Burroughs-Wellcome Fund
Maryland Genetics, Epidemiology and Medicine (MD-GEM) Pre-doctoral Training Program

Abstract Book
Genetics Research Day

February 17, 2017

Contents

Dear Participants,

On behalf of the Maryland-Genetics, Epidemiology, Medicine Training Program (MD-GEM) it is our pleasure to welcome you to the fourth annual Genetics Research Day at Johns Hopkins University. MD-GEM includes faculty spanning the Mckusick-Nathans Institute of Genetic Medicine, the Johns Hopkins Bloomberg School of Public Health, the Johns Hopkins School of Medicine and the National Human Genome Research Institute, who join together to train doctoral students in population and laboratory sciences focused on genetics.

This Genetics Research Day provides the greater JHU community an opportunity to promote discussion and collaboration across JHU/NHGRI and to integrate students from different disciplines into the wide breadth of genetics research. We welcome all faculty, post-doctoral fellows and students, especially those new to the field of genetics. We look forward to continued partnerships and new relationships across the fields of Epidemiology, Biostatistics, Human Genetics, Biology, Computer Science, Mathematics and more. The posters represent the Departments of Biostatistics, Epidemiology, and Mental Health in the Johns Hopkins Bloomberg School of Public Health; Departments of Molecular Biology and Genetics, Neuroscience, Oncology, Pathology, Pediatrics, Psychiatry and Behavioral Sciences, and the Divisions of Allergy and Clinical Immunology, Nephrology and Molecular Medicine in the Department of Medicine in the Johns Hopkins School of Medicine; the Berman Institute of Bioethics, Center for Epigenetics, Center for Inherited Disease Research, Clinical Pharmacology Analytical Laboratory, Genetics Resources Core Facility, Greenberg Center for Skeletal Dysplasias, Lieber Institute for Brain Development, McKusick-Nathans Institute for Genetic Medicine, Sidney Kimmel Comprehensive Cancer Center, and Wendy Klag Center for Autism and the Wilmer Eye Institute of the Johns Hopkins University; and the Computational and Statistical Genomics Branch and Medical Genetics Branch of the National Human Genome Research Institute.

A very special thank you to Dr. Sekar Kathiresan, Massachusetts General Hospital, Harvard Medical School, for joining us as our plenary speaker. Thank you to all faculty judges who have generously lent us their expertise and time and to whom we are indebted. We extend our sincere thanks to Sandy Muscelli, Jon Eichberger and Nicole Thornton for all of their help in organizing and promoting this event. We are especially grateful for the tireless efforts of Jennifer Deal who graciously attended to every detail to bring this day together.

Thank you for participating.

Sincerely,

Priya Duggal, PhD, MPH
Director, MD-GEM
Johns Hopkins Bloomberg School of Public Health

David Valle, MD, PHD
Director, MD-GEM
Mckusick-Nathans Institute of Genetics Medicine

Dani Fallin, PhD
Associate Director, MD-GEM
Johns Hopkins Bloomberg School of Public Health

Sekar Kathiresan, M.D.

Sekar Kathiresan, a physician scientist and a human geneticist, is the Director of the Center for Genomic Medicine (CGM) at Massachusetts General Hospital (MGH), Ofer and Shelley Nemirovsky MGH Research Scholar, Director of the Cardiovascular Disease Initiative at the Broad institute, and an Associate Professor of Medicine at Harvard Medical School.

Dr. Kathiresan leverages human genetics to understand the root causes of heart attack and to improve preventive cardiac care. Among his scientific contributions, Dr. Kathiresan has helped highlight new biological mechanisms underlying heart attack, discovered mutations that protect against heart attack risk, and developed a genetic test for personalized heart attack prevention.

Dr. Kathiresan received his B.A. in history and graduated *summa cum laude* from the University of Pennsylvania in 1992 and received his M.D. from Harvard Medical School in 1997. He then completed his clinical training in internal medicine and cardiology at MGH, where he served as Chief Resident in Internal Medicine from 2002-2003. Dr. Kathiresan pursued research training in cardiovascular genetics through a combined experience at the Framingham Heart Study and the Broad Institute. In 2008, he joined the faculties of the MGH Cardiology Division, Cardiovascular Research Center, and Center for Genomic Medicine.

Burroughs Wellcome Fund

The *Burroughs Wellcome Fund* is an independent private foundation dedicated to advancing the biomedical sciences by supporting research and other scientific and educational activities. Within this broad mission, BWF has two primary goals:

- To help scientists early in their careers develop as independent investigators
- To advance fields in the basic biomedical sciences that are undervalued or in need of particular encouragement

BWF's financial support is channeled primarily through competitive peer-reviewed award programs. A Board of Directors comprising distinguished scientists and business leaders governs BWF. BWF was founded in 1955 as the corporate foundation of the pharmaceutical firm Burroughs Wellcome Co. In 1993, a generous gift from the Wellcome Trust in the United Kingdom, enabled BWF to become fully independent from the company, which was acquired by Glaxo in 1995. BWF has no affiliation with any corporation.

http://www.bwfund.org/

Maryland Genetics, Epidemiology and Medicine (MD-GEM) Training Program

The *Maryland Genetics, Epidemiology and Medicine (MD-GEM)* is a pre-doctoral training program that comprehensively integrates Genetics, Epidemiology, and Medicine (GEM). Funded by the Burroughs-Wellcome Fund, the MD-GEM training grant brings together the expertise and training infrastructure of the Johns Hopkins Schools of Public Health and Medicine and the National Human Genome Research Institute. Together, these three institutions can provide laboratory, methodological and clinical expertise and coursework to train the next generation of scientists who can forge new avenues of research and address the rapidly changing field of human genetics. This program trains pre-doctoral students through integration of these important areas by partnering with established mentors and offering integrated learning. We envision a training program that will prepare scientists for the next generation of genetics research.

http://www.hopkinsgenetics.org/

MD-GEM Faculty

Priya Duggal, Co-Director
David Valle, Co-Director
M. Daniele Fallin, Associate Director
Dan Arking
Dimitrios Avramopoulos
Joan E. Bailey-Wilson
Terri Beaty
Aravinda Chakravarti
Debra Mathews
Ingo Ruczinski
Steven Salzberg
Diane M. Becker
Lewis Becker
Larry Brody
Nilanjan Chatterjee
Josef Coresh
Jennifer Deal
Hal Dietz
Andrew Feinberg
Gail Geller
Loyal A. Goff
Ada Hamosh
Kasper Hansen
Julie Hoover-Fong
William Isaacs
Lisa Jacobson
Corrine Keet
Alison Klein
Christine Ladd-Acosta
Jeffrey Leek
Justin Lessler
Brion Maher
Rasika Mathias
Shruti Mehta
Elaine A. Ostrander
Elizabeth A. Platz
Stuart Ray
Robert Scharpf
Alan Scott
Margaret Taub
David Thomas
Kala Visvanathan
Jeremy Walston
Xiaobin Wang
Alexander Wilson
Robert Wojciechowski
Peter Zandi

Poster No.	Presenter	Title	Page No.
1	Stephen Cristiano	Approaches for non-invasive detection of cancer	13
2	Claire Malley	Gene enrichment analysis of whole-genome sequenced patients to identify rare and deleterious genetics determinants of eczema herpeticum	25
3	Ryan Longchamps	Examining the Relationship Between mtDNA Quantity, Quality and Human Disease	23
4	Samantha Bomotti	Identifying rare and low-frequency variants associated with axial length in the Beaver Dam Eye Study (BDES)	10
5	Carrie Wright	Evaluation of isomiRNA expression in schizophrenia using small RNA sequencing of postmortem dorsolateral prefrontal cortex brain tissue	41
6	Yan Zhang	Estimating effect-size distribution from summary-level statistics for large genome-wide association studies	46
7	Michael Chou	Statistical methods for analysing Copy Number Variants associated with Hirschsprung disease	12
8	Rachel Dvoskin	Long-term impact of Huntington's presymptomatic genetic testing: interviews with at-risk individuals 20-30 years after testing	14
9	Alexandra Winters	Genetic determinants of peanut-specific immunoglobulins in the Learning Early About Peanut Allergy (LEAP) study	40
10	Yanzi Xiao	Evidence of APOBEC3 editing in the HPV16 genome	43
11	Rebecca Mitchell	Investigating the role of the SCD1 locus on Sudden Cardiac Death risk	27
12	Bracha Avigdor	Whole exome sequencing of multiple metastatic breast cancer tumors from a rapid autopsy series suggests a single origin and low heterogeneity between Metastatic sites	9
13	Johanna Robertson	Neuronal differentiation potential of HT22 cells with Kabuki Syndrome mutations in vitro	36
14	Kipper Fletez-Brant	Cross-sample normalization of HiC contact matrices	15
15	Jing You	Understanding the genetic basis of very early onset inflammatory bowel disease (VEOIBD) by using whole exome sequencing	44
16	Julie Jurgens	Investigation of a novel genetic basis for strabismus through whole genome sequencing	18
17	Woori Kim	Elucidation of the Relationship among COPD, Genes and Smoking	20
18	Christina Peroutka	Retrospective Lab Database & EMR Analysis to Identify Patients at risk of HPP	30
19	Deyana Lewis	Family-Based Rare Variant Association Study of Familial Myopia in Amish and Ashkenazi Jewish Families	22
20	Guanghao Qi	Genetic Principal Component Analysis Using Summary Level Data Identifies New Blood Lipid Loci	33
21	Kelsey Stuttgen	Perspectives on Genetic Testing and Return of Results from the First Cohort of Presymptomatically Tested Individuals At-Risk for HD	39
22	Nina Rajpurohit	Investigating Expression of Epistatic Interaction Networks in Bipolar Disorder in Human Dorsolateral Prefrontal Cortex	34
23	Julius S. Ngwa	Differential Expression Analysis of Gene and Transcript Abundance for Single Cell RNA-Seq Data using STAR and HISAT Aligners	29
24	Candace D. Middlebrooks	Identification of Candidate Genes that may Underlie a Familial Lung Cancer-linked Region within 6q	26

25	Arianna Franca	Review of Return of Results Policy, As Reflected in Clinical Trial Informed Consent Documents	16
26	Weiyan Li	Maternal use of oral contraceptives before and after conception, mid-pregnancy hormone and protein markers, and risk of autism spectrum disorder in children	21
27	Genevieve Stein-O'Brien	Genome specific transcriptional signatures predict differentiation biases in Human ES/IPS cells	38
28	Shan Andrews	Placenta methylation and autism risk in the Early Autism Risk Longitudinal Investigation (EARLI)	8
29	Jack Fu	Detection of De-Novo Copy Number Deletion from Targeted Sequencing Data	17
30	Norazlin Kamal Nor	Growth abnormalities and CNV burden in the SEED Study	19
31	Haoyu Zhang	Testing for genetic association in case-control studies incorporating multivariate disease characteristics	45
32	Tyler Bryant	FMO3 exome sequencing variants as quantitative trait loci for systolic and diastolic blood pressure	11
33	Julia Pringle	Leveraging genomics to track malaria transmission dynamics in varied epidemiologic settings in Zambia	32
34	Amanda Price	Dynamic, cell type-specific cis-regulatory element use in the developing human frontal cortex	31
35	Anthony Musolf	Chromosome 11p is Significantly Linked to Myopia in Caucasian Families	28
36	Stephanie Loomis	Genome-wide association study of serum fructosamine and glycated albumin in adults without diagnosed diabetes: results from the Atherosclerosis Risk in Communities Study	24
37	Nicholas Roberts	Transcriptomic profiling of normal pancreatic duct, precursor, and pancreatic adenocarcinoma organoids	35
38	Alan Scott	Synthetic Long Read Sequencing with Optical Mapping To Produce a High Quality de novo Genome	37

Placenta methylation and autism risk in the Early Autism Risk Longitudinal Investigation (EARLI)

Shan V. Andrews[1,2], Kelly M. Bakulski [3], Jason I. Feinberg[1,2], Rakel Tryggvadottir[4], Lisa A. Croen[5], Irva Hertz-Picciotto[6], Kasper D. Hansen[1,7], Christine Ladd-Acosta[1,2], Craig J. Newschaffer[8], M. Daniele Fallin[1,2] and Andrew P. Feinberg[4]

[1] Johns Hopkins Bloomberg School of Public Health
[2] Wendy Klag Center for Autism and Developmental Disabilities, JHSPH
[3] University of Michigan School of Public Health
[4] Center for Epigenetics, Johns Hopkins School of Medicine
[5] Kaiser Permanente Northern California
[6] UC Davis MIND Institute
[7] McKusick-Nathans Institute of Genetic Medicine
[8] A.J. Drexel Autism Institute

Presented by Shan Andrews

Background: Epigenetic mechanisms are of increasing interest in Autism Spectrum Disorder (ASD) etiology. Previous studies have established an association of altered epigenetic marks in brain and lymphoblastoid tissues of ASD cases compared to controls, including DNA methylation (DNAm) and histone 3, lysine 9 trimethyl (H3K9me3) modifications. The placenta is an important mediator of stress and environmental exposures during the gestational period, shown to be a critical risk window for neurodevelopmental disorders, and is therefore of interest for etiologic investigations of ASD. However, to date, there have been no previous genome-wide studies of placenta DNAm and ASD.

