed in the African-Americans, these families were smaller and therefore less informative. Nonetheless, cumulative LOD scores across families of greater than 3 were observed at 3 SNPs on chromosome 7.

Multipoint linkage analyses are ongoing. Overall, this work identifies multiple interesting linkage signals for myopia across four discrete populations, using both two-point and multi-point analyses. Many of these signals are located within genes, which may be promising candidates for further study.

* Eligible for poster award

ABSTRACT 21

Differential Expression Analysis of Gene and Transcript Abundance for Single Cell RNA-Seq Data using STAR and HISAT Aligners

Julius S. Ngwa[*1], Melissa Liu[2], Robert Wojciechowski[2], Terri Beaty[3], Don Zack[2], Ingo Ruczinski[1]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,
2. Johns Hopkins Wilmer Eye Institute, Johns Hopkins School of Medicine,
3. Department of Epidemiology, Johns Hopkins University School of Public Health

Single-cell RNA-Seq is becoming one of the most widely used methods for transcription profiling of individual cells. Currently there are a number of algorithms available for mapping high-throughput RNA-Seq reads against a reference genome, and for quantifying the abundance of gene transcripts. Accurate characterization of these spliced transcripts is critical in determining functionality in normal and disease cells. Here we compare gene/transcript counts obtained from Hierarchical Indexing for Spliced Alignment of Transcript (HISAT2) and Spliced Transcripts Alignment to Reference (STAR) algorithms. HISAT2 implements a large set of small graph Ferragina-Manzini (FM) indexes, spanning the whole genome to enable rapid and accurate alignment of sequencing reads. STAR aligner consists of a seed searching step and a clustering/stitching/scoring step, and is capable of mapping full-length RNA sequences. We analyzed expression profiles of human and mouse cells from the publicly available Gene Expression Omnibus NCBI database (Series GSE63473 - http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63473). We compared the Digital Gene Expression (DGE) matrix from the aligned library as well per-cell information indicating number of genes and transcripts observed. Some large differences were found in the number of transcripts between STAR and HISAT2 aligners. In particular, the gene counts tended to be higher using HISAT2 compared to STAR. DGE matrices obtained from these aligners showed larger differences in mouse cells compared to human cells. STAR and HISAT2 aligners provide information on the number of reads that map to a particular genomic position, but lack information on which of the overlapping transcripts they originate from. With the presence of ambiguous reads, uncertainties in counts can result in false differential expression calls of transcripts with similar isoforms in the same gene. Resolving potential fragment assignment ambiguity may be an essential issue to address in RNA-Seq data.

* Eligible for poster award

ABSTRACT 22

Multivariate models from RNA-Seq SNVs yield candidate molecular targets for biomarker discovery: SNV-DA

*Matt R. Paul[1,5], Nicholas P. Levitt[1], David E. Moore[1], Patricia M. Watson[2], Robert C. Wilson[2,3], Chadrick E. Denlinger[3,4], Dennis K. Watson[2,3] and Paul E. Anderson[1].

1. Department of Computer Science, College of Charleston, 66 George St., Charleston, SC, USA.
2. Hollings Cancer Center, Medical University of South Carolina, 165 Canon St., Charleston, SC, USA.
3. Department of Pathology, Medical University of South Carolina, 165 Canon St., Charleston, SC, USA.
4. Department of Surgery, Medical University of South Carolina, 165 Canon St., Charleston, SC, USA.
5. Department of Cancer Biology, University of Pennsylvania, 421 Curie Blvd, Philadelphia, PA, USA.

It has recently been shown that single nucleotide variants (SNVs) can be reliably called from RNA-Seq data. These may provide another source of features for multivariate predictive modeling of disease phenotype for the prioritization of candidate biomarkers. The continuous nature of SNV allele fraction features allows the concurrent investigation of several genomic phenomena, including allele specific expression, clonal expansion and/or deletion, and copy number variation.

The proposed software pipeline and package, SNV Discriminant Analysis (SNV-DA), was applied on two RNA-Seq primary tumors datasets: twenty patients with different disease outcomes in lung adenocarcinoma and tumors from two major breast cancer subtypes, estrogen receptor positive and triple negative. Predictive models were generated using the machine learning algorithm, sparse projections to latent structures discriminant analysis. Training sets composed of RNA-Seq SNV features limited to genomic regions of origin (e.g. exonic or intronic) and/or RNA-editing sites were shown to produce models with accurate predictive performances, were discriminant towards true label groupings, and were able to produce SNV rankings significantly different from than univariate tests. Furthermore, the utility of the proposed methodology is supported by its comparable performance to traditional models as well as the enrichment of selected SNVs located in genes previously associated with cancer and genes showing allele-specific expression. As proof of concept, we highlight the discovery of a previously unannotated intergenic locus that is associated with epigenetic regulatory marks in cancer and whose significant allele-specific expression is correlated with ER+ status; hereafter named ER+ associated hotspot (ERPAHS)

* Eligible for poster award

ABSTRACT 23

A Chemogenomics Approach to Identify Bacterial Metabolites with Immune-modulatory Effects via Human Host Receptors

*S. Saha[1], C. Jayawickreme[2], J. P. Marino[3], Andrew Benowitz[3], D. K. Rajpal[1] and J. R. Brown[1]

1. Computational Biology, Target Sciences, R&D
2. Screening Profiling & Mechanistic Biology, NCE Molecular Discovery, R&D
3. Pipeline Futures Group, Alternative Discovery & Development, GlaxoSmithKline, 1250 S. Collegeville Road, PA 19426, USA

Recent metagenomics studies suggest that that microbiome shifts are associated with disease incidences. To understand the microbial basis of several gut-related immune-inflammatory disorders, we focused on the metabolites secreted by the microbiome. These metabolites are truly the interaction currencies between the microbiome and the human host. From a drug discovery perspective, microbial metabolites or their mimics can be explored as a novel drug class because of their high affinity and selectivity for human receptors. Being endogenous in origin, they also presumably mitigate safety and tolerance issues of drugs in human system. Despite progress in the field, few specific metabolite-human ligand interactions are known and there are lack of systematic screens. Here we discuss the identification of potentially bioactive bacterial metabolites using a combination of computational chemistry, bioinformatics and functional screening using in vitro assays representing different disease phenotypes. We searched GlaxoSmithKline's (GSK) pharmaceutical collection of ~4.5 million compounds for structural homologs of 10 known bacterial metabolites as the starting point. Significant biological activities associated with these analogs were systematically mined against human receptors from historical in-house assays. Over 421 compounds were found with a significantly high similarity to the parent metabolites. Of these compounds, over 101 putative human protein target associations were found in GSK assays. A subset of these metabolites and their corresponding structural analogs were then tested in 12 different human primary cell assay systems which mimic various disease phenotypes (BioMAPR assay systems, DiscoveRx). From these screens, several novel immune-modulatory compounds were identified which warrant further investigation.

* Eligible for poster award

ABSTRACT 24

Gene-based analysis identified the gene *ZNF248* is associated with late-onset asthma in African Americans

*Leyao Wang, Yasmmyn D. Salinas, Andrew T. DeWan

Department of Chronic Disease Epidemiology, Yale School of Public Health

Late-onset asthma (LOA) has distinct characteristics and may rely on unique pathogenesis pathways. While current studies are primarily focused on childhood asthma, more research is needed to reveal the mechanisms underlying LOA. To conduct genome-wide association analysis and gene-based analysis to identify single nucleotide polymorphisms (SNPs) and genes that are associated with LOA. The Women's Health Initiative (WHI) Observational Cohort and the Multi-Ethnic Study of Atherosclerosis (MESA) were used to identify subjects with late-onset asthma. The association between LOA and body mass index (BMI) and smoking were evaluated. In the discovery stage of genetic analysis, 1218 African American subjects from WHI with genotype data (n=271 cases and n=947 controls) were used for both single SNP and gene-based association analyses. Significant or suggestive results were subsequently investigated in an independent African American population from MESA (n=38 cases and n=806 controls). In WHI, the relative odds for LOA in obese vs. normal weight subjects was 2.55 (95% confidence interval (CI): 1.74 to 3.76). Ever smokers also had greater odds for LOA compared to never smokers (odds ratio=1.59, 95% CI: 1.21, 2.09). The same trends were observed in MESA. In WHI, six SNPs were associated with LOA at a genome-wide suggestive significance level ($p < 1.0 \times 10\text{-}5$). The gene, *ZNF248* was associated with LOA and reached genome-wide significance ($p = 4.0 \times 10\text{-}7$). In MESA, the association between *ZNF248* and LOA successfully replicated (p=0.015). Both smoking and obesity are risk factors for LOA. *ZNF248* confers increased susceptibility to LOA in African Americans.

* Eligible for poster award

ABSTRACT 25

The Power of Population Diversity in Probing the Biology of Positive Selection in Asthma Disparities

*Bradford Wilson[1], Tshela E Mason[1], Luisel Ricks-Santi [2], Georgia M. Dunston[1,3]

1. National Human Genome Center, Howard University, Washington, DC, USA
2. Cancer Center, Hampton University, Hampton, VA, USA
3. Department of Microbiology, Howard University, Washington, DC USA

BACKGROUND: Genome-wide association studies (GWAS) have been informative for identifying single nucleotide polymorphisms (SNPs) statistically associated with asthma, a chronic inflammatory disease common in African Americans. The interrelationship of these statistically associated SNPs in the etiology of asthma is unknown. In characterizing these SNPs, recent signatures of strong positive selection were used to study genomic adaptation underlying the pathophysiology of asthma.

METHODS: The GWAS Catalog was queried for all signals that were significantly associated with asthma, asthma (childhood onset) and atopy. These signals were stratified based on their signatures of recent strong positive selection. Next, interaction and pathway analyses were conducted on the selection signals of interest. Lastly, the literature was mined for information linking the selection signals of interest to the pathophysiology of asthma.

RESULTS: One of the GWAS signals located in a gene associated with asthma had a strong signature of recent positive selection among African descent populations in the International Haplotype Map Project. Twelve additional genes within a two mega base region of the GWAS signal were found to have the same selection profile. Through interactions analysis and literature mining, eleven of the thirteen associated genes and several of their interaction partners demonstrated a correlation with airway dysregulation in asthma.

CONCLUSIONS: These results illustrate that positive selection can be used as a tool for interrogating the biology associated with common complex diseases.

* Eligible for poster award

ABSTRACT 26

A Powerful Procedure for Pathway-based Meta-Analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations

*Han Zhang[1], William Wheeler[2], Paula L Hyland[1], Yifan Yang[3], Jianxin Shi[1], Nilanjan Chatterjee[4] , Kai Yu[1]

1. Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892, USA
2. Information Management Services Inc., Calverton, MD 20904, USA
3. Department of Statistics, University of Kentucky, Lexington, KY 40508, USA
4. Department of Biostatistics, Bloomberg School of Public Health and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

Meta-analysis of multiple genome-wide association studies (GWAS) has become an effective approach for detecting single nucleotide polymorphism (SNP) associations with complex traits. However, it is difficult to integrate the readily accessible SNP-level summary statistics from a meta-analysis into more powerful multi-marker testing procedures, which generally require individual-level genetic data. We developed a general procedure called Summary based Adaptive Rank Truncated Product (sARTP) for conducting gene and pathway meta-analysis that uses only SNP-level summary statistics in combination with genotype correlation estimated from a panel of individual-level genetic data. We demonstrated the validity and power advantage of sARTP through empirical and simulated data. We conducted a comprehensive pathway-based meta-analysis with sARTP on type 2 diabetes (T2D) by integrating SNP-level summary statistics from two large studies consisting of 19,809 T2D cases and 111,181 controls with European ancestry. Among 4,713 candidate pathways from which genes in neighborhoods of 170 GWAS established T2D loci were excluded, we detected 43 T2D globally significant pathways (with Bonferroni corrected p-values < 0.05), which included the insulin signaling pathway and T2D pathway defined by KEGG, as well as the pathways defined according to specific gene expression patterns on pancreatic adenocarcinoma, hepatocellular carcinoma, and bladder carcinoma. Using summary data from 8 eastern Asian T2D GWAS with 6,952 cases and 11,865 controls, we showed 7 out of the 43 pathways identified in European populations remained to be significant in eastern Asians at the false discovery rate of 0.1. We created an R package and a web-based tool for sARTP.

* Eligible for poster award

Notes