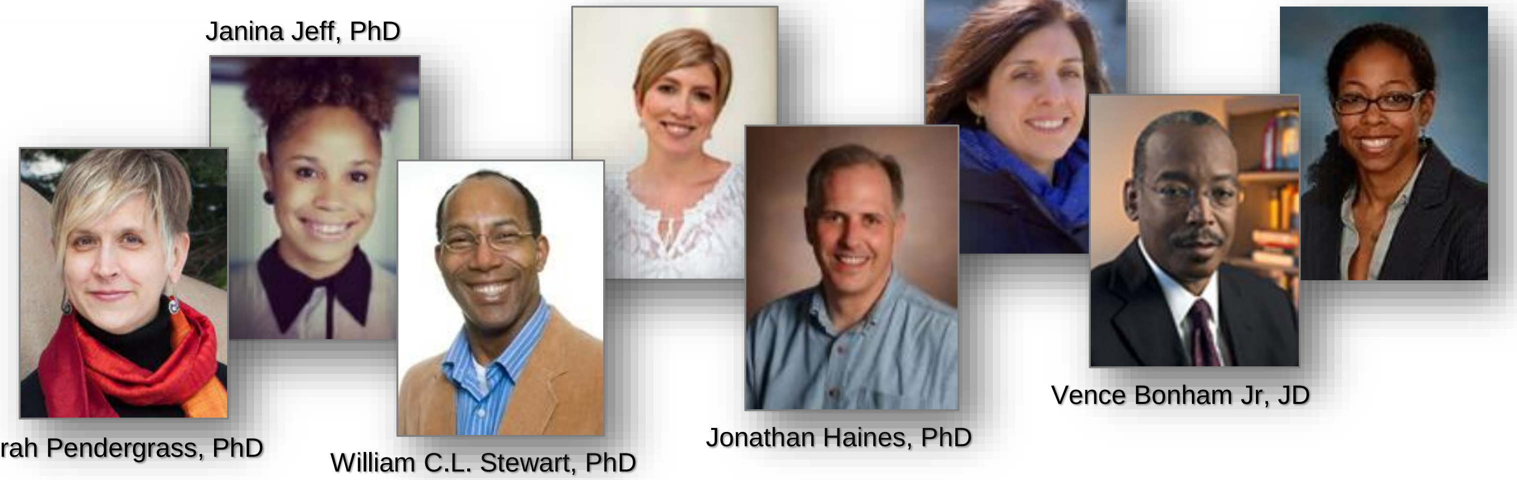


Janina Jeff, PhD

Marylyn Ritchie, PhD

Blanca Himes, PhD

Casey Overby, PhD

Sarah Pendergrass, PhD

William C.L. Stewart, PhD

Jonathan Haines, PhD

Vence Bonham Jr, JD

CWRU Institute for Computational Biology
presents

PRESENT-DAY PROBLEMS AND POTENTIALS FOR PRECISION MEDICINE

and workshop

PUTTING THE PIECES TOGETHER: PRECISION MEDICINE DISCOVERY FROM ELECTRONIC HEALTH RECORDS

INSTITUTE FOR COMPUTATIONAL BIOLOGY

SCHOOL OF MEDICINE
CASE WESTERN RESERVE UNIVERSITY

September 22, 2015

Welcome to the first annual Case Western Reserve University (CWRU) Institute for Computational Biology (ICB) Symposium!

This year, we are highlighting emerging topics in precision medicine. Precision or personalized medicine is the incorporation of 'omic data into clinical practice to better predict, prevent, and treat disease at the individual level. Earlier this year, the White House announced the Precision Medicine Initiative (PMI), an ambitious research endeavor that aims to ascertain and follow 1 million Americans with the purpose of studying variations in genes, environment, and lifestyle that impact risk of disease and response to interventions or treatments. A major component of PMI will be the collection of clinical data on participants through electronic health records and the linkage of 'omic data to these records for clinical care.

The recent availability of electronic health records and the affordability of 'omic data generation make precision medicine possible. However, before precision medicine can be fully and effectively implemented for the benefit of all patients, a host of topics must be explored through the PMI and other research efforts. To discuss these emerging issues, we have brought together several leading experts in precision medicine to discuss specific topics ranging from use of electronic health records in research settings, application of research findings in a clinical setting, ethics and health disparities, and statistical and computational challenges and opportunities.

Immediately following the symposium is the interactive workshop focusing on approaches for electronic phenotyping using electronic health records. Workshop leaders will challenge participants with complex phenotypes to illustrate the process required to extract high quality data from electronic health records for research purposes. The workshop leaders will summarize with brief examples of their own experiences in data extraction and downstream studies in precision medicine research.

We hope you find this year's symposium and workshop enjoyable and educational. Please check our website this summer for next year's symposium at www.icompbio.net. You can also follow us on Instagram (smartpeoplesciencing) and Twitter (@compbio). See you next year!

Dana C. Crawford, PhD
Chair, Organizing Committee
Associate Professor
Epidemiology and Biostatistics
Institute for Computational Biology

SCHEDULE

8:00 – 8:30 Registration (upstairs foyer) and breakfast (Ballroom A)

8:30 – 8:40 Welcome and Introduction; Dana Crawford, PhD; Associate Professor, Department of Epidemiology and Biostatistics, and member, Institute for Computational Biology, Case Western Reserve

8:45 – 9:25 Electronic Health Records and Genomics – a Dynamic Duo; Marylyn Ritchie, PhD, MS; Professor of Biochemistry and Molecular Biology; Director, Center for Systems Genomics at Pennsylvania State University and Director, Biomedical & Translational Informatics Program, Geisinger Research

9:25 – 10:05 Precision Medicine in Cleveland; Jonathan Haines, PhD; Mary W. Sheldon, MD Professor of Genomic Sciences; Chair, Department of Epidemiology and Biostatistics, and Director, Institute for Computational Biology, Case Western Reserve University

10:05 – 10:35 Break (30 min)

10:35 – 11:15 Big Data, Big Problems, and Big Potential; William Stewart, PhD; Assistant Professor, The Ohio State University; Battelle Center for Mathematical Medicine, Nationwide Children's

11:15 - 11:55 An Implementation Model for Point-of-Care Genomic Clinical Decision Support; Casey Overby, PhD; Assistant Professor, University of Maryland, Program for Personalized and Genomic Medicine, Center for Health-related Informatics and Bio-imaging

Noon – 1:30pm Lunch (Ballroom A) and poster session (upstairs foyer)

1:30pm – 2:10pm Precision Medicine in Asthma; Blanca Himes, PhD; Assistant Professor of Epidemiology in Biostatistics and Epidemiology Department, Perleman School of Medicine, University of Pennsylvania

2:10pm – 2:50pm Will Precision Medicine Reduce or Increase Health Disparities? Vence Bonham, Jr., JD; Senior Advisor to NHGRI Director on Genomics and Health Disparities

3:00 – 5:00 pm Putting the Pieces Together: Precision Medicine Discovery from Electronic Health Records (Interactive Workshop)

> PheWAS: Embracing Complexity for Discovery; Sarah Pendergrass, Assistant Professor, Biomedical & Translational Informatics Program, Geisinger Research
>
> Using Electronic Medical Records to Study High-Temporal Phenotypes for Pharmacogenomics; Janina Jeff, post-doctoral fellow at Icahn School of Medicine, Mount Sinai

5:00 – 7:00 pm Dinner in Ballroom B

2015 INVITED SPEAKERS

Vence Bonham, Jr, JD

Dr. Vence Bonham is Senior Advisor to the National Human Genome Research Institute (NHGRI) Director on Genomics and Health Disparities. Dr. Bonham is also Associate Investigator in the NHGRI Social and Behavioral Research Branch and Head of the Health Disparities Unit. Dr. Bonham's research interests the translation of genomic findings into the clinic and how that translation impacts health disparities. Dr. Bonham's research team uses qualitative and quantitative methods to develop measurement tools grounded in theory (race and social cognitive), public health law, law, genomics, and clinical decision-making.

Jonathan L. Haines, PhD

Dr. Jonathan Haines is Mary W. Sheldon, MD, Professor of Genome Sciences; Chair of the Department of Epidemiology and Biostatistics; and Director of the Institute for Computational Biology at Case Western Reserve University School of Medicine. Dr. Haines is a genetic epidemiologist with a focus on adapting and applying statistical computational approaches to identify genetic variants and their modifiers that impact common human diseases with a focus on ocular and neurological diseases including age-related macular degeneration, primary open angle glaucoma, multiple sclerosis, autism, Parkinson's disease, Alzheimer's disease, to name a few. Dr. Haines is also active in Big Data science and electronic health records research and was the previous principal investigator of the National Human Genome Research Institute (NHGRI)-funded electronic MEdical Records & GEnomics (eMERGE) Coordinating Center. More recently, Dr. Haines founded the CWRU Institute for Computational Biology (ICB) as part of a multi-institutional funded effort involving Case Western Reserve University (CWRU), University of Hospital, Cleveland Clinic Foundation, and MetroHealth. Physically located at CWRU, the ICB aims to bring together electronic health records data in a de-identified environment to be accessed by investigators for research purposes. The ICB also aims to provide educational opportunities and resources such as databasing capabilities and the development of statistical methods for big data analysis. Dr. Haines has several accolades in recognition of his discoveries including the Zenith Award for Excellence in Alzheimer's Disease Research (1993) and the Vanderbilt University School of Medicine Sidney P. Colowick Award for Research in Diverse Areas (2005). Dr. Haines was named a fellow in the American Association for the Advancement of Science in 2010.

Blanca E. Himes, PhD

Dr. Blanca Himes is Assistant Professor in the Department of Biostatistics and Epidemiology in the Perelman School of Medicine at the University of Pennsylvania. Dr. Himes is also affiliated with the Graduate Group in Genomics and Computational Biology. As a computational scientist, Dr. Himes' research interests focus on using biomedical informatics approaches to better understand complex respiratory diseases such as asthma. Among the on-going projects in the Himes lab include understanding the function of disease-associated genes by altering cell-specific transcriptomes, integrating diverse genomic datasets to model cellular transcriptome changes involved in disease, and using geospatial demographic and environmental data to better understand factors that contribute to asthma prevalence and exacerbation. In addition to research, Dr. Himes is active in mentoring students, and she blogs occasionally to address topics posed by young students such as career advice, personal motivation for doing research, and parenthood. Dr. Himes' research potential and leadership skills have been recognized with the UCSD Sang-keng Ma Memorial Award (2001) and the Harvard-MIT Division of Health Sciences and Technology Student Leadership Award (2007).

Casey L. Overby, PhD

Dr. Casey Overby is Assistant Professor in the Division of Endocrinology, Diabetes and Nutrition in the Department of Medicine at the University of Maryland and an Adjunct Investigator I in the Geisinger Health System. At the University of Maryland, Dr. Overby is a faculty member in the Program for Personalized and Genomic Medicine and the Center for Health-related Informatics and Bio-imaging, and she is affiliated with the University of Maryland Institute for Advanced Computing Studies. As an informatics researcher, Dr. Overby's interests intersect at public health genetics and biomedical informatics. Dr. Overby is currently developing applications that support translation of genomic research to clinical and population-based healthcare settings and delivering health information and knowledge to the public. Dr. Overby is also developing knowledge-based approaches to use Big Data such as electronic health records for population health.

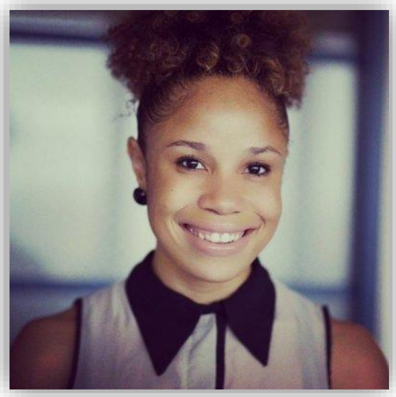
Marylyn D. Ritchie, PhD, MS

Dr. Marylyn Ritchie is Professor in the Department of Biochemistry and Molecular Biology and Director, Center for Systems Genomics at the Pennsylvania State University. Dr. Ritchie is also Director of the new Biomedical and Translational Informatics at Geisinger Clinic. Dr. Ritchie's research interests as a statistical geneticist include the development and application of novel statistical and computational methods to identify genetic variants associated with human diseases. Dr. Ritchie's lab places a special emphasis on the development of methods to detect gene-gene interactions, gene-environment interactions, and network/pathway effects associated with disease. Dr. Ritchie has extensive experience in Big Data science and the use of electronic health records in genomic research. Dr. Ritchie has been the electronic MEdical Records & GEnomics (eMERGE) Coordinating Center genomics lead for the past eight years. Dr. Ritchie's other accomplishments include being named Genome Technology's "Rising Young Investigator" (2006), a Sloan Research Fellow (2010), and a Kavli Frontiers in Sciences fellow by the National Academy of Science (2011-2014). Dr. Ritchie was most recently named Thomas Reuters Most Highly Cited Researchers in 2014.

William C.L. Stewart, PhD, MS

Dr. William Stewart is Assistant Professor of Pediatrics at the Ohio State University. Dr. Stewart is also Principal Investigator at The Research Institute of Nationwide Children's Hospital and a faculty member of the Battelle Center for Mathematical Medicine. Dr. Stewart's broad research interests include the development of statistical models for quick and efficient analyses of complex datasets, deterministic math models, gene networks, and problems in computational molecular biology.

2015 WORKSHOP LEADERS



Janina M. Jeff, PhD, MS

Dr. Janina Jeff is a post-doctoral fellow at the Icahn School of Medicine at Mount Sinai. As a genetic epidemiologist, Dr. Jeff has an interest in identifying genetic variants that explain disease disparities observed across populations. Dr. Jeff has applied her interests to several common human diseases and traits including quantitative cardiovascular traits, anthropometric traits, hematological traits, reproductive outcomes, chronic kidney disease, type 2 diabetes, and pharmacogenomics in highly structured admixed populations such as African Americans and Hispanics. More recently, Dr. Jeff has developed approaches to extract data from a series of clinical data warehouses to define novel clinical phenotypes, for example extracting real-time physiologic response to short acting catacholamines administered during surgery for pharmacogenomic studies. Dr. Jeff is a previous National Human Genome Research Institute Genome Scholar (2006) and a recipient of the Women of Excellence and Leadership award (2006) and Levi Watkins Award for Commitment to Diversity (2011).

Sarah A. Pendergrass, PhD, MS

Dr. Sarah Pendergrass is Assistant (Investigator I) Professor in the Biomedical and Translational Informatics Program at Geisinger Health System. Dr. Pendergrass is a genetic bioinformatician who focuses on high-throughput data analysis and data-mining approaches to studying complex human diseases and traits. Dr. Pendergrass has extensive experience in using both epidemiologic and clinic-based resources to perform phenome-wide association studies (PheWAS) to identify cross-phenotype associations and pleiotropy. Dr. Pendergrass also develops software tools to visualize complex data. In recognition for her innovative work, Dr. Pendergrass was named one of Genome Technology’s PIs of Tomorrow (2013).

ABSTRACTS

Analyzing pathway specificity of variants associated with Alzheimer's disease from the scientific literature corpus

Mariusz Butkiewicz[1], Margaret A. Pericak-Vance[2], Richard Mayeux[3], Lindsay A. Farrer[4], Li-San Wang[5], Gerard D. Schellenberg[5], Will S. Bush[1], Jonathan L. Haines[1]
[1]Case Western Reserve University, Institute for Computational Biology, Cleveland, Ohio; [2]University of Miami, Hussman Institute for Human Genomics, Miami, Florida; [3]Columbia University, New York; [4]Boston University School of Medicine, Boston, Massachusetts; [5]Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania;

Background: Alzheimer's disease (AD) is clinically characterized by progressive cognitive impairment and imposes a tremendous burden on society with an aging population. While the exact cause of the disease is still being investigated, AD therapeutics are being developed to target known pathways associated with e.g. reducing Amyloid beta (Aβ) production, aggregation, and clearance as well as inhibition of tau protein phosphorylation. However, pathway specificity is a valid concern when developing novel therapeutics with limited side effects. In this study, we investigate specificity of annotated variants associated with AD towards involved protein pathways using the scientific literature corpus.

Methods: A set of 1,932 publications in the NHGRI GWAS Catalog were categorized by NCBI MeSH terms. All publications associated with AD or proximal neurological conditions, e.g. Amyotrophic lateral sclerosis and Parkinson's disease, were filtered and associated SNPs were extracted. This list of SNPs was extended by mapping out nearby SNPs using patterns of linkage disequilibrium (LD) from European populations. Using LD, we mapped GWAS associations to SNPs that alter gene expression within the brain using expression quantitative trait loci (eQTL) data, and promoter and enhancer regions identified from brain cell lines within the Fantom5 dataset. The resulting collection of transcripts are hypothesized to influence AD-related conditions, and were selected as high-priority candidates for examining coding and missense variants for AD risk. A pathway enrichment analysis using the PARIS method utilizing KEGG database is currently in progress. We expect to quantify pathway specificity of AD related SNPs through the number of associated pathways.

Results: Starting from 18,939 reported SNPs in the filtered publication list of GWAS catalog, A subset of 369 SNPs was related to AD and other neurological disorders. Further filtering for brain eQTL involvement identified 84 SNP candidates suitable for further investigation. A pathway enrichment analysis investigating SNP candidates related to AD is ongoing and we expect a wide distribution of involved biological pathways to quantify pathway specificity.

Conclusions: The proposed annotation approach poses a promising filtering mechanism to investigate genetic variants associated with AD. Analyzing pathway specificity of these variants can be a powerful technique to prioritize novel biological targets.

Novel Genetic Loci Identified in Meta-Analysis of Glaucoma Genome-Wide Imputed Dataset

Cooke Bailey JN, Gharahkhani P, Pasquale LR, Kang JH, Craig JE, MacGregor S, Burdon K, Haines JL, Wiggs JL, the NEIGHBORHOOD Consortium

Glaucoma is a phenotypically and genetically complex neurodegenerative disease that is the second leading cause of blindness worldwide. Though genetic factors are known to contribute to glaucoma and associated endophenotypes (e.g. intraocular pressure, cup-to-disc ratio, optic nerve parameters, and central corneal thickness), identified risk loci fail to fully account for the genetic component of glaucoma. To extend the proportion of the genome testable for association with glaucoma as well as the power to detect association at numerous loci, we evaluated imputed genome-wide data in the NEIGHBORHOOD Consortium dataset which includes more than 37,000 individuals of European descent from eight studies. Eight datasets, typed on different genome-wide SNP arrays, were imputed to the March 2012 version of the 1000 Genomes data using IMPUTE2 and/or MaCH/miniMaC. We then performed dosage analysis (using ProbABEL) evaluating the estimated genotypic probabilities from the imputation step. The logistic regression model included age, gender, and significant Eigenvectors as covariates for each dataset, along with, where necessary, study-specific covariates. Each dataset was individually evaluated for genomic inflation and then filtered for minor allele frequency (MAF≥0.05) and imputation quality (r^2 or info metric ≥ 0.7). These results were meta-analyzed applying the inverse variance weighted method in METAL and applying genomic control correction. The meta-analyzed dataset includes 3,853 POAG cases and 33,495 controls with high-quality imputed genotypes at 6,425,680 variants. Significant associations were found for genomic regions previously associated with POAG (*TMCO1*, *AFAP1*, *CDKN2BAS*, *ABCA1*, and *SIX6*) as well as novel regions on chromosomes 6p, 12q, 17p, and 22p. This study confirmed several known POAG loci and identified three novel regions of interest. Follow up of genes in these novel regions indicates expression in relevant ocular tissues. These results identify new pathways underlying disease susceptibility and suggest novel targets for preventative therapies.

Functional variants in a clinical setting: an example using *APOC3* R19X and extreme triglyceride levels extracted from electronic health records

Dana C. Crawford[1], Kirsten E. Diggins[2], Nicole A. Restrepo[3], Eric Farber-Eger[4], Quinn S. Wells[5]

[1]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2527, Cleveland, OH 44106; [2]Cancer Biology, Vanderbilt University, 742 Preston Research Building, 2220 Pierce Avenue, Nashville, TN 37232; [3]Center for Human Genetics Research, Vanderbilt University, 519 Light Hall, 2215 Garland Avenue, Nashville, TN 37232; [4]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 600, Nashville, TN 37203; Departments of Medicine and Pharmacology, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 300, Nashville, TN 37203, USA

In a personalized or precision medicine setting, patients with extreme triglyceride (TG) levels may be flagged for further evaluation for cardiovascular disease risk assessment (TG ≥200 mg/dL) or for the presence of hyperthyroidism, malnutrition, or malabsorption disease (TG ≤10 mg/dL). We hypothesize that the addition of functional genetic variants such as *APOC3* R19X, a variant (rs76353203) associated with low TG levels and cardioprotection, can further facilitate the triage process in assessing disease risk among flagged patients. To test this approach, we surveyed BioVU, the Vanderbilt University Medical Center's biorepository linked to de-identified electronic health records (EHRs), for adult European American patients (>45 and >55 years of age for men and women, respectively) with the lowest percentile of TG levels. The initial search identified 262 patients with the lowest TG levels in the biorepository; among these, 184 patients with sufficient DNA and the lowest TG levels were chosen for Illumina ExomeChip genotyping. The average first mentioned TG level for these patients was 39.3 mg/dL. A total of two patients were identified as heterozygotes of *APOC3* R19X for a minor allele frequency (MAF) of 0.55% in this patient population. Both heterozygous patients had only a single mention of TG in the EHR (31 and 35 mg/dL, respectively), and one patient had evidence of previous cardiovascular disease. In this patient population, the inclusion of *APOC3* R19X genotypes in the EHR did not assist in assessing hyperthyroidism, malnutrition, or malabsorption given no TG levels ≤10 mg/dL were identified in the EHR. Among the two patients that were carriers of a null variant strongly associated with cardioprotective lipid profiles, only one lacked evidence of disease highlighting the challenges of inclusion of functional genetic variation in clinical risk assessment.

Inference about sets of hypotheses

Omar de la Cruz

Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, 1311B, Cleveland, Ohio 44106

The analysis of high throughput data is often set up as a set of parallel hypothesis tests (e.g., one of each gene, or for each variant under consideration), and the result is a list of discoveries, together with an estimate of the false discovery rate. Usually, it is desirable to go further and interpret such results in biologically meaningful terms, by comparing the set of discoveries with pre-existing, annotated sets (e.g., gene pathways, or genomic regions); however, any conclusions derived this way must be statistically valid. Here we present a general setup for principled inference about sets of hypothesis, with relevance for result interpretation and for data integration.

Extracting socioeconomic data from electronic health records for gene-environment studies of blood pressure.

Brittany Hollister[1], Eric Farber-Eger[1], Dana C. Crawford[3], Melinda C. Aldrich[1,4,5], Amy Non[1,2]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN; [2]Department of Anthropology, Vanderbilt University, Nashville, TN; [3]Department of Epidemiology & Biostatistics, Institute for Computational Biology, Case Western Reserve University, Cleveland, OH; [4]Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, TN; [5]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN

Socioeconomic status (SES) is a fundamental variable contributing to health, particularly when analyzing racial disparities in health. SES data are rarely included in genetic studies, due in part to the difficultly of collecting these data when studies were not designed for that purpose. The emergence of large clinic-based biobanks linked to electronic health records (EHRs) provides research access to large populations with longitudinal phenotypic and exposure data captured in structured fields as billing codes, procedure codes, and prescriptions. SES in the EHRs, however, is often not explicitly recorded in structured fields. Rather, SES data are recorded in the free text of clinical notes and the content and completeness of these data vary widely by practitioner. To enable gene-environment studies that consider SES as an exposure, we sought to extract SES variables from BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified EHRs. We developed an index of SES using information available within the de-identified EHR, including broad categories of occupation, education, insurance status, and homelessness. We performed free-text search across 15,865 individuals (11,521 African Americans; 1,714 Hispanics, 1,122 Asians, and 1,412 others) in BioVU for 22 different categories (534 unique terms) to extract relevant SES data. We identified 14,186 individuals with education information (89%) and 14,523 individuals with occupation information (92%). Examples of information collected include years of education an individual received, degrees earned, and latest recorded occupation. Insurance status was found for 75% of the individuals, and the term homelessness was found in the EHR for 43% of the individuals. We are in the process of investigating how the extracted SES data contribute to hypertension through interactions with genetic variants previously associated with hypertension from the Illumina Metabochip in African Americans. The SES data extraction approach and index developed here will enable future EHR-based genetic studies to feasibly integrate SES data into statistical analyses. Ultimately, increased incorporation of SES measures into genetic studies and examination of gene-SES interactions will help elucidate the impact of the social environment on common diseases.

G6P missense variant (rs1671152) and risk of primary open-angle glaucoma in African Americans from a biorepository linked to de-identified electronic medical records

Restrepo, Nicole[1]; Goodloe, Robert[1]; Eger-Farber[2], Eric; Crawford, Dana C[3].
[1]Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232; [2]Vanderbilt Institute for Clinical and Translational Research, Nashville, TN 37203; [3]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106

Purpose: Primary open-angle glaucoma (POAG) is the second leading cause of permanent vision loss and blindness in the U.S. Genome-wide and candidate gene association studies have identified loci associated with POAG risk in European-descent and Japanese populations. African Americans are ~15 times as likely to develop permanent vision impairment from glaucoma vs. European Americans, yet few studies have been performed in this population. We as the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study accessed the Vanderbilt University Medical Center's BioVU, a DNA repository linked to de-identified electronic medical records, to identify cases and controls of POAG among African Americans to perform a genetic association study.

Methods: Using a combination of International Classification of Diseases diagnostic codes, Current Procedural Terminology billing codes, and manual review of clinical records, we identified 114 African American POAG cases and 1341 controls. Cases/controls were genotyped on the Metabochip, an Illumina genotype array targeting ~200,000 SNPs chosen with an emphasis on metabolic diseases and traits. After quality control, ~116k SNPs were included in analyses. We performed single SNP tests of association for common variants (MAF>0.05) using logistic regression in PLINK assuming an additive genetic model adjusted for age, sex, and first three principal components.

Results: While none of the tests of association passed a strict Bonferroni correction (p<4.3x10-7), a number were nominally significant. The five most significant associations are shown in the table.

Table: Most significant associations in African American POAG association study

						CAF(%)	
CHR	SNP	Allele	OR	CI	p-value	cases	controls
3	rs4678836	A	1.95	(1.41-2.70)	4.31x10-5	35.1	24.6
19	rs1671152	A	1.87	(1.38-2.53)	4.53x10-5	44.8	30.6
21	rs9982695	A	2.01	(1.43-2.81)	4.59x10-5	33.1	22.7
6	rs9479660	G	1.88	(1.38-2.55)	4.42x10-5	36.4	25.5
6	rs11155927	G	1.93	(1.40-2.66)	5.28x10-5	30.5	20.4

coded allele frequency (CAF); confidence interval (CI); odds ratio (OR)

Conclusions: Our study did not detect a strong association for POAG risk in African Americans on the Metabochip. Small sample sizes and lack of genome wide coverage are major limitations in this study. Of interest for future studies is rs1671152 (OR=1.87; p =4.53x10-5), a known missense variant in the glycoprotein VI (GP6) gene. GP6, a collagen receptor, is involved in platelet aggregation but is expressed in the eye and brain. Potential implications may include scleral collagen organization and integrity of the blood-retinal barrier in glaucoma susceptibility.

A machine learning based technique for predicting patient response to small molecule activators of protein phosphatase 2

Elena Svenson, Mehmet Koyuturk, PhD, Goutham Narla, MD, PhD

Integrating multi-platform multi-cancer information is becoming increasingly important for cancer drug development. In considering best positioning for a new drug compound, researchers are increasingly starting to consider molecular markers, or mutational and gene signatures, instead of traditional histological subtypes of cancer to identify predictive signatures for drug response. Here we examine mutational and gene expression information in the context of drug sensitivity, aiming to create a genetic and mutational signature predictive of cell line response to a novel series of small molecule PP2A activators. PP2A is known to dephosphorylate multiple key oncogenic signaling proteins. Since it is so central to counteracting the activation of kinase driven pathways that lead to cancer progression, PP2A has been identified as a critical target for cancer therapy. This series of SMAP compounds appear to be broadly active across all cancer types, which is perhaps to be expected considering its target. Sensitive and non-sensitive cancer cell lines do not stratify based on lineage or mutational markers alone, indicating a need for a more refined signature. In order to model patient stratification, we are using cell line data from the Cancer Cell Line Encyclopedia to survey gene expression and genetic mutations alongside a drug panel screening of an early derivative of our SMAP against 240 cell lines. We used this data to first do feature selection based on differential expression and alteration occurrences between the sensitive and non-sensitive groups, and to train a neural net with k-fold cross validation. We then fed genetic features for new cell lines (cell lines from the CCLE that were not screened for our compound into the network, to make new predictions on sensitive and non-sensitive lines. These new predictions will then be validated functionally by determining cell responsiveness using viability analysis by dosing the predicted sensitive and non-sensitive lines, and validating their relative sensitivities. Additionally, this study represents a chance to use predictive modeling of anticancer compound sensitivity in a way that can inform patient treatment stratification since it only uses pre-dosed information, predictively, it best approximates the data that will be available when a patient presents for treatment. To best approximate how this method will translate to patients, we will apply the trained model from cell lines to patient samples from The Cancer Genome Atlas. This approach is uniquely well suited for a broadly active compound whose pathway-level impacts are still relatively unknown.