Objectives: We measured DNAm across the genome in placenta tissue to identify genomic regions at which DNAm differed according to autism risk as quantified by the Autism Observational Scale for Infants (AOSI) administered at 12 months.

Methods: We isolated genomic DNA from the fetal side of 133 placenta samples from the Early Autism Risk Longitudinal Investigation (EARLI), an ongoing autism-enriched pregnancy cohort which enrolls families with a previously diagnosed ASD child during a new pregnancy. Families are followed throughout the gestational period and infants are followed from birth through 36 months. Recruitment was carried out at 4 sites: Drexel University School of Public Health & Children's Hospital of Philadelphia, University of California Davis & MIND Institute, Johns Hopkins Bloomberg School of Public Health & Kennedy Krieger Institute, and Northern California Kaiser Permanente. Sequencing libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina by New England BioLabs Inc. Whole-genome bisulfite sequencing at 13x coverage used 125 base pair, paired-end reads with the Illumina HiSeq 2500. We are currently performing alignment using Bowtie2, calculating methylation at single-nucleotide resolution, and searching for differentially methylated regions (DMRs) according to AOSI score, adjusting for ancestry and sex using the BSmooth algorithm as implemented in the R package 'bsseq'.

Results: AOSI score was available on 115 of the 133 children with available placental samples (range 0-19, mean [sd] = 5.25 [3.86]). We will report at the meeting the top-ranked DMRs and explore their implicated regions for their potential functional relevance to ASD and towards placenta functionality more generally.

Conclusion: This study comprises the largest and most comprehensive survey of the placenta methylome in the context of ASD to date. Discovered regions may help define the role of placenta methylation in ASD etiology and may support the development of a placenta-based DNAm biomarker for autism risk.

Content Area: Computational Genetics, Genetic Epidemiology
Keywords: Placenta, Autism, DNA methylation

Whole exome sequencing of multiple metastatic breast cancer tumors from a rapid autopsy series suggests a single origin and low heterogeneity between Metastatic sites

Bracha Erlanger Avigdor[1,] Pegram Argani[2], Sarah J. Wheelan[1] and Ben Ho Park[1]

[1] Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine

[2] Department of Pathology, The Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine

Presented by Bracha Avigdor

Largely incurable and not very well elucidated, Tumor metastasis continues to be the leading cause of cancer related deaths. While metastatic tumor samples are not easily obtained, such sample from rapid autopsy series present a unique opportunity to study trajectory and progression of tumor metastasis. Recent studies have highlighted the need to understand the mechanism of tumor development and genomic landscape of metastatic breast cancer to inform clinical decisions; for example, therapy resistant ESR1 mutations may be present in one lesion and not others..

Here we survey metastatic breast cancer tumors from five patients who participated in a rapid autopsy program, including three patients with ER positive breast cancer and two patients with Triple Negative (TN) breast cancer. For each patient we performed whole exome sequencing on high quality DNA from three flash frozen tumors harvested from different organ sites.

We employed a pipeline we developed to identify somatic mutations, harnessing the ExAC dataset as a panel of normals in the absence of normal samples, to identify single nucleotide (SNV) and copy number (CNV) variants in each metastatic tumor.

While the patients shared few variants, we observed low heterogeneity among metastases within each patient, both for SNVs and CNVs. Shared SNVs included hotspot variants in TP53, PIK3CA (p.His1047Arg) and BRCA1. In one patient, we identified a p.Asp538Gly ESR1 mutation that occurred in only one metastatic tumor. These results were corroborated by copy number analysis.

These results support a model in which metastatic tumors arise from the same clone in the primary tumor, and disseminate either at an early stage or do not gain substantial numbers of mutations (possibly through tumor cell dormancy) before developing at distal sites.

Content Area: Human Genetics, Computational Genetics

Keywords: Breast Cancer, Cancer Metastasis, Exome Sequencing, Cancer genetic variation

Identifying rare and low-frequency variants associated with axial length in the Beaver Dam Eye Study (BDES)

Samantha Bomotti[1] , Priya Duggal[1], Fei Chen[1], Barbara E.K. Klein[2], Kristine E. Lee[2], Barbara Truitt[3], Ronald Klein[2], Sudha K. Iyengar[3] and Alison P. Klein[1,4,5]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[2] Department of Ophthalmology and Visual Sciences, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA
[3] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA
[4] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland, USA
[5] Department of Pathology, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Presented by Samantha Bomotti

Refractive errors, myopia and hyperopia, are the most common causes of visual impairment in the US and the world. Although studies suggest a strong genetic component, only ~3.4% of the genetic variation in refractive errors has been explained to date. The primary determinant of refractive errors is ocular axial length. The goal of this analysis is to identify rare and low-frequency variants (minor allele frequency (MAF) ≤ 5%) associated with axial length.

A subset of 1,908 individuals aged 58-94 years from the population-based Beaver Dam Eye Study (BDES) cohort were successfully genotyped using the Illumina exome array. Following standard quality control procedures, 10,397 autosomal genes and 34,981single variants from 874 individuals were available for analysis. Axial length measurements were obtained from the fourth follow-up visit of the BDES using partial coherence laser interferometry (IOL Master, Carl Zeiss, Jena, Germany). We conducted a gene-based analysis using SKAT-O to increase our power to detect the effects of rare and low-frequency variants on axial length variation. A single variant analysis was then conducted in PLINK under an additive model. All analyses were adjusted for age, sex, height, and education, and were conducted in the right eye followed by the left eye to ensure consistency of results. The Bonferroni-corrected threshold for significance was P < 4.81 × 10-6 for the gene-based analysis and P < 1.84 × 10-6 for the single variant analysis.

In the gene-based test, the RAD17 checkpoint clamp loader component gene (MIM 603139) on chromosome 5q13.2 was significantly associated (P = 3.44 × 10-6) with axial length. The single variant test supported this finding with a suggestively associated missense variant (rs35440980, c.1778G>A [MAF = 0.76%], p.R593K) in RAD17 in the right (P = 7.48 × 10-6) and left (P = 1.65 × 10-5) eyes. The adhesion G protein-coupled receptor G7 gene (ADGRG7 [MIM 612307]) on chromosome 3q12.3 was also significantly associated with axial length (P = 1.56 × 10-6). A missense variant (rs61742836, c.187G>A [MAF = 0.68%], p.D63N) in ADGRG7 was suggestively associated in the right eye (P = 8.18 × 10-5) and significantly associated (P = 1.26 × 10-6) with axial length in the left eye.

These results suggest that rare and low-frequency variants within the RAD17 and ADGRG7 genes are associated with variation in axial length. These results may help to elucidate the biological mechanism underlying ocular refraction.

Content Area: Genetic Epidemiology
Keywords: Axial Length, Refractive Error, Genetic Epidemiology, Exome Array, SNP Analysis/Discovery

FMO3 exome sequencing variants as quantitative trait loci for systolic and diastolic blood pressure

Tyler Bryant[1], Priya Duggal[1], Tariq Shafi[2], Josef Coresh[1,3] and Adrienne Tin[1,3]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
[2] Division of Nephrology, Johns Hopkins University School of Medicine, Baltimore, MD
[3] Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins Medical Institutions, Baltimore, Maryland

Presented by Tyler Bryant

Introduction: Individuals with rare loss of function variants in FMO3 have trimethylaminuria and often present with hypertension. FMO3 has been implicated in regulating blood pressure. However, studies of a common missense variant in FMO3 and blood pressure have found conflicting results.
Objectives: The goal of this study was to characterize the variants in FMO3 and examine the associations of rare and low frequency variants with systolic and diastolic blood pressure. A replication study of the association of a common missense variant, E158K, in FMO3 with hypertension was also performed.

Methods: 7272 European and 2790 African American participants in the Atherosclerosis Risk in Communities study with exome sequencing data for the FMO3 gene passing quality control and complete data on covariates were included in this analysis. We performed single-variant tests using multiple linear regressions for all variants and gene-based tests including variants with minor allele frequency < 5% to evaluate the association of FMO3 variants with systolic and diastolic blood pressure. The association of the common variant, E158K, with hypertension was examined using logistic regression.

Results: FMO3 was found to be significantly associated with systolic blood pressure in Europeans in a Sequence Kernel Association Test ($p = 0.007$). The replication study resulted in a non-significant association of the common variant E158K with hypertension.

Conclusion: Rare and low frequency variants in FMO3 may affect the ability of the enzyme to metabolize catecholamines, resulting in the dysregulation of blood pressure. Further studies of the effects of these variants on enzyme function could help elucidate the mechanisms behind this process.

Content Area: Genetic Epidemiology
Keywords: Blood Pressure, Hypertension, FMO3

Statistical methods for analysing Copy Number Variants associated with Hirschsprung disease

Michael Chou[1], Robert B. Scharpf[2], Terri H. Beaty[1] and Aravinda Chakravarti[3]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA
[3] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, 21205, USA

Presented by Michael Chou

Hirschsprung disease (HSCR) is highly heritable and rare deleterious coding variants in 24 different genes have been identified as influencing risk to this complex birth defect. Previous studies have also demonstrated different karyotypic anomalies in HSCR patients, and we have recently identified an enrichment of large deletions and duplications in at least 15 different chromosomal regions. Smaller copy number variants may also increase HSCR risk but these remain poorly studied and inherent limitations for discovering CNVs mean that only larger variants (>500 kilobases) have been adequately documented to date.

Broadly, we implement two approaches to identifying copy number variants in HSCR. One approach uses structural variants characterised and validated in the 1000 Genomes Project (1000GP) as a reference to identify a set of polymorphic Copy Number Variants (CNVs) in 295 HSCR patients and their families using exome sequence data. Our HSCR dataset has the added advantage of having genome-wide markers from the Illumina Omni2.5 platform (which was also used on the 1000GP data). We employ Bayesian hierarchical Gaussian mixture models to identify these copy number variants. We also implement Hidden Markov Modelling (HMM) to identify additional polymorphic copy number variant regions potentially not represented in 1000GP. We augment these approaches with statistical filters including assessing heterozygosity as well as the signal-to-noise ratio (SNR) of the mixture models applied to these copy number data. In addition, we refine our copy number inferences by using the family-based design of our dataset.

Based on the mixture models, we identify 309 polymorphic CNV regions consisting of 292 deletions, 15 duplications and 2 regions with both deletions and duplications. A median of 61 polymorphic copy number regions are identified for each HSCR sample. The statistical filters used for refining this collection of polymorphic CNVs form a set of rules for identifying CNVs in HSCR patients and their family members. These rules may be applied to the actual 1000GP Omni2.5 data. Therefore, the 1000GP data may serve as a reference for estimating the sensitivity of our CNV genotyping algorithms in the HSCR dataset. Our results show that this set of filtering methods provides a more comprehensive set of CNVs compared to existing methods as implemented in packages like PennCNV. These results demonstrate how reference maps may be leveraged for analysis of structural variants.

Content Area: Human Genetics, Genetic Epidemiology
Keywords: Hirschsprung Disease, Copy Number Variants

Approaches for non-invasive detection of cancer

Stephen Cristiano[1,2], Jillian Phallen[2], Jacob Fiksel[1,2], Mark Sausen[2], Vilmos Adleff[2], Rob Scharpf[1,2] and Victor Velculescu[2]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21287, USA
[2] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Presented by Stephen Cristiano

Non-invasive diagnosis, monitoring, and prognosis of cancer remain challenging aspects of clinical care. Recent approaches towards biomarker development based on the identification of tumor derived sequence mutations have met mixed success due to the major challenge of detecting circulating tumor DNA (ctDNA) at levels below 1% of total cell-free DNA (cfDNA). To overcome this challenge, we explore the promise of two orthogonal approaches for biomarker development: the identification of structural variants from deep sequencing of targeted regions and the elucidation of epigenetic signatures of cancer from whole genome sequencing. We analyzed structural alterations using targeted next generation sequencing of the HER2/neu region and utilized characteristics of rearranged paired end tags such as aberrant orientation and separation to identify rearrangements in three breast cancer patients and three healthy individuals. Our analyses of cfDNA revealed a stark difference in the genomes; rearrangements indicative of HER2/neu amplification were found in the plasma of all breast cancer patients, whereas no structural alterations were identified in plasma of healthy individuals. Analysis of the matched tumor tissue for the breast cancer cases revealed concordance with the rearrangements identified through analyses of ctDNA. To assess epigenetic signatures in cancer, we analyzed DNA fragmentation patterns through whole genome sequencing of cfDNA from 11 individuals with colorectal cancer and 20 healthy individuals. Our analyses suggest differential genomic representation of cfDNA in cancer patients compared to healthy individuals. The development of methods for non-invasive detection of structural alterations and fragmentation patterns in cfDNA represents a novel approach for screening, early detection and monitoring with significant impact for cancer patients.

Content Area: Other
Keywords: cancer genomics, circulating tumor DNA, nucleosomes, next-generation sequencing, structural variants

Long-term impact of Huntington's presymptomatic genetic testing: interviews with at-risk individuals 20-30 years after testing

Rachel Dvoskin[1], Juli Bollinger[1], Allison McCague[1], Kelsey Stuttgen[1,2], Barnett Shpritz[3], Jason Brandt[3] and Debra Mathews[1]

[1] Johns Hopkins Berman Institute of Bioethics
[2] Johns Hopkins Institute of Genetic Medicine
[3] Johns Hopkins Department of Psychiatry and Behavioral Science

Presented by Rachel Dvoskin

Background: To learn about the long-term impacts of genetic testing, we are conducting a follow-up study of participants from the presymptomatic testing program of the Johns Hopkins' Baltimore Huntington's Disease Project (BHDP), one of the first such programs in the U.S.

Methods: For the first phase of the study, we held semi-structured interviews with at-risk probands who enrolled in the BHDP from 1986–1998. We used indirect recruitment methods (flyers posted in the HD clinic, outreach through Huntington Disease Society of America and its support groups, and Facebook) and sent recruitment letters directly to BHDP participants. Here we report on interviews with 35 participants, 20 of whom tested with a normal repeat and 15 of whom tested with an expanded repeat. We asked participants how their testing experience and knowledge of the result had influenced their life over the past 20-30 years, including decisions about career, children, and relationships. A codebook was developed by the study team through an iterative process. After inter-coder reliability was obtained, each interview transcript was coded by two team members, reconciled, and entered into Nvivo 11 software. Code reports were run, and analysis was performed for identification of themes.

Results: Nearly all participants reported positive feelings about their decision to enter the BHDP and their testing experience. Most believed knowing their risk status had a major impact on their life; some said they did not live life differently as a result. While people varied in how they reacted to, processed, and communicated their result, many participants spontaneously brought up family relationships as they discussed the most salient impacts of testing on their life. Some individuals felt closer to certain family members after knowing their result, and some reported being better informed/able to help children with the testing decision process or to support relatives through HD or other presymptomatic testing. On the other hand, a number of people reported strained family relationships related to their decision to be tested: for example, a few discussed anger among family members about their testing decision; rifts between at-risk siblings (and other relatives) who wanted to be tested and those who did not; or the experience of survivor's guilt or a feeling of being "different" from other siblings after receiving a normal repeat result.

Conclusions: Many of us think of the decision to learn one's HD status as a highly personal, individual choice. Yet we learned that family communication and relationships played a large role in both the decision making and impact of HD testing on people's lives. Whether there is open communication or silence around the topic of HD—and whether there is support and acceptance of personal decision making around testing—will contribute to the impacts of testing (and test results) on individuals and their family relationships. These findings highlight the importance of considering an individual's family history and current dynamics, as well as an individual's communication style and current social/familial environment to be able to provide appropriate counseling around presymptomatic testing.

Content Area: Human genetics, Other
Keywords: ELSI, presymptomatic testing, Huntington's disease

Cross-sample normalization of HiC contact matrices

Kipper Fletez-Brant[1], David Gorkin[2], Yunjiang Qiu[2] and Kasper Hansen[1]

[1] JHU IGM and Biostatistics
[2] Ludwig Institute for Cancer Research, UCSD

Presented by Kipper Fletez-Brant

The 3D structure and arrangement of the genome within the nucleus is the focus of active research and has spurred the development of a variety of assays, including HiC, which uses a unique strategy to map pairs of interacting loci genome-wide. Current normalization methods for HiC experiments treat individual samples one at a time, but do not have capabilities for normalizing data across samples. We address this issue and develop a novel technique to normalize multiple samples together, including the ability to correct for batch effects.

Content Area: Human Genetics, Computational Genetics, Statistical Genetics
Keywords: genetics, genomics, statistics

Review of Return of Results Policy, As Reflected in Clinical Trial Informed Consent Documents

Arianna Franca[1] and Debra Mathews[2]

[1] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine
[2] Johns Hopkins Berman Institute of Bioethics, Johns Hopkins University

Presented by Arianna Franca

The recent launch of the Precision Medicine Initiative, the move toward learning health systems, and the changing role of patients in research necessitates that we reconsider the responsibility of an investigator to communicate research results to participants. This reconsideration has been taking place in earnest for several years, through theoretical and empirical work on the return of results and incidental findings (IFs), resulting in several recommendations and guidances from scholars and organizations. In an effort to further inform this debate, we sought to understand the current landscape of practice regarding the return of results, as assessed through an analysis of informed consent documents. We began with a survey of institutional guidelines and/or policies (G/P) regulating informed consent at the top ten NIH-funded U.S. research institutions (USRI). Of these institutions, 8/10 had a G/P mentioning the return of individual results (IR). About half of these (4/10) were specific to genetic results. Five institutions had a G/P regarding IFs. Of these five, one was specific to imaging results, three were specific to genetic results, and one covered both genetic and overall IFs. We next analyzed 39 publicly available informed consent forms (ICFs) from clinical trials published in the New England Journal of Medicine (NEJM) between January 2014 and March 2016. Of the 39 ICFs, 35 addressed the return of research results. Of these 35 ICFs, 22 stated that they would return IR. Of these 22, 13 included specific exceptions regarding the return of results. A second set of ICFs derived from genetics-focused research projects were also analyzed, and included 10 ICFs; 7 that were procured by emailing corresponding authors of clinical trials performed at USRI and published in NEJM between January 2014 and March 2016 with the word “gene” in the title or abstract, and three from three of the five Centers for Mendelian Genomics. Seven of the 10 ICFs mentioned the return of IR. Of these, 6 stated that IR would be returned, and all 6 included specific exceptions. Our findings indicate that there is a tremendous amount of variation in policies, guidelines and practice with regard to the return of IR. We recommend additional guidance and outreach to provide uniformity among institutions in order to decrease variation and ensure research subjects at minimum understand what results will and will not be returned from any trial in which they are being asked to participate.

Content Area: Human Genetics
Keywords: ELSI, genomics, clinical testing, policy issues

Detection of De-Novo Copy Number Deletion from Targeted Sequencing Data

Jack Fu[1], Terri H. Beaty[2], Alan F. Scott[3], Mary L. Marazita[4], Elizabeth J. Leslie[4], Ingo Ruczinski[1], Robert B. Scharpf[5]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public
[2] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
[3] Center for Inherited Disease Research and Institute of Genetic Medicine, Johns Hopkins School of Medicine
[4] Department of Oral Biology, University of Pittsburgh
[5] Department of Oncology, Johns Hopkins School of Medicine

Presented by Jack Fu

Copy number variants (CNVs) defined as gains or losses of large genomic segments that alter the copy-neutral diploid state of DNA are a major contributor of genome variability in humans, and frequently underlie the etiology of disease. De novo CNVs delineated in case-parent trios, albeit very rare, are of particular interest for their potential to have a functional role in the genesis of the disease phenotype. We developed a novel method based on the MinimumDistance statistic (Scharpf et al, 2012) for delineating de novo copy number deletions simultaneously across multiple trios from targeted sequencing data, dramatically lowering the false positive rate while maintaining sensitivity. We applied our method to 1,305 case-parent trios with targeted sequencing data of regions previously implicated in oral cleft. Across the 6.3Mb of capture, we detected 1 de-novo deletion in gene TRAF3IP3 [chr1: 209,945,711-209,947,360], in addition to 1 rare inherited deletion at chr8: 130,113,679-130,132,634, and 2 copy number polymorphic regions at chr1: 210,078,418-210,085,977 & chr8:129,762,791-129,766,160, respectively. These calls are further supported by read-pairs with correspondingly long insert lengths.

Content Area: Statistical Genetics, Genetic Epidemiology
Keywords: Targeted Sequencing, CNV, Trios, Minimum Distance, Oral Cleft

Investigation of a novel genetic basis for strabismus through whole genome sequencing

Julie Jurgens[1], Martha Brucato[2], Dimitri Avramopolous[1], Elizabeth Pugh[1], Dane Witmer[1], Hua Ling[1], Kurt Hetrick[1], Kim Doheny[1], Christine Ladd-Acosta[2], David Valle[1], Nara Sobreira[1] and Jeremy Nathans[3]

1 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD
2 Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
3 Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD

Presented by Julie Jurgens

Strabismus, or misalignment of the eyes, is a common disorder affecting 1-4% of the population. Despite evidence for high heritability, the molecular basis of strabismus remains poorly understood. Parikh et al. (2003) described a 3-generation pedigree (19 affected, 26 unaffected), in which strabismus appeared to be inherited as an autosomal dominant trait but was linked to a region on chromosome 7p14.1 -7q21.11 under an autosomal recessive inheritance model. To search for novel genetic causes of strabismus, we performed whole exome sequencing (WES) on 7 affected family members from this pedigree. This strategy failed to identify rare functional variants co-segregating with strabismus inside or outside of the previously identified linked region on chromosome 7. Next, we repeated linkage analysis using SNP array data (maximum LOD for the samples studied = 2.4) and identified four linked regions with positive LOD scores. We identified three linked regions with LOD scores of 2.4 on chromosomes 6p22.3-6p23, 17q21.33-17q24.3, and 19p13.11-19q13.31 that fit an autosomal dominant model of inheritance. We also identified a linked region on chromosome 7p21.1-7p21.3 fitting an autosomal recessive model of inheritance and overlapping the previously described linked region. To investigate variants in the linked regions that were missed by WES, we performed whole genome sequencing (WGS) on four severely affected and genetically distant members of this pedigree. For WGS, libraries were prepared according to Illumina TruSeq PCR-Free protocol and sequenced on an Illumina HiSeq X Ten using 125 bp paired-end reads. Illumina Real Time Analysis software was used for intensity analysis and base calling. Files were aligned to the human reference genome using BWA-MEM. GATK was used for realignment around indels and base call quality score recalibration. Bina RAVE software was used to convert FASTQ files into VCF files. Variants were filtered using Variant Quality Score Recalibration, genotype refinement was performed using CalculateGenotypePosteriors (GATK), and ANNOVAR was used for variant annotation. Using the PhenoDB variant analysis tool (Sobreira et al., 2015), we started by analyzing the WGS data in a similar manner to WES, prioritizing rare (MAF<0.01) functional variants (missense, nonsense, splicing, and indels) that are shared among affected individuals inside or outside of the four linked intervals. This strategy yielded no variants co-segregating with strabismus that had not been identified by the WES approach. Currently, we are applying annotation tools such as ENCODE, Roadmap Epigenomics, GWAVA, CADD, VISTA Enhancer, and miRdSNP to prioritize noncoding variants within the linked intervals. We are using BRAVO (TOPMed) and gnomAD databases for additional frequency-based filtering of noncoding variants. We are also considering alternative genetic models, including potential oligogenic contributions of variants in the linked interval(s). Once novel candidate genes and variants are identified, we will proceed with their validation and segregation analysis in additional family members. We anticipate that identification of the causal genetic variant(s) in this family will identify relevant biological systems for strabismus.

We would like to thank the Burroughs-Wellcome Fund, the Maryland-Genetic, Epidemiology and Medicine Training Program, the Visual Neuroscience Training Program, and the Baylor-Hopkins Center for Mendelian Genomics for their support.

Content Area: Human Genetics
Keywords: Whole genome sequencing, strabismus, linkage analysis, next-generation sequencing, noncoding variant annotation

Growth abnormalities and CNV burden in the SEED Study

Norazlin Kamal Nor[1], Dani M. Fallin, Ph.D.[2], Terri Beaty, Ph.D.[1], Julie Hoover-Fong, M.D., Ph.D.[3] and Chris Ladd-Acosta, Ph.D.[1]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
[2] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
[3] Greenberg Center for Skeletal Dysplasias, Johns Hopkins Medical Institutes

Presented by Norazlin Kamal Nor

Introduction: The SEED Study is a multi-site case-control study for risk factor identification in autism spectrum disorder. Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with variable presentation. Children were classified as ASD cases or as having developmental delay (DD) or normal development (POP). Multiple risk factors have been implicated in ASD including genetic, environmental and gene-environment interactions. Understanding the construct of ASD involves consideration of ASD phenotypes such as dysmorphology and abnormal growth. The aim of this study is to assess (i) the association between CNV burden and abnormal growth, and (ii) the association between ASD and abnormal growth, in the SEED Study.

Methods: This is a cross-sectional study with outcomes assessed at one point in time. The eligibility criteria were children born in the study catchment area between September 1st. 2003 and August 31st. 2006, aged 30-68 months at the completion of the clinical developmental assessment. Exclusion criteria included children with chromosomal abnormalities, genetic syndromes, birth defects and congenital abnormalities. ASD status was determined using Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R).

Methods for Aim (i): Biospecimens were collected and a total of 956 total cases and controls were genotyped at 1 million SNP markers on Illumina Human Omni1 array. Using the PennCNV calling algorithm, a hidden Markov model approach was used to detect CNVs using SNP array data. Based on self-reported race and ethnicity, the genotyped study sample was stratified by ancestral population to European Caucasian (EUR), African-American (AFR) and mixed (ADM). Quality control measures were instituted. Samples were excluded if < 98% of markers were called successfully, cryptic relatedness was suspected by IBS/IBD estimation, sex inconsistencies were noted and if there was excess heterozygosity/homozygosity present. SNPs were excluded if the call rate was < 0.95, minor allele frequency (MAF) was < 0.01 and the Hardy-Weinberg p-values < 1.0x10-8 in controls. CNV calls based on <10 SNPs and < 30 kb in size, and those occurring near or in centromere and telomere regions were filtered out. The association between genome-wide CNV burden and abnormal growth was assessed using logistic regression.

Methods for Aim (ii): SEED study participants underwent a physical examination including anthropometric measurements as part of dysmorphology assessments. The growth measures were plotted on CDC growth charts. One of the child's parents also underwent height and head circumference measurements.

Growth abnormalities were assessed in three ways, firstly the estimation of the three modalities of growth (height, weight, head circumference) and BMI, secondly the estimation of the genetic potential for growth or 'growth cordance', and finally the estimation of growth symmetry amongst the three growth modalities ('trivariate growth phenotype').

The association between ASD and abnormal growth was assessed using logistic regression for each growth abnormality measure, stratifying by sex.

Conclusion: The results of the analysis on the association between CNV burden and abnormal growth, and the estimation of the association between ASD and abnormal growth and growth patterns in the SEED Study is hoped to add to our understanding of ASD construct and etiology.

Content Area: Genetic Epidemiology
Keywords: Growth abnormalities, Autism Spectrum Disorder, Copy Number Variants, SEED Study

Elucidation of the Relationship among COPD, Genes and Smoking

Woori Kim[1], Margaret Parker[2], Michael Cho[2,3] and Terri Beaty[1], COPDGene Investigators

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA
[3]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA

Presented by Woori Kim

Chronic Obstructive Pulmonary Disease (COPD) is a major public health problem as the third leading cause of death in the United States (Hoyert and Xu 2012). COPD is a complex disease influenced by multiple environmental factors, genetic factors and their potential interactions. Smoking is the strongest environmental risk factor for COPD, but only 15-20% of smokers ever develop COPD. Genetics may explain why some smokers will develop COPD, but others will not. Although genome-wide association studies (GWAS) have identified several common variants associated with COPD, collectively these markers explain only small portion of the total phenotypic variance in spirometric measures of lung function, called to the 'missing heritability'. One of the possible explanation is potential gene-by-environment (GxE) interaction. A GxE interaction study will help define the most susceptible population to COPD and could help establish a tailored intervention strategy for COPD. Moreover, from recent GWAS results, some genetic variants were associated with not only decline in lung function but also nicotine addiction (Wain et al. 2015). It is conceivable that the relationship between genetic variants and risk of COPD is mediated through smoking behaviors. This proposal aims to elucidate the relationship between genes, smoking behavior and risk of COPD using the COPDGene study and the UK BiLEVE study. We will investigate genetic variants yielding evidence of gene-by-smoking interaction on risk to COPD using GWAS data, and assess whether smoking behavior acts as a mediator on the association between genetic factors and risk of COPD. This proposal will enhance our understanding of the genetic etiology of COPD and help establish personalized prevention and treatment strategies that might reduce the burden of the COPD.

Content Area: Human Genetics
Keywords: Chronic Obstructive Pulmonary Disease (COPD), Gene by Environment Interaction, Smoking, Genetics

Maternal use of oral contraceptives before and after conception, mid-pregnancy hormone and protein markers, and risk of autism spectrum disorder in children

Weiyan Li[1], Brian K. Lee[2], Nicole Gidaya[2], Craig Newschaffer[2], Laura A. Schieve[3], Diana E. Schendel[4], Nicole Jones[5], Julie L. Daniels[6], Gayle C. Windham[7], Lisa A. Croen[8], Andrew P. Feinberg[9], M. Daniele Falllin[10] and Christine Ladd-Acosta[1]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
[2] Department of Epidemiology and Biostatistics and the A.J. Autism Institute, Drexel University School of Public Health, Philadelphia, PA
[3] National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, GA
[4] Department of Public Health, Institute of Epidemiology and Social Medicine, Aarhus University, Aarhus, Denmark, Department of Economics and Business, National Centre for Register-based Research, Aarhus University, Aarhus, Denmark, and Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Denmark
[5] Biomedical Research Informatics Core, Michigan State University, East Lansing, MI
[6] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC
[7] California Department of Public Health, Richmond, CA
[8] Kaiser Permanente Division of Research, Oakland, CA
[9] Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, MD
[10] Department of Mental Health, Johns Hopkins School of Public Health, Baltimore, MD

Presented by Weiyan Li

Autism spectrum disorder (ASD) is a group of devastating neurodevelopmental syndromes that affect up to 1 in 68 children. The most striking and consistent observation in ASD is a highly skewed male to female ratio (approximately 4 to 1). The extreme male brain theory proposes that characteristics of ASD are presentations of male-typical characteristics on the extreme end of a spectrum. Estrogen plays a role in masculinization of the brain as well as sex-dimorphic behavior, and may be involved in the etiology of ASD. Both animal and human studies suggest that estrogen exposure may be associated with ASD risk. In zebrafish mutants, a behavior phenotype caused by an autism gene was alleviated by estrogenic compounds. Also mothers of ASD children were shown to have lower unconjugated estriol (E3) level in blood during pregnancy compared that of mothers of controls. In the US, over 10 million women are exposed to highly potent synthetic estrogen and progesterone through use of oral contraceptives (OC). Failures in contraception lead to unintended exposure of highly potent synthetic estrogen among over a half of a million fetuses. Even with this wide spread use of OC, no studies have examined the relationship between OC use around the time of conception and ASD in children. In this study, we plan to investigate the relationship between estrogen related exposure and ASD risk by looking at (1) the maternal OC use before and after conception, and (2) mid-pregnancy hormone and protein biomarker levels, respectively. Also, since both genes and environment heavily influence ASD risk, we will also investigate how genetic and epigenetic mechanisms contribute to estrogen associated ASD risk. DNA methylation regulates gene expression and is responsive to environmental exposure, thus we will examine whether maternal OC use leads to changes in DNAm profiles in children, and whether this change mediates ASD associated with maternal OC use. Lastly, we will examine whether polymorphisms in genes involved in the synthesis, metabolism, transport and ligand binding of estrogen affects the potential relationship between OC use and ASD risk. This study aims at investigating the effect of several estrogen related hormone and protein exposures during pre-conception and prenatal period on ASD risk in offspring through environmental, genetic and epigenetic measurements to assess the hormone etiology of ASD.

Content Area: Genetic Epidemiology
Keywords: Autism spectrum disorders, estrogen, oral contraceptives, DNA methylation

Family-Based Rare Variant Association Study of Familial Myopia in Amish and Ashkenazi Jewish Families

Deyana Lewis[1], Claire L. Simpson[2], Anthony M. Musolf[1], Laura Portas3, Federico Murgia[3], Dwight Stambolian[4] and Joan E. Bailey-Wilson[1]

[1] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland
[2] Genetics, Genomics and Informatics, University of Tennessee Health Sciences Center, Memphis, Tennessee, United States
[3] Institute of Population Genetics, CNR, Li Punti, Sassari, Italy Ophthalmology
[4] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States

Presented by Deyana Lewis

Myopia is a common refractive error (RF) which affects at least a third of most populations. Genome-wide association studies (GWAS) and linkage studies have identified loci influencing the risk of developing myopia but, few causal variants have been identified.

We have performed family based association analyses examining common and rare variants for 37 Amish and 63 Ashkenazi Jewish families with strong family history of myopia from the Penn Family Study using exome-targeted genotyping array. Myopia was defined if their average RF was <= -1 Diopter (D) and were considered unaffected if their average RF was > 0.0 D; others were coded as having an unknown phenotype. Stringent rules were used to code children as unaffected since children's eyes become more myopic during childhood and adolescence. Two variants (rs135 and rs136) for Amish families and one variant (rs6972578) for Ashkenazi Jewish families, all in the same gene (OSBPL3) were suggestively associated (p < 1.4 x10-4) with common myopia under a significant linkage peak previously detected in a set of African American families from the Penn Family Study.

To follow-up this observation of association of two variants in the same gene in two different samples, we are using rare-variant transmission disequilibrium test (RV-TDT) to perform gene-based tests with the rare variants in these exome chip data. The RV-TDT framework can control for both admixture and substructure and thus avoid spurious associations. This method has the potential to improve our power to detect causal genetic variants.

Content Area: Genetic Epidemiology
Keywords: Myopia, complex disease, genetic epidemiology, rare variants, exome chip

Examining the Relationship Between mtDNA Quantity, Quality and Human Disease

Ryan Longchamps[1,2], Rebecca Mitchell [1,2], Megan Grove[3], Eric Boerwinkle[3] and Dan Arking[1]

[1] Mckusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore MD
[2] Predoctoral Training Program in Human Genetics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD
[3] Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX

Presented by Ryan Longchamps

Mitochondria play a critical role in energy metabolism and have been linked to human disease and mortality. Several biological process have been hypothesized to explain the role of mitochondrial dysfunction in disease, such as declines in energy production, altered rates of apoptosis, and elevated free radical production. Previous reports have shown mitochondrial DNA (mtDNA) may be the key component underlying mitochondrial dysfunction, and thus human disease. Our lab has recently shown mtDNA quantity, also known as mtDNA copy number (mtDNA CN) is associated with coronary artery disease (CAD), cardiovascular disease (CVD), sudden cardiac death (SCD), and overall mortality. However, little is known about how an accumulation of mtDNA mutations, resulting in increased heteroplasmy, affect the quality of mtDNA. We hypothesized as the heteroplasmic content of an individual's mtDNA increased; mitochondrial dysfunction would increase resulting in human disease. Additionally, we hypothesized mtDNA quality may affect mtDNA replication resulting in lower mtDNA CN.

Utilizing whole genome sequence data from 404 participants of the Atherosclerosis Risk in Communities (ARIC) cohort we initiated a pilot study to test our hypotheses. mtDNA CN and heteroplasmy were called using mitoAnalyzer – a software package specifically developed to analyze mtDNA sequence data. mtDNA CN is calculated as the observed ratio of sequence coverage between the mtDNA and autosomal DNA. Heteroplasmy was called as any site with at least 4X nuclear DNA coverage and 4% of reads with the alternative allele. We developed a mtDNA quality score whereby we summed the percentages of all heteroplasmic sites for an individual. The final mtDNA quality score represented the standardized residuals from a linear model adjusting for mtDNA coverage and ARIC collection center.

Although an age and sex effect is often seen in mtDNA CN, no such effect was observed with our mtDNA quality score (P = 0.45, P = 0.31 respectively). Initial analyses in our cohort revealed mtDNA quality and mtDNA quantity are not correlated (R = -0.008, P = 0.87). Additionally, mtDNA quality was not associated with prevalent or incident CAD (P = 0.42, P = 0.65), prevalent or incident stroke (P = 0.13, P = 0.11), prevalent CVD (P = 0.24), or prevalent diabetes (P = 0.52). Interestingly, our mtDNA quality score showed nominal significance with incident CVD (P = 0.09). Although none of our findings are statistically significant, the direction of effect of each association supports our hypothesis. Given our modest sample size we believe this finding deems further investigation with a larger sample and more refined quality score.

Content Area: Human Genetics, Molecular Genetics, Statistical Genetics
Keywords: Mitochondria, Heteroplasmy, Cardiovascular Disease

Genome-wide association study of serum fructosamine and glycated albumin in adults without diagnosed diabetes: results from the Atherosclerosis Risk in Communities Study

Stephanie J. Loomis, MPH, Nisa M. Maruthur , MD, MHS, Man Li, MS, PhD, Kari North, PhD, Hao Mei, MD, PhD, Alanna Morrison, PhD, April Carson, PhD, MSPH, Laura Rasmussen-Torvik, PhD, Abigail Baldridge, James Pankow, PhD, Eric Boerwinkle, PhD, Robert Sharpf, PhD, Priya Duggal, PhD, Josef Coresh, MD, PhD, Elizabeth Selvin, PhD, MPH, Anna Kottgen, MD, MPH

Presented by Stephanie Loomis

Introduction: Fructosamine and glycated albumin are potentially important alternatives to hemoglobin A1c (HbA1c) for diabetes management and diagnosis, but the genetics of these biomarkers has not been investigated.

Methods: We performed a genome-wide association study for the log of fructosamine and log of glycated albumin (total and percent) among persons without diabetes in blacks (N=2,090) and whites (N=7,586) in the Atherosclerosis Risk in Communities (ARIC) study. Significant findings were replicated in the Coronary Artery Risk Development in Young Adults (CARDIA) study. We also evaluated genetic variants associated with traditional measures of hyperglycemia (fasting glucose or HbA1c) for association with fructosamine and glycated albumin.

Results: We found four significant associations, three of which have not been shown to be associated with glycemic traits. Among whites, rs10419198, an intronic SNP in the *RCN3* gene, was associated with fructosamine (beta=-0.12, p=7.8x10^{-10}) and total glycated albumin (beta=-0.13, p=2.5x10^{-10}). Among blacks, an intergenic SNP, rs2438321, was associated with fructosamine at a genome-wide significant level (beta=0.29; p=3.5x10^{-9}). Also among blacks, rs59443763, an intronic variant in *PRKCA*, was associated with percent glycated albumin (beta=0.36, p=5.3x10^{-9}). Among whites, rs1260236, a known missense mutation in the *GCKR* gene, was associated with percent glycated albumin (beta=0.10, p=7.7x10^{-10}) as well as with fasting glucose (beta=0.09, p=3.6x10^{-8}). Over one-half of previously-identified fasting glucose or HbA1c SNPs were nominally associated with fructosamine or glycated albumin.

Conclusions: We found both glycemia-related and novel genetic variants associated with fructosamine and glycated albumin, adding the understanding of these biomarkers.

Content Area: Genetic Epidemiology

Gene enrichment analysis of whole-genome sequenced patients to identify rare and deleterious genetics determinants of eczema herpeticum

Claire Malley[1] and Rasika Mathias[1]

[1] Johns Hopkins University

Presented by Claire Malley

Current understanding of the genetic loci and pathways implicated in the occurrence atopic dermatitis (AD) is limited, and particularly so for the rarer condition of of recurrent eczema herpeticum (EH) in AD. Through the NIAID sponsored Atopic Dermatitis Research Network (ADRN) we use high-throughput whole genome sequencing data to bridge these gaps in knowledge. We hypothesize that AD subjects with EH represent a rare disease phenotype potentially caused by mutations with functionally damaging effects as compared to a AD or non-atopic (NA) control population. We expect such mutations to belong to skin barrier and immune activation pathways. DNA from 491 NA, 48 ADEH, and 238 AD subjects were whole-genome sequenced on the Illumina HiSeq platform. We generated sequence annotations for SNPs and indels with ANNOVAR software. A custom pipeline screened for exonic SNPs with functional annotations that were frameshift, nonsynonymous, missense, stopgain, or stoploss. We required a damaging prediction from either, or both, the SIFT and PolyPhen 2 algorithms. In these two sets, we compared carrier status for all damaging variants versus those that are novel or with less than or equal to 5% global frequency in the 1000 Genomes Project database. Carrier status for rare indels with exonic annotations was separately analyzed and compared between the groups. Interim results point to enrichment in EH for damaging SNPs in cell growth and calcium ion binding pathways. Next steps will be to validate the finding in-silico and with targeted sequencing of patients.

Content Area: Human Genetics, Computational Genetics, Genetic Epidemiology
Keywords: eczema, dermatitis, WGS, annotation, immune system

Identification of Candidate Genes that may Underlie a Familial Lung Cancer-linked Region within 6q

Candace D. Middlebrooks[1], Anthony M. Musolf[1], Claire L. Simpson[2], Susan M. Pinney[3], Mariza deAndrade[4], Colette Gaba[5], Ping Yang4, Ming You[6], Anne G. Schwartz[7], Christopher I. Amos[8] and Joan E. Bailey-Wilson[1]

[1] National Human Genome Research Institute, National Institutes of Health, Baltimore, MD
[2] University of Tennessee Health Science Center, Memphis, TN
[3] University of Cincinnati College of Medicine, Cincinnati, OH
[4] Mayo Clinic, Rochester, MN
[5] University of Toledo Dana Cancer Center, Toledo, OH
[6] Medical College of Wisconsin, Milwaukee, WI
[7] Karmanos Cancer Institute, Wayne State University, Detroit, MI
[8] Geisel School of Medicine, Dartmouth College, Lebanon, NH

Presented by Candace D. Middlebrooks

Linkage studies using highly aggregated lung cancer (LC) families from the Genetic Epidemiology of Lung Cancer Consortium (GELCC) have identified a region on 6q that shows genome-wide significant linkage to LC. Targeted DNA sequencing was performed under the linkage peak in the 9 most strongly linked families to identify rare variants (RVs) that may explain the signal in each family. However, there are numerous candidate genes in this region and tissue samples for functional validation of this analysis are non-existent. This region shows different broad linkage peaks in each family which may be due to diversity in the genetic source of the signal between the families. The strongest linkage signals in each family are due to RVs located in intergenic regions. To advance this study, we used expression, genotype and clinical data from The Cancer Genome Atlas (TCGA). We sought to determine if there are any common variants near the most significant RVs from each family that play a regulatory role in expression of any genes in the region. If any of the most significant RVs in our families are located in regulatory elements that have been shown to regulate nearby cancer-relevant genes, then these RVs will be considered good candidates for follow-up by targeted sequencing in additional familial cases and model organism studies by the GELCC

We first extracted Affymetrix SNP 6 chip genotype data from 582 adenocarcinoma (the most common lung cancer subtype) patients in TCGA as well as the accompanying RNASEQ2 expression and clinical data. We then performed expressed quantitative trait (eQTL) analysis using R Matrix eQTL. We found several eQTL near or within the family-specific linkage peaks that were associated with expression of various nearby genes. Additionally, we found multiple eQtLs within high linkage peaks for several families that were associated with NOX3 and ALDH8A1. This analysis warrants further laboratory analysis of their putative functional role in lung cancer.

Content Area: Genetic Epidemiology
Keywords: Genetic Epidemiology, Cancer Biology, Human Genetics, Lung Cancer

Investigating the role of the SCD1 locus on Sudden Cardiac Death risk

Rebecca Mitchell[1], Foram Ashar[1], Juhani Junttila[2], Heikki Huikuri[2,3] and Dan E. Arking[1]

[1] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[2] Division of Cardiology, Department of Internal Medicine, University of Oulu, Finland
[3] Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

Presented by Rebecca Mitchell

Recently, there has been considerable effort into identifying genetic variants that influence risk for coronary artery disease (CAD). Despite advances in preventative therapies and treatment of CAD, sudden cardiac death (SCD) remains to be one of the leading causes of death with approximately 4-5 million cases worldwide per year.1 Given that SCD is often the first manifestation of heart disease, and is largely fatal (8.3% survival rate in US in 2012)2, early identification of individuals at increased risk for SCD has the potential to be life-saving. SCD occurs as a result of multiple different pathologies, including heart diseases such as CAD and cardiomyopathies, as well as electrical defects.3 This heterogeneity contributes to the challenge in identifying variants associated with SCD risk, such as those identified through genome-wide association studies (GWASs). However, a GWAS performed by Arking et al. identified a SNP within the 2q24.2 locus, which they named the SCD1 locus, associated with a moderate increase in risk for SCD (1.92-fold per allele increase in SCD risk).4 The identified SNP was located within an intron; was not found to be in high LD with a known functional variant; and was not found to be an eQTL. While this GWAS implicated the SCD1 locus, it was not able to implicate a casual gene. This requires further investigation of the SCD1 locus using sequencing. We propose to perform sequencing of the SCD1 region, as well as other loci of interest, in additional SCD cases and controls. We believe this will provide us the opportunity to interrogate the entire SCD1 region to attempt to identify the potential functional variant(s) which underlie the GWAS signal. We further hypothesize that the casual gene within the SCD1 locus will be enriched for rare functional variants in SCD cases as compared to controls. Identifying such genes would provide evidence towards their association with SCD risk. We hypothesize that one or more genes within the SCD1 locus influence SCD risk.

Content Area: Human Genetics
Keywords: Sudden cardiac death, cardiovascular disease, sequencing

Chromosome 11p is Significantly Linked to Myopia in Caucasian Families

Anthony Musolf[1], Claire L. Simpson[2], Bilal A. Moiz[1], Kyle A. Long[1], Laura Portas[3], Federico Murgia[3], Elise Ciner[4], Dwight Stambolian[5] and Joan E. Bailey-Wilson[1]

[1] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland
[2] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee
[3] Institute of Population Genetics, CNR, Li Punti, Sassari, Italy
[4] Salus University, Philadelphia, Pennsylvania
[5] Department of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania

Presented by Anthony Musolf

Myopia is one of the most common forms of visual impairment in the United States, with nearly one in four Americans affected. The genetics involved in myopia have yet to be completely understood. This study uses Caucasian families with a history of myopia to find linkage between the disease and variants in the genome. Our sample data consisted of 56 Caucasian families with a history of myopia. Individuals were genotyped on the Illumina Exome Array and merged with previously obtained microsatellite data. Individuals were coded as either affected with myopia, unaffected, or unknown. Three types of parametric linkage analyses were performed: standard single variant two-point linkage, multipoint linkage, and collapsed haplotype pattern variant linkage (CHP). CHP linkage involves creating a multi-allelic pseudomarker from multiple single variants to boost information content. Two-point linkage is then run on the pseudomarker, which corresponds to a genetic region. Family-based association analyses using rare and common variants are underway. Single variant two-point analysis revealed three genome-wide significant SNPs (HLOD >= 3.3) located at 11p15.4, 11p15.1, and 11p11.2. An additional 47 suggestive SNPs were also found in the 11p region. CHP linkage analysis identified three suggestive variants on 11p and multipoint analysis identified two suggestive variants. 11p is an excellent candidate region for myopia, as it contains the known myopia locus (MYP7) at 11p13 and several other interesting genes. Our highest overall SNP is located in an antisense RNA to BBOX1. Two other highly suggestive SNPs were located in antisense RNA to BDNF which is known to have retinal healing properties. These two SNPs were particularly strong in a one family (LOD score > 1). The most interesting exonic SNPs occurred in PTPRJ, one of which was predicted damaging. While PTPRJ has not been implicated in myopia, other protein tyrosine phosphatases have been. We have identified a genome-wide significant linkage peak on 11p for myopia in Caucasian families. The region contains a large number of quality candidate genes, and we plan targeted sequencing on the region to further narrow down the causal variant(s).

Content Area: Genetic Epidemiology
Keywords: Myopia, linkage, family-based data, exome data

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

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

[1] Department of Biostatistics, Johns Hopkins University 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

Presented by Julius S. Ngwa

Purpose: 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. Our aim is to compare gene/transcript counts obtained from Hierarchical Indexing for Spliced Alignment of Transcript (HISAT2) and Spliced Transcripts Alignment to Reference (STAR) algorithms.

Methods: 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). The data entailed highly parallel genome-wide expression profiling from individual cells in mouse retinal tissue obtained by separating them into nanoliter-sized aqueous droplets. We compared the Digital Gene Expression (DGE) matrix from the aligned library, as well per-cell information which indicates the number of genes and transcripts observed.

Results: 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.

Conclusions: STAR and HISAT2 aligners provide information on the number of reads that map to a particular genomic position, but lack information about 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 within the same gene. Resolving potential fragment assignment ambiguity may be an essential issue to address in RNA-Seq data.

Content Area: Statistical Genetics
Keywords: Drop-Seq, Single-Cell RNA Seq, Differential Expression, HISAT, STAR

Retrospective Lab Database & EMR Analysis to Identify Patients at risk of HPP

Christina Peroutka[1], Julie Hoover Fong[2,3], Adekemi Yewande Alade[2,3], Kerry Schulze[2,3], Mark Marzinke[4] and John McGready[5]

[1] McKusick Nathans Institute of Genetic Medicine, Johns Hopkins Hospital
[2] Greenberg Center for Skeletal Dysplasias, McKusick Nathans Institute of Genetic Medicine
[3] Johns Hopkins University
[4] Clinical Pharmacology Analytical Laboratory, Johns Hopkins Hospital
[5] Bloomberg School of Public Health, Johns Hopkins University

Presented by Christina Peroutka

Introduction: Hypophosphatasia (HPP) is an inherited metabolic condition caused by mutations in the tissue-nonspecific alkaline phosphatase gene (TNSALP/ALPL) with pleiotropic manifestations ranging from perinatal lethal bone fragility, respiratory failure and seizures to early tooth loss and adult-onset musculoskeletal pain and low bone density. Low serum alkaline phosphatase (ALP) is a key biochemical indicator of HPP, though often unrecognized as pathologic. Enzyme replacement therapy (ERT) with asfotase-alpha was recently approved by the FDA for perinatal lethal HPP. Given the spectrum of HPP manifestations and unappreciated significance of low ALP, we hypothesize HPP is underdiagnosed. Furthermore, misdiagnosis of idiopathic osteoporosis and empiric treatment with bisphosphonates is worrisome since this treatment is contraindicated in HPP. Electronic medical records (EMRs) provide a unique opportunity to query large amounts of data to identify abnormal laboratory data or clinical findings which may be consistent with a rare or under-recognized condition such as HPP. This is the first EMR-related risk prevalence study of HPP in a large hospital population.

Methods: With a waiver of consent from the Johns Hopkins Medical Institution (JHMI)Review Board, retrospective ALP results generated by the Johns Hopkins Hospital (JHH)Core Laboratories were collected over 3 years (January,2013 - December,2015) and maintained on a password-protected, institutionally-supported server. Data from all inpatients and clinic patients with at least one serum ALP measurements were included. Patients were excluded from subsequent analysis if demographics were incomplete or the ALP results were cancelled. Based on age- and sex-specific references values for ALP activity, low measurements were categorized as abnormal. Next, all ICD-9 codes associated with each patient were merged with ALP measurements. Patients with at least 1 of 150 ICD-9 codes for specific hematologic, oncologic, drug and transplant causes of low serum ALP were excluded from further consideration. The remaining patients with at least one low ALP measurement and allowable ICD-9 codes were prioritized for EMR review based on the total number and proportion of low ALP values. Data extraction included history and physical features suggestive of HPP.

Results: There were 1,030,121 ALP measurements from 163,016 patients performed at the JHH Laboratories from January,2013 through December,2015. After exclusion of ALP levels with cancelled results or incomplete demographic data, there were 983,458 observations from 156,459 patients remaining for further consideration. After application of age- and sex-specific ALP references values, 45,233 low measurements (22,175 in patients <19 years; 23,058 in patients >19 years)were identified. Exclusion of patients based on ICD-9 codes associated with low ALP levels yielded a cohort of 8,098 patients eligible for chart review.

Conclusions: EMR query is a powerful and novel source of large data across a patient population. After EMR review, we hope to identify a subcohort of patients with a high likelihood of having HPP. We plan to seek new JHMI IRB approval to contact these at-risk patients to offer molecular testing, genetic counseling and ERT. This type of study serves as a proof of principle that EMR-related prevalence studies may be applied more broadly to help improve identification and treatment of patients with rare conditions.

Content Area: Human Genetics
Keywords: Bioinformatics/Methodology, Bone/Joint Abnormalities, Databases, Delineation of Diseases/Phenotype, Risk Assessment

Dynamic, cell type-specific cis-regulatory element use in the developing human frontal cortex

Amanda J. Price[1,2], Joo Heon Shin[1], Ran Tao[1], Thomas M. Hyde[1,3,4], Joel E. Kleinman[1, 4], Andrew E. Jaffe[1,5,6], Daniel R. Weinberger[1,2,3,4,7]

[1] Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, Maryland, USA
[2] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[3] Department of Neurology, Johns Hopkins School of Medicine, Baltimore, Maryland, USA
[4] Department of Psychiatry, Johns Hopkins School of Medicine, Baltimore, Maryland, USA
5 Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
6 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
7 Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Presented by Amanda J. Price

Current evidence suggests that cis-regulatory DNA elements such as enhancers may play an underappreciated role in neurodevelopmental psychiatric disorders such as schizophrenia (SZ); cataloguing dynamic enhancer use over cortical development in different cell types in human post-mortem brain therefore has the potential to illuminate currently unknown mechanisms underlying SZ risk by identifying putative non-coding regulatory sequences critical to normal neuronal and glial development and the timeframes in which they act. To assess cis-regulatory element use over human cortical development, accessible chromatin in two cell types from dorsolateral prefrontal cortex (DLPFC) of donors aged 0.2–22.71 (N=22) was assessed using the Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq). Enriched neuronal and non-neuronal (i.e., glial) populations were first isolated using NeuN antibody labeling and fluorescence-activated sorting of nuclei purified from the homogenate tissue. 1-3 technical replicates were generated from each ATAC-seq library. Peaks called using MACS2 passing an Irreproducible Discovery Rate cutoff of 0.15 were merged and coverage at each peak was calculated. Developmental analyses compared bins spanning ages 0-1 (Neonate), 1-12 (Child), 13-17 (Teen), and 18-23 (Young Adult). We identified over 100,000 regions of accessible chromatin across the genome. 89% of regions fall in intronic or intergenic sequence, although open regions are enriched around TSSs. PCA shows at least 31.2% of the variance in accessibility across samples is defined by cell type, identifying cell type as the largest source of variation in the accessible chromatin landscape. Regions differentially accessible by cell type were enriched near genes with neuronal- and glial-associated biological processes. In contrast, few developmentally regulated accessible regions were detected; interestingly however, most changes in accessibility were between neonates and the other age groups, as would be expected given the large amount of remodeling that occurs from that age. Regions more open in teens than neonates exist nearer to genes involved in RNA splicing, an important mechanism of brain development. Ongoing analysis includes more in-depth focus on loci associated with genetic risk for SZ.

Content Area: Human Genetics, Other
Keywords: accessible chromatin, ATAC-seq, schizophrenia, epigenomics, neurodevelopment

Leveraging genomics to track malaria transmission dynamics in varied epidemiologic settings in Zambia

Julia C. Pringle[1], Giovanna Carpi1, Kelly M. Searle[1], Christine M. Jones[1], Tamaki Kobayashi[1], Ben Katowa[2], Jennifer C. Stevenson[1,2], Philip E. Thuma[2], Douglas E. Norris[1] and William J. Moss[1,2]

[1] Johns Hopkins Bloomberg School of Public Health
[2] Macha Research Trust
[3] Southern Africa International Center of Excellence in Malaria Research

Presented by Julia C. Pringle

The World Health Organization has a goal to eliminate malaria from 35 countries by 2030, with many target countries in sub Saharan Africa. Those attempting to achieve elimination have faced difficulties due to high levels of national and regional heterogeneity in transmission. Investigating the genetic diversity of both Plasmodium falciparum parasites and vector populations across different epidemiologic settings can guide interventions to control and eliminate malaria across heterogeneous transmission intensities. The Southern Africa International Center of Excellence for Malaria Research (ICEMR) has two study sites in Zambia investigating the epidemiology of malaria transmission and evaluating the impact and effectiveness of control and elimination efforts. Nchelenge District in northern Zambia shares a border with the Democratic Republic of Congo (DRC) and has perennial high malaria transmission (holoendemic), with limited impact of control measures. Parasite and vector genomics are being used to understand how transmission is sustained in the face of control interventions. An amplicon deep sequencing approach using P. falciparum human-derived samples collected from Nchelenge District and bordering towns, Kilwa and Kashobwe, in the DRC will enable assessment of cross-border malaria transmission in Nchelenge District. Whole-genome sequencing of vectors from Nchelenge District, DRC, and Tanzania will enable population genetic and demographic analyses to inform the extent to which vector population structure contributes to sustaining high transmission. Whole-genome capture techniques are being used to assess patterns of genomic diversity of P. falciparum sampled from mosquito vectors to understand the role of vectors in driving parasite diversity, assess the genetic relatedness of mosquito-derived strains to those isolated from humans, and monitor transmission changes in response to vector control efforts. Macha, in southern Zambia has low-level, seasonal transmission (hypoendemic) and is approaching elimination. In Macha, parasites have been genotyped to describe residual endemic transmission, investigate the role of imported infections, and identify barriers to achieving elimination. A 24-SNP molecular barcode was used to investigate the relatedness between actively and passively detected infections. Microsatellites are being used to determine the relatedness between infections identified through reactive case detection and construct focal transmission networks. P. falciparum histidine-rich protein II (Pfhrp2) based rapid diagnostic tests (RDTs) are widely used for diagnosis and the deletion of the pfhrp2 gene could be a barrier to case identification. In both sites, investigating the prevalence of Pfhrp2 deletions will enable us to understand the potential for false negative RDT results. These findings can be used to inform programs to more effectively implement interventions and achieve malaria control in high transmission settings and achieve and sustain elimination in low transmission settings.

Content Area: Pathogen Genetics, Other
Keywords: Plasmodium, malaria, hrpII, Zambia

Genetic Principal Component Analysis Using Summary Level Data Identifies New Blood Lipid Loci

Guanghao Qi [1] and Nilanjan Chatterjee[1,2]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public
[2] Department of Oncology, Johns Hopkins School of Medicine

Presented by Guanghao Qi

Atherosclerosis depends heavily on levels of low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides and total cholesterol. The blood lipid levels were shown to be associated with over 150 loci by single-trait analysis of GWAS data. We developed a summary level data based multi-trait approach called the genetic principal component analysis (gPCA). gPCA reduces the number of phenotypes while keeping most of the genetic information, by finding the linear combination of traits that maximizes the heritability. We use LD-score regression methods to estimate variance-covariance matrix of multivariate trait decomposed into genetic and non-genetic factors, and then perform suitable matrix eigenvalue decomposition to find the directions that have maximum variances due to genetic components. We applied gPCA to the joint meta-analysis results of Metabochip and GWAS data from the Global Lipids Genetics Consortium. The association tests for the first and second genetic principal components show increased power. This method discovered 22 new genome-wide significant loci that were not identified by standard methods. The use of summary level data makes the method very easy to implement.

Content Area: Statistical genetics
Keywords: Genetic principal component analysis, summary level data, heritability, blood lipids

Investigating Expression of Epistatic Interaction Networks in Bipolar Disorder in Human Dorsolateral Prefrontal Cortex

Nina Rajpurohit[1], Carrie Wright[1], Richard E. Straub[1], Andrew E. Jaffe[1,2,3], Ran Tao[1], Fernando S. Goes[4], Peter P. Zandi[3], Thomas M. Hyde[1,4,5], Joel E. Kleinman[1,4], Daniel R. Weinberger[1,4,5,6,7]

[1] Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD
[2] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore
[3] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[4] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD
[5] Department of Neurology, Johns Hopkins School of Medicine, MD
[6] The Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, MD
[7] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Presented by Nina Rajpurohit

Bipolar disorder (BD) is a common, highly heritable psychiatric illness whose genetic underpinnings remain unclear. Recent genome-wide association studies (GWAS) have identified multiple risk loci, but the complex genetic architecture of BD makes it necessary to evaluate findings in a broader biological context. Investigation of epistatic interactions between disease-associated genes has the potential to illuminate this area. Such interactions can be elucidated by the recently developed W-test, which identifies pair-wise epistasis effects, and which in a recent publication identified two interaction networks when applied to BD GWAS datasets. To explore whether the identified interaction networks are altered in BD, we characterized their expression using RNA sequencing and gene set analysis of post-mortem brain tissue from patients and controls.

Content Area: Statistical Genetics
Keywords: Bipolar Disorder, epistasis, gene set

Transcriptomic profiling of normal pancreatic duct, precursor, and pancreatic adenocarcinoma organoids

Nicholas J. Roberts[1+], Fei Chen[2+], Miaozhen Qiu[1], Alison P. Klein[1,2,3] and Ralph H. Hruban[1,2]

[1] Department of Pathology, Johns Hopkins Bloomberg School of Medicine, Baltimore, Maryland, USA
[2] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[3] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland, USA
[+] Authors contributed equally

Presented by Nicholas Roberts

The discovery of susceptibility genes for pancreatic ductal adenocarcinoma (PDAC) is hindered by the lack of understanding of the gene functions in normal pancreatic duct. This project aimed to generate a high-resolution transcriptomic profile of normal pancreatic duct through RNA-sequencing, and to identify differentially expressed genes (DEGs) between groups of samples with different phenotypes.

We have established pancreatic duct organoids from 8 patients undergoing pancreatic resection at the Johns Hopkins Hospital. Organoids were derived from pathologically normal duct in 4 patients, intraductal papillary mucinous neoplasms (IPMN, a macroscopic PDAC precursor lesion) in 2 patients, and PDAC in a further 2 patients. Normal pancreatic duct, IPMN, and PDAC samples were digested to single cells before being suspended in Matrigel^{TM} and grown in media containing growth factors. Each organoid line was expanded *in vitro* before DNA and RNA isolation, and cryopreservation.

RNA-seq sample libraries were prepared using polyA selection and were sequenced on an Illumina HiSeq 2500 V4 system. Possible adapter sequences and poor quality reads were removed. Remaining sequence reads were aligned to the hg38 reference genome. After mapping, the total gene hit counts were measured and normalized, and the reads per kilobase of transcript per million mapped reads (PRKM) values were calculate using DESeq2. Comparisons were made between normal pancreatic duct vs. IPMN samples, normal pancreatic duct vs. PDAC samples, and IPMN vs. PDAC samples. Genes with adjusted p-value < 0.05 and absolute log2 fold change > 1 were considered as differentially expressed genes (DEGs).

In summary, these eight samples on generated 38-41 million paired reads per sample, of which over 92% had a quality score > 30 and >96% of sequenced fragments were mapped to the reference genome. In the comparison between normal pancreatic duct vs. IPMN, normal pancreatic duct vs. PDAC, and IPMN vs. PDAC samples, our analysis identified 19, 205 and 35 DEGs, respectively. *PRSS1*, a known susceptibility gene for familial pancreatic cancer was down-regulated (adjusted P = 0.049) in PDAC patients relative to IPMN patients. None of these genes were located in reported GWAS loci. We also found several genes were differentially expressed across comparisons. For example, gene *PITX1* had a 1.60 fold increased expression among IPMN patients in comparison to normal individuals (adjusted $P = 5.46 \times 10^{-5}$), and was up-regulated by 1.73 fold among PDAC patients in comparison to IPMN patients (adjusted $P = 8.40 \times 10^{-5}$). Gene *NPC1L1* and *SIM1* suggested a dose-response in their expression intensity in comparison between IPMN vs. PDAC samples, and normal pancreatic duct vs. PDAC samples.

Derivation and RNA-sequencing of additional pancreatic duct samples and further analysis is required to determine the function of these genes. Results from this project, in combination with our future analysis of whole genome sequence data, will facilitate the susceptibility gene discovery in pancreatic cancer.

Neuronal differentiation potential of HT22 cells with Kabuki Syndrome mutations in vitro

Johanna Robertson[1] and Loyal Goff[1]

[1] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Presented by Johanna Robertson

Kabuki Syndrome is a genetic disorder caused by a haploinsufficiency of one of two genes which promote the opening of chromatin: KMT2D and KDM6A. These genes encode a histone methyltransferase and a histone demethylase respectively, both of which are required to maintain the proper balance of open vs closed chromatin and ultimately the regulation of gene expression at target loci. Kabuki syndrome is characterized by specific facial characteristics, decreased immune function, growth retardation, and intellectual disability. Patients with Kabuki syndrome demonstrate significant deficits in learning and memory, which may be attributed to reduced neurogenesis in the adult hippocampus. Indeed, recent studies in a mouse model of Kabuki syndrome have identified a significant reduction in neurogenic divisions relative to wild type. To identify the specific target genes of one KS-associated gene, KMT2D, and how these target genes are disrupted in hippocampal neurogenesis in KS patients, we used CRISPR-Cas9 to introduce a KMT2D loss of function mutation into an immortalized mouse hippocampal cell line, HT22. WT and mutant cells were differentiated towards mature neurons, and cells were collected at days 0, 1, 3, and 5 in triplicate. Total RNA from collected cells and used to create RNA-Seq libraries for differential expression analysis. Consistent with KMT2D's global role in maintaining open chromatin, we identified a global decrease in gene expression in the KMT2D-editied cell line. Specific differential gene expression analysis between mutant and WT lines is ongoing.

Content Area: Human Genetics
Keywords: Kabuki Syndrome, Neuronal differentiation, Adult neurogenesis, KMT2D

Synthetic Long Read Sequencing with Optical Mapping To Produce a High Quality de novo Genome

Alan Scott[1] and David Mohr[1]

[1] Johns Hopkins Genetic Resources Core Facility

We have been studying methods for assembling high quality genomes from non-human mammals. Current short-read methods have come to dominate genome sequencing because they are cost-effective, rapid and accurate. However, short reads are most applicable when data can be aligned to a known reference rather than be assembled with more traditional methods. Because standard methods are time-consuming, costly and inefficient, we have explored new approaches to de novo genome assembly. In particular, we evaluated 10X Genomics Chromium Linked-Read sequencing, with ~1M molecular indices, combined with BioNano Genomics (BNG) optical mapping and hybrid assembly to the genome of the endangered Hawaiian monk seal. We show that the Chromium data, assembled with Supernova v1.1 software, provided long sequence blocks and, when used for de novo assembly, produced scaffolds with an N50 of 22.23 Mb with the longest individual scaffold at 84.06 Mb. When combined with BNG optical maps the scaffold N50 increased to 29.65 Mb and the longest individual scaffold increased to 84.78 Mb. Because both BNG and 10X technologies interrogate single DNA molecules, they can also be used to construct haplotypes and to detect larger scale structural variants between parental chromosomes. Combining the two orthogonal methods of Chromium Linked-reads with BNG optical maps is likely to make the assembly of high quality genomes routine and significantly improve our understanding of comparative genome biology.

Content Area: Other
Keywords: Long DNA sequencing, optical maps, de novo genome, comparative genomics

Genome specific transcriptional signatures predict differentiation biases in Human ES/IPS cells

Genevieve Stein-O'Brien[1,2], Amritha Jaishankar[1], Suel-Kee Kim[1], Seungmae Seo[1], Joo Heon Shin[1], Daniel Hoeppner[1], Josh Chenoweth[1], Thomas Hyde[3,4], Joel Kleinman[1,3], Daniel Weinberger[1,3,4,5], Ronald McKay[1], Elana J Fertig[6] and Carlo Colantuoni[5]

[1] Lieber Institute for Brain Development, Baltimore, MD
[2] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD
[3] Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD
[4] Department of Psychiatry, Johns Hopkins School of Medicine, Baltimore, MD
[5] Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD

Presented by Genevieve Stein-O'Brien

Starting from a single fertilized egg, the compendium of human cells is generated via stochastic epigenetic changes. The resulting changes to global gene expression drives variation in cellular phenotypes while canalization of developmental pathways regulate the nature of this variation to ensure viability of the individual as a whole. Thus, it is unsurprising that major deviations in global gene expression early in life have been linked to complex human diseases. Human pluripotent stem cells (hPSCs) are a highly robust and uniquely human experimental system in which to model the sources and consequences of this variability in gene expression. Further, variation in hPSCs' transcriptomes has been directly linked to both genomic background and biases in differentiation efficiency. Taking advantage of this link between genomic background and developmental phenotypes, we developed Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS), the first robust whole genome Bayesian non-negative matrix factorization (NMF), to find conserved transcriptional signatures representative of the functional effect of human genetic variation. Using time course RNA-seq data obtained from three human embryonic stem cells (ESC) and three human induced pluripotent stem cells (IPSC) in three different experimental conditions, GWCoGAPS distinguished shared developmental trajectories (called dynamic signatures) from unique transcriptional signatures of each of the cell lines (called identity signatures). Enrichment in gene targets of developmental pathway in identity signatures was predictive of lineage biases during neuronal differentiation. Additionally, lineage biases were consistent with early differences in morphogenetic phenotypes within monolayer culture, thus, linking transcriptional genomic signatures to stable quantifiable cellular features. To test whether the cell line signatures were genome specific, we next developed the projectoR algorithm to assess a given signatures robustness in independent data sets. By using the identity signatures as inputs to projectoR, we were able to identify samples from the same donor genome in datasets from multiple tissues and across technical platforms, including RNA-seq data from post-mortem brain and microarray data from embryoid bodies. Further interrogation with projectoR using DNA methylation data showed that genes with high rankings in identity signatures were significantly hypomethylated when compared to other individual's genomes. Therefore, we conclude that the gene expression patterns found with GWCoGAPS are robust and reflect epigenetic regulation of gene expression for individual biases in development.

Content Area: Molecular Genetics, Computational Genetics, Statistical Genetics
Keywords: Human Pluripotent Stem Cells, Computational Methods, human development, post mortem brain, whole genome RNAseq

Perspectives on Genetic Testing and Return of Results from the First Cohort of Presymptomatically Tested Individuals At-Risk for HD

K.M. Stuttgen[1,2], R.L. Dvoskin[1], J.M. Bollinger[1], A. McCague[1,2], B. Shpritz[3] and D.J.H. Mathews[1]

[1] Berman Institute of Bioethics
[2] Institute of Genetic Medicine
[3] Department of Psychiatry and Behavioral Science

Presented by K.M. Stuttgen

Technological advances have made genetic and genomic testing possible for a variety of Mendelian diseases, including adult-onset diseases1-4, warranting examination of how we think about genetic testing and its benefits and risks. For decades, Huntington's disease (HD) has served as a model for how we think about genetic testing, and its benefits and harms for at-risk individuals and their families. In 1983, HD was the first genetic disease mapped using DNA polymorphisms. Shortly thereafter, presymptomatic genetic testing for HD began in the context of two clinical trials. The Baltimore Huntington's Disease Project (BHDP) at Johns Hopkins University began in 1986 and enrolled 180 individuals. Experiences with this cohort influenced collective thinking about issues related to genetic testing.

The current study is obtaining opinions on and attitudes toward genetic testing from people who enrolled in the BHDP 20-30 years ago. One-hour semi-structured interviews are being conducted with 20 people found to carry an expanded HD repeat, 20 with normal repeats, and 10 who dropped out of the BHDP before disclosure of test results. As part of the interview, participants are asked their opinions on the importance of autonomy in the decision to be tested, whether a formal testing protocol is necessary, whether online direct-to-consumer genetic testing for HD would ever be acceptable, and whether incidental findings (including the presence of HD or other genetic risk factors) should be returned in the context of whole exome/genome sequencing.

As increasing numbers of genetic tests are being used to both directly and incidentally assay genes for adult-onset neurodegenerative disease, and as large-scale genetic testing is increasingly integrated into clinical care, it is critical that we understand the implications of presymptomatic testing, not just in the near term, but over the course of the lives of at-risk individuals and their families. Because of the relatively recent introduction of presymptomatic genetic testing into clinical care, few such studies exist.

Results to date suggest that participants believe that the decision to undergo testing should be autonomous, though a small number believe everyone at risk for HD should know his/her genetic status. Additionally, most participants believe that a formal protocol is necessary when undergoing genetic testing for HD. A small number felt that the protocol was helpful but that ultimately a patient has the right to know his/her genetic information without having to go through a formal protocol.

Thus far, most participants have stated that DTC genetic testing related to serious disease is unacceptable. The majority of participants believe that those undergoing testing should be asked before testing whether s/he wants to be informed of any incidental or secondary findings. A small number of individuals believe patients should be told of all incidental findings so that they can 1) prepare themselves for the onset of a serious disease, such as HD, and 2) make lifestyle changes in an effort to prevent or mitigate their risk for diseases such as diabetes.

Content Area: Others
Keywords: ELSI, Genetic Testing, Genetic Policy, Huntington's Disease

Genetic determinants of peanut-specific immunoglobulins in the Learning Early About Peanut Allergy (LEAP) study

Alexandra Winters[1], Meher Boorgula[2], Claire Malley[1] and Rasika Mathias[1] for The LEAP study

[1] Johns Hopkins University
[2] University of Colorado

Presented by Alexandra Winters

Rationale: The LEAP study showed a protective effect of early peanut exposure for infants at high risk of developing peanut allergy. Peanut-specific IgG4:IgE ratio increased over time in peanut exposure as compared to avoidance subjects, and was lower in individuals who developed peanut allergy compared to those who remained tolerant. In this study we undertake whole genome sequencing (WGS) to understand the genetic determinants of clinical outcomes in the LEAP Study.

Methods: First, we leverage genomewide genotype array (GWA) data generated as part of our WGS and perform a GWAS on 378 European ancestry participants in the LEAP study. Tests for association were performed on all single nucleotide variants (SNPs) that passed quality control and age, sex, and the first three principal components for ancestry for included in the model.

Results: The peak association signal for peanut-specific IgG4:IgE was located on chromosome 3 mapping to the CD80 locus: rs12330581 and rs3915039 with p= 2.15E-07 and 8.40E-07, respectively. Both the SNPs are common intronic variants with minor allele frequencies >10% in European populations.

Conclusions: In this late breaking abstract we present early results from a genomewide interrogation of clinical outcomes in the LEAP Study. Our results are promising implicating a key candidate gene, CD80 that works to prime T cells as a determinant of baseline peanut-specific IgG4:IgE. We are currently analyzing all sequence-identified variants to follow up on this early result, identify addition rare and novel variants, as well as extending our analysis to other ancestries represented within the LEAP Study.

Content Area: Human Genetics, Genetic Epidemiology
Keywords: peanut allergy, gwas, immunoglobulins, CD80

Evaluation of isomiRNA expression in schizophrenia using small RNA sequencing of postmortem dorsolateral prefrontal cortex brain tissue

Carrie Wright[1,2], Joo Heon Shin[1], Anandita Rajpurohit[1], Courtney Williams[1], Andrew E. Jaffe[1,3,4], Nicholas J. Brandon[5], Thomas M. Hyde[1,6,7], Joel E. Kleinman[1,6], Alan J. Cross[5] and Daniel R. Weinberger[1,6,7,8,9]

[1] Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA
[2] AstraZeneca Postdoc Program, Innovative Medicines and Early Development Biotech Unit, 141 Portland St, Cambridge, MA 01239 USA
[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[4] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
[5] AstraZeneca Neuroscience, Innovative Medicines and Early Development Biotech Unit, 141 Portland St, Cambridge, MA 01239 USA
[6] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, USA
[7] Department of Neurology, Johns Hopkins School of Medicine, MD, USA
[8] The Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, MD, USA
[9] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Presented by Carrie Wright

MicroRNAs are small regulatory RNAs that individually modulate the expression of many genes. miRNAs participate in the regulation of nearly every biological process and are highly implicated in a wide variety of disease. IsomiRNA (IsomiRs) are miRNAs with slight variation in length and sequence from that of the canonically described microRNAs, due to altered miRNA biogenesis, genetic variants, posttranscriptional editing, and exonuclease degradation. Such changes in the sequence of miRNAs, can lead to drastic functional alterations by shifting the repertoire of binding partners, therefore leading to potentially significant downstream consequences. These slightly altered miRNAs are now known to be functional and associated with various disease states, however characterization remains limited. Alterations in miRNA expression and the expression and function of miRNA biogenesis related enzymes has previously been identified in schizophrenia. To explore whether the magnitude or diversity of isomiR production is altered in schizophrenia, we evaluated miRNA and isomiR expression in postmortem brain tissue of cases and healthy controls. This is to our knowledge, the first characterization of isomiR expression in schizophrenia.

We evaluated RNA sequencing data of post-mortem dorsolateral prefrontal cortex (DLPFC) samples from 92 subjects (30 subjects with schizophrenia and 62 control subjects) using the BIOO Scientific NextFlex small RNA sequencing library preparation, with 500ng of starting total RNA. Fifty base pair single-end sequencing was run on the HiSeq 3000, using the Illumina Real Time Analysis (RTA) module to perform image analysis and base calling, and the BCL Converter (CASAVA v1.8.2) to generate the sequence reads. Sequencing depth was over 20 million reads per sample. Reads were aligned to known miRNAs and isomiRs using miRge. The influence of diagnosis status on miRNA and isomiR abundance was then evaluated using a linear model regression covarying for principal components to capture latent sources of variation. Abundance estimates were normalized using Reads per Million mapped reads (to all miRNA sequences) (RPM) and filtered to include only those with expression values greater than 10 RPM. Multiple testing correction was performed using the Benjamini Hochberg method.

The expression of individual miRNAs and isomiRs, as well as the overall magnitude and diversity of isomiR expression was then evaluated for group differences between cases and controls.
Following multiple testing correction, 7 miRNAs were found to be significantly differentially expressed between cases and controls. Additionally, 56 individual isomiRs were found to be differentially expressed. No differences were found when evaluating the magnitude or diversity of overall isomiR expression. Several isomiRs were found to be higher expressed than their respective canonical sequences, however no significant difference in the ratio between the expression of the top expressed isomiRs and respective canonical sequences was identified between cases and controls.

This preliminary study further suggests that isomiR and miRNA expression are altered in schizophrenia. Despite genetic associations of miRNA biogenesis enzymes, global alterations in isomiR expression were not identified. We plan to further analyze isomiR expression with additional samples for replication purposes. Our results will further clarify the role of miRNA and isomiR expression in schizophrenia.

Content Area: Computational Genetics
Keywords: miRNA, schizophrenia, isomiRNA, RNA sequencing, gene expression

Evidence of APOBEC3 editing in the HPV16 genome

Yanzi Xiao[1], Bin Zhu[1], Meredith Yeager[1,2], Michael Cullen[1,2], Joseph F. Boland[1,2], Nicolas Wentzensen[1], Tina Raine-Bennett[3], Zigui Chen[4], Kai Yu[1], Qi Yang[1,2], Mia Steinberg[1,2], David Roberson[1,2], Sara Bass[1,2], Laurie Burdette[1,2], Thomas Lorey[5], Philip E. Castle[6], Robert D. Burk[6,7], Mark Schiffman[1] and Lisa Mirabello[1]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD
[2] Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick, MD
[3] Women's Health Research Institute, Division of Research, Kaiser Permanente Northern California, Oakland CA
[4] Department of Microbiology, The Chinese University of Hong Kong, Hong Kong
[5] Regional Laboratory, Kaiser Permanente Northern California, Oakland CA
[6] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY
[7] Departments of Pediatrics, Microbiology and Immunology, and Obstetrics & Gynecology and Women's Health, Albert Einstein College of Medicine, Bronx, NY

Presented by Yanzi Xiao

Human papillomavirus (HPV) is a very common sexually transmitted infection, however only a small proportion of women progress to cervical precancer or cancer. HPV16 is the most carcinogenic type, causing more than half of the cervical cancer globally. The HPV16 genome is 7,906base-pairs coding for 8 genes (E6, E7, E1, E2, E4, E5, L2, L1) and one upstream regulatory region (URR). Within HPV16 there are 4 main lineages (A, B, C, D) that are strongly associated with disease risk.

Human APOBEC3A (hA3A) cytidine deaminases have been shown to have antiviral effects. The APOBEC mutational process results in a C to T base change at specific motifs (5' [C/T]•C>T•W 3'). Previous studies established that there was evidence of APOBEC3 editing on a small number of HPV16 samples. It's unknown how these mutations are related to infection clearance or the long-term accumulation of genomic mutations that contribute to HPV-associated cancers.

We conducted detailed analyses to comprehensively evaluate APOBEC3 editing on HPV16 genomes using HPV16 whole-genome sequencing data from 3,215 HPV16-infected women in the NCI-HPV Persistence and Progression (PaP) cohort. Cases were defined as women with cervical precancer (CIN3, N = 1,093) or cancer (N = 109) and controls were women with no histologic evidence of precancer or cancer (<CIN2, N = 1,107). HPV16 DNA was extracted from banked specimens and whole-genome sequenced using a high-throughput assay. We evaluated all rare variants, defined as having a minor allele frequency of <1%, for matching an APOBEC3-associated variant. An APOBEC3-associated variant was defined as a variant having one of the eight possible motifs (5' [C/T]•C>T•W 3') out of 96 potential motifs of 3 base-pairs. Using logistic regression, we compared the number of APOBEC3-associated variants in cases and controls, among HPV16 variant lineages, and among genome regions of the virus.

We discovered that there is evidence of APOBEC3 editing throughout the HPV16 genome. Specifically, we observed that women with precancer or cancer had less APOBEC3-associated variants compared to the controls (OR = 0.84, p-value = 0.06). We further showed that women with an HPV16 A lineage infection have more APOBEC3-associated variants compared to those with a non-A lineage infection (OR = 1.35, p-value = 0.02). After controlling for the number of APOBEC3 vulnerable loci, we observed that the L1 (OR = 0.23, p-value = 0.04) and E7 (OR = 0.29, p-value = 0.07) genes have less APOBEC3 footprints overall, and particularly in the cases compared to the controls in these regions, compared to the viral non-coding upstream regulatory region (URR).

Overall, we found that APOBEC3 is not affecting the HPV16 genome in a uniform way, and instead, it appears to be targeting specific regions which could suggest antiviral activity. Importantly, we determined that APOBEC3-associated variants are less prevalent in cases which could be related to disease progression in these individuals. Further evaluation is underway.

Content Area: Genetic Epidemiology
Keywords: Human papillomavirus (HPV), cervical cancer, APOBEC

Understanding the genetic basis of very early onset inflammatory bowel disease (VEOIBD) by using whole exome sequencing

Jing You[1,2], Nara Sobreira[2,3], David Valle[2,3] and Anthony Guerrerio[4]

[1] Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[2] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[3] Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[4] Pediatric Gastroenterology and Nutrition, Johns Hopkins Children's Center, Baltimore, MD 21205, USA

Presented by Jing You

Inflammatory bowel diseases (IBDs, OMIM 266600) are chronic, relapsing intestinal inflammatory disorders of heterogeneous etiology. Like most complex traits, IBD results from a combination of genetic factors and environmental contributors. VEOIBD (OMIM 612567) is defined by an age of onset < 6 years, a severe course and frequently a positive family history. Together, these characteristics suggest a strong genetic contribution.

As part of the Baylor Hopkins Center for Mendelian Genomics (BHCMG), we performed WES on 18 probands with VEOIBD aiming to uncover new disease causing variants and genes. Five probands are from multiplex families and 13 probands are isolated cases. The 5 multiplex families fit an autosomal dominant mode of inheritance with at least 2 generations affected.

To analyze the WES we used the PhenoDB variant analysis tool (Sobreira et al., 2015) to select, in each proband, the heterozygous and homozygous rare (MAF<1%) functional variants (missense, nonsense, splicing, indels). Next, we applied four different analysis strategies. 1. We analyzed the genes known to cause IBD and genes known to cause monogenic disorders with VEOIBD as part of their phenotype. 2. We selected genes that were mutated in the homozygous or compound heterozygous state in 2 or more probands. 3. We selected genes mutated (heterozygous variants) in 4 or more families. 4. We selected variants in genes previously associated to IBD by GWAS.

The analyses described above have identified candidate genes that are now being further investigated. Variants in NOD2, PLCG2, GUCY2C, and WAS (genes known to cause IBD or disorders characterized by VEOIBD) have been identified in 8 probands. The segregation analysis of the variants in these genes showed incomplete penetrance associated to NOD2, PLCG2, and GUCY2C. The recessive analysis identified only one gene, the CNGA2 on chromosome X, as a novel candidate gene in two male probands, each with different variants, p.G113V (MAF=0.6%) and p.R453H (MAF=0.01%). Eleven genes has heterozygous rare functional variants in 4 or more probands and their relevance to the VEOIBD phenotype needs to be further investigated. Eight genes were selected because of their previous association with IBD by GWAS. These genes were also mutated in 3 or more probands.

Currently we are working on two additional analysis strategies: a polygenic analysis, in which we are investigating 2 or more genes mutated in two or more probands; and a pathway analysis in which we are investigating different mutated genes involved in the same pathways.

Content Area: Human Genetics
Keywords: Inflammatory bowel diseases (IBDs), very early onset, inflammatory bowel disease (VEOIBD), whole exome sequencing (WES)

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

Haoyu Zhang[1], Thomas U.Ahearn[2], Montserrat García-Closas[2] and Nilanjan Chatterjee[1,3]

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

Presented by Haoyu Zhang

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 association in GWAS of cancers utilizing multivariate tumor characteristics. The framework can be used to test for overall genetic association and evidence of etiologic heterogeneity, overall or by specific 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 (Ncase=96317, Ncontrol=111357) of breast cancer incorporating ER, PR and HER2 status, three clinically relevant tumor characteristics.

Content Area: Statistical Genetics
Keywords: GWAS, Heterogeneity, Two-Stage Modelling

Estimating effect-size distribution from summary-level statistics for large genome-wide association studies

Yan Zhang[1] and Nilanjan Chatterjee[1,2]

[1] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University
[2] Department of Oncology, School of Medicine, Johns Hopkins University

Presented by Yan Zhang

It has been shown that widely available GWAS summary-level statistics, i.e., estimates of SNP effects and standard error from one SNP at a time analysis, can be used to estimate heritability and co-heritability of traits that can be explained by common variants. We describe a novel likelihood based approach for analyzing summary-level statistics to estimate the effect-size distribution of underlying causal SNPs incorporating linkage disequilibrium (LD) scores for the SNPs which are available from public databases. We show that, under certain assumptions, a mixture normal distribution for the underlying effect of the SNPs in a joint model leads to another mixture normal distribution for the marginal effects of the SNPs, the summary-level statistics available from GWAS. In the resulting mixture model, the mixing proportions are determined by the probability distribution of number of underlying causal variants a SNP may tag due to linkage disequilibrium and the variance component parameters account for the LD-score of the SNPs. Thus, the parameters of the underlying effect-size can be conveniently estimated by an EM where M-step have closed form solutions. We will present results from applications of the methods using summary-level results from large GWAS across several traits. Preliminary analysis of summary level data from GWAS of height and schizophrenia shows that there are about 0.17% and 0.37% underlying common susceptibility SNPs among the 1000 Genome SNPs, explaining an average per-SNP variances of 0.60 (linear scale) and 3.32 (logistic scale), respectively, for these two traits complex traits.

Content Area: Statistical Genetics
Keywords: heritability, likelihood, linkage disequilibrium, summary-level statistics, GWAS