Published on: 28 September 2016

DOI: 10.1002/gepi.22001

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

ABSTRACTS

The 2016 Annual Meeting of the International Genetic Epidemiology Society

INVITED SPEAKERS

1 | Detecting and Correcting for Sample Contamination in DNA and RNA Sequencing Studies

Michael Boehnke[1]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America

Sample contamination is a frequent problem in genome sequencing studies, and may result in genotyping errors and reduced power or increased false-positive rates in downstream analyses. In this talk, I will describe mixture-model-based methods to identify within-species sample contamination using DNA sequencing read data and/or array-based genotype data, and a method to model contamination during genotype calling as an alternative to discarding contaminated samples. I will demonstrate that the contamination detection methods can be used effectively in RNA sequencing studies. This is joint work with Matthew Flickinger, Goo Jun, Hyun Min Kang, and Goncalo Abecasis.

2 | At Last, One Size does not Fit All: Progress Meets Practicality in Translating Genetics into New Medicines

Lon Cardon[1]

[1]GlaxoSmithKline, King of Prussia, Pennsylvania, United States of America

The substantial progress over the past decade in identifying genetic loci underlying both rare and common human diseases comes at an opportune time for drug discovery, as the failure rates due to lack of efficacy are increasing and occurring at the worst possible time, in late-stage clinical trials when the maximum costs and time have been spent. Much of the pharmaceutical industry is now returning to genomics in the hope to translate the genetic discoveries to new drug targets and thereby reduce the failure rates. Further enthusiasm for translation is coming from the advances in sequencing, genome editing, imaging and data analytics, all in the hope of moving from association/mutation discovery to better functional understanding for drug discovery. To date, however, most of the genetic findings have not led to novel medicines, apart from a few well-cited examples. Similarly, studies of genetic predictors of drug response have revealed relatively few major insights on responders vs. non-responders, apart from predictors of adverse drug response and somatic mutations in oncology. I will review both the promise and recent challenges of genetics in drug discovery and discuss some of the exciting opportunities presently emerging, particularly in the context of targeted therapies based on data and methods from genetic epidemiology.

3 | "Moving the Genome to the Clinic"

Gail Jarvik[1]

[1]Division of Medical Genetics, University of Washington Medical Center, Seattle, Washington, United States of America

There are numerous obstacles to genomic medicine. These include developing best practices for when and how to use genomic tests, recognizing patient and provider preferences, educating primary care providers, finding cost-effective delivery options, legal and regulatory issues, and overcoming barriers to insurance coverage.

A major challenge is the large number of rare and novel genomic variants per individual. The American College of Medical Genetics and Genomics (ACMG) has recommended that all pathogenic variants in 56 gene-disease pairs that are identified incidentally in a genomic test be offered to the patient (Green et al., 2013, PMID:23788249). We considered an expanded list of 112 actionable gene-disease pairs, ones where medical intervention is possible to prevent or detect disease early. We estimate the rate of these Incidental Findings (IFs) in European and African Ancestry groups. However, we found high discordance between classifications of expert reviewers. We have reported both inconsistency across labs in variant classification and a bias towards overcalling pathogenicity (Amendola et al., 2015, PMID:25637381).

Thus, there is a need to standardize classification of genomic variants in medical sequencing. In the past genomics laboratories have used non-standard classification systems. The ACMG published guidelines for variant classification for

Abstract #93 (Is there a relationship between fetal sex and placenta cord characteristics in twin gestations by Shayesteh Jahanfar & Kenneth Lim) has been removed on 14th october after original online publication.

Mendelian disorders designed to increase consistency among labs (Richards et al., 2015, PMID:25741868). The Clinical Sequencing Exploratory Research (CSER) Consortium evaluated the use of these rules by nine of the CLIA laboratories supporting CSER projects, considering 99 germline variants. The results were examined to evaluate intra-laboratory differences between variant classifications using the labs own criteria vs. adopting ACMG criteria and inter-laboratory differences using either the lab's own system or the ACMG guidelines. Agreement among labs did not differ whether using the laboratory specific vs. ACMG criteria (*p*=0.9); i.e. the ACMG criteria did not yield more consistent variant classification in this exercise. We further analyzed sources of disagreement in the use of the ACMG criteria and identified causes of variance in classifications. In addition to providing useful analyses of how variant classifications approaches vary among laboratories, these data should allow clarification and refinement of the ACMG criteria that may increase consistency in variant classification (Amendola et al., 2016, PMID:27181684).

5 | Exposomics: Lifestyle, Chemical, Physical and Social Exposures

Joel Schwartz[1]

[1]Departments of Environmental Health and Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

The Exposome is the concept that people are exposed to many external stresses and that their response depends on timing, co-exposures, and cumulative exposures in complex ways. To convert this from an idea to research is challenging. Even if all relationships were linear and there were no interactions, there is the challenge of identifying timing of exposure and critical windows for multiple correlated exposures. If multiple exposures influence the same pathway, this may appear as nonlinearity in the 'last exposure', or a single interaction, whereas the truth is more complex. Finally, the exposome interacts with the genome and epigenome and may be mediated by epigenomic changes. I will discuss examples of critical windows, interactions, mediation analysis, pathway analyses, and ways of dealing with cumulative exposures.

6 | Discovery of Genetic Variants for Cardiometabolic Disease: Lessons from Population Sequencing

Nicole Soranzo[1,2]

[1]Wellcome Trust Sanger Institute, Hinxton, United Kingdom; [2]Department of Haematology, University of Cambridge, Cambridge, United Kingdom

Whole-genome and whole-exome sequencing studies are increasingly used to study the contribution of low-frequency (MAF 1–5%) and rare (MAF<1%) variants to complex traits and diseases, and contribute to a greater understanding of differences in genetic architecture across different traits and diseases. Starting from our experience with the UK10K study, I will discuss how genome sequencing and imputation based on genome-sequence empowers discovery of low-frequency and rare variants and boosts interpretation of findings from genome-wide association studies.

ORAL PRESENTATIONS

7 | Assessing the Genetic Effect Mediated through Gene Expression from Summary eQTL and GWAS Data

Richard Barfield[1,2], Bogdan Pasaniuc[3,4], Peter Kraft[1,2]

[1]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [3]Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America; [4]Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California, United States of America

Integrating GWAS and eQTL data can boost power to detect novel disease loci or pinpoint the susceptibility gene at a known disease locus. However, it is often the case that multiple eQTL genes co-localize at disease loci (an effect of linkage disequilibrium, LD), making the identification of the true susceptibility gene challenging. To distinguish between true susceptibility genes (i.e. when the genetic effect on phenotype is mediated through expression) and spurious co-localizations, we developed LD-aware MR Egger regression (LD-MRE). LD-MRE is an extension of MR Egger regression (MRE) to the case where multiple SNPs in LD are associated with gene expression. LD-MRE only requires summary GWAS and eQTL, along with LD from reference panels. Through simulations we show that when eQTLs have direct (pleiotropic) effects on disease, LD-MRE provides adequate control of Type I error, more power, and less bias than previously-proposed methods. When there is no effect of gene expression on disease, LD aware MR-Egger regression has the desired Type I Error, while LD aware MR (LD-MR), which assumes no pleiotropy, can have a 13.6× inflated Type I Error. When there was no pleiotropy, LD-MRE had power nearly equal to that of LD-MR. In the presence of pleiotropy, LD-MR is not a valid test, while LD-MRE had up to 3× greater power than MRE while properly controlling Type I Error. Our method can also be used for general Mendelian Randomization analyses when SNPs in the genetic instrument are correlated.

8 | Using Data-Driven Approaches to Address Clinical Heterogeneity in Complex Traits using COPD

Anna O. Basile[1], H. Lester Kirchner[2], Joseph B. Leader[2], Catarina B. Manney[2], Anurag Verma[2], Marylyn D. Ritchie[1,2]

[1] *Department of Biochemistry and Molecular Biology, Center for Systems Genomics, The Pennsylvania State University, University Park, Pennsylvania, United States of America;* [2] *Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, Pennsylvania, United States of America*

Complex diseases are heterogeneous in nature; instead of representing a single disease, these conditions may be comprised of multiple disorders, each with varying symptoms, clinical presentations, and differing etiology. Trait heterogeneity is a confounding variable often overlooked in traditional genetic approaches. It complicates patient diagnoses, and can decrease statistical power to detect associations, and risk attributed to susceptibility variants. The copious amounts of biomedical data in the Electronic Health Record (EHR) system can be leveraged using data-driven approaches to overcome heterogeneity and detect homogeneous patient subgroups. We performed *in silico* evaluations to examine the performance of multiple machine learning algorithms in homogeneous data classification. The evaluated algorithms included hierarchical clustering, k-means analysis, random forests, principal component analysis and t-distributed stochastic neighbor embedding. Various proximity and similarity metrics were evaluated to account for mixed data types within the EHR, e.g., binary, continuous, and missing variables. We demonstrate the utility of these algorithms by using Chronic Obstructive Pulmonary Disease (COPD), a condition with known heterogeneity. 49,000 patients from the Geisinger Health System MyCode® Community Health Initiative with known COPD diagnosis were subset using clinical lab variables, comorbidities, clinical diagnosis codes, metabolic panels, medication usage, vitals, spirometry and exercise testing measures. Using COPD as a proof of concept model, we demonstrate that data-driven approaches can be used to overcome clinical heterogeneity by embracing the complexity and wealth of EHR data. These methods have the potential to build more accurate predictive disease models that can aid in uncovering the genetic basis of complex traits.

9 | A Model for Interpretable High Dimensional Interactions

Sahir R. Bhatnagar[1,2], Yi Yang[3], Mathieu Blanchette[4], Celia M.T. Greenwood[1,2]

[1] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada;* [2] *Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada;* [3] *Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada;* [4] *Department of Computer Science, McGill University, Montréal, Canada*

There are several applications where interaction models can reflect biological phenomena and improve statistical power. Furthermore, diseases are now thought to be the result of entire biological networks whose states are affected by environmental factors. These systemic changes can induce or eliminate strong correlations between elements in a network without necessarily affecting their mean levels. We propose a multivariate penalization procedure for detecting interactions between high dimensional data ($p \gg n$) and an environmental factor, where the effect of this environmental factor on the high dimensional data is widespread and plays a role in predicting the response. Our approach improves on existing procedures for detecting such interactions in several ways; 1) it simultaneously performs model selection and estimation 2) it automatically enforces the strong heredity property, i.e., an interaction term can only be included in the model if the corresponding main effects are in the model and 3) it reduces the dimensionality of the problem and leverages the high correlations by transforming the input feature space using network connectivity measures. An extensive simulation study shows that our method outperforms LASSO, Elastic Net and Group LASSO in terms of both prediction accuracy and feature selection. We apply our methods to the NIH pediatric brain development study to refine estimates of which regions of the frontal cortex are associated with intelligence scores, and a sample of mother-child pairs from a prospective birth cohort to identify epigenetic marks observed at birth that help predict childhood obesity. Our method is implemented in an R package: http://sahirbhatnagar.com/eclust/.

10 | Investigating Fine-Scale Population Structure in the United Kingdom BioBank

James P. Cook[1], Andrew P. Morris[1]

[1] *Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom*

The United Kingdom (UK) is a genetically diverse population with strong genetic differences between regions, which may adversely affect Genome-wide Analysis Study (GWAS) of complex traits if not fully accounted for in the analysis. Principal components, calculated from a genetic relatedness matrix, are routinely included in regression models to account for population structure in GWAS.

The UK BioBank provides an opportunity to examine fine-scale UK population structure in unprecedented detail, with a first release of genetic data consisting of ∼150,000 genotyped individuals recruited from 23 centres. Easting and Northing coordinates were collected for every participant at recruitment and birth, and genotypes were imputed up to a combined 1000 Genomes and UK10K reference panel.

We have performed GWAS using the Easting and Northing coordinates at birth as continuous phenotypes, in univariate analyses, with and without adjustment for principal components. In analyses adjusted for six principal components,

variants mapping in *TLR1* showed the strongest signal of association genome-wide with both Northings (rs4833095, previously associated with asthma and hay fever, p=3.5×10^{-122}) and Eastings (rs5743614, previously associated with alcohol tolerance, p=3.5×10^{-13}), representing a North-West to South-East cline in allele frequencies across the UK. However, after adjusting for 15 principal components, the strongest *TLR1* association with Northings was reduced to p=1.5×10^{-18}, while no genome-wide significant associations (p<5×10^{-8}) with Eastings remained.

Our study has important implications for accounting for population structure in large scale GWAS performed in the UK BioBank, and highlights the need for caution in interpreting association results from regression models incorporating principal components.

11 | Simplified Power Calculations for Rare Variant Association Tests and Implications for Association Studies

Andriy Derkach[1], Nilanjan Chatterjee[2]

[1]National Cancer Institute, Rockville, Maryland, United States of America; [2]Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America

Many statistical tests have been proposed for association studies with rare genetic variants. These methods can be divided into two classes: tests based on a linear composite statistic (e.g., burden tests) and tests based on a quadratic statistic (e.g., variance-component tests). Power calculations for these tests require specifying a large number of parameters: the number of rare variants in a locus, the proportion of causal variants, the effect size of each causal variant, and, for linear statistics, the direction of the effects. With limited knowledge about the genetic architecture and the corresponding parameters, existing power calculations have very limited utility in practice. Here we propose fast and accurate methods to calculate the power of test statistics in both classes. Through theory and simulations, we demonstrate that power can be approximated using at most three parameters: the proportion of phenotypic variation explained by the locus, the number of variants in a locus, and, for linear statistics, and the proportion of causal variants that are deleterious. We then use this simplified framework to characterize power of association tests that may pre-select variants based on prior functional annotation. These derivations allow us to study effect of sensitivity and specificity of extraneous information on the power. We show that the power can be increased if the AUC for identifying functional variants exceeds 0.70. Because current annotation offers AUCs between 0.70 and 0.75, we suggest using less stringent criteria for variant selection and focusing on removing variants that are most probable to be neutral.

12 | Recessive Selection in Complex Disease: Implications for Variant Discovery and Disease Architecture

Daniel Jordan[1], Daniel Balick[2], Shamil Sunyaev[2], Ron Do[1]

[1]Charles Bronfman Institute for Personalized Medicine, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [2]Department of Medicine, Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

Methods to identify genetic associations for complex diseases generally assume that effects of alleles are additive, meaning that a homozygous genotype confers twice the risk of a heterozygous genotype. Indeed, the majority of GWAS for complex diseases have assumed additive models of risk. Genes under recessive selection exhibit detectably different population dynamics. Leveraging this property, we have developed a novel method to quantify the strength and recessivity of selection of all protein-coding genes across the human genome by comparing European population sequencing data from the Exome Aggregation Consortium (N=35,000) with simulated evolutionary histories for both additive and recessive alleles. This method could inform model choice by identifying genes and pathways likely to be under recessive selection. We find a variety of biological categories enriched in the recessive class, including glycoproteins (p=6.3×10^{-13}), immunoglobulin domains (p=0.023), and inflammatory response (p=0.0052). The enrichment for inflammatory genes suggests that complex diseases with inflammatory and autoimmune components may be under recessive selection, such as Crohn's Disease (CD) and Rheumatoid Arthritis (RA). In the case of CD, we find that well-validated genes are under recessive selection, while genes implicated by GWAS do not show such enrichment. Similarly, RA loci discovered by GWAS show no enrichment for recessive selection, despite RA being known to involve inflammatory and immune pathways that are highly enriched for recessive selection. These examples highlight the usefulness of our catalog of recessive selection as a tool for variant discovery and more broadly, for gaining biological insight into complex disease architecture.

13 | Evidence of Hybrid Vigor in a Human Population from PheWAS

Todd L. Edwards[1,2], Digna R. Velez Edwards[1,3], Chun Li[4], Yaomin Xu[5], Eric S. Torstenson[1,2], Lisa Bastarache[6], Josh C. Denny[6], Dan M. Roden[6,7], Tracy L. McGregor[8]

[1]Vanderbilt Epidemiology Center and Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [2]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [3]Department of Obstetrics and Gynecology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [4]Department of

Epidemiology and Biostatistics, Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, United States of America; [5]Department of Biostatistics, Center for Quantitative Sciences, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [6]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [7]Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [8]Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America

Heterosis, or hybrid vigor, is a phenomenon utilized in agriculture where a hybrid strain has superior characteristics to either pure-bred parental strain. Size, growth rate, yield, fertility, and resistance to disease are often cited as features of heterosis. Heterosis may have also played an evolutionary role in humans by selecting for traits that influence reproductive success. Models propose that heterotic traits exhibit this property due to heterozygosity at causal loci. We observed increasing levels of heterozygosity over time in a clinical population of 33,672 patients from Vanderbilt University Medical Center spanning over 100 birth years and genotyped with exome arrays. Increasing admixture proportions, long-range LD, and variance of admixture proportions over time provide evidence that admixture is recent, increasing, and ongoing. We performed a PheWAS of 1,116 common clinical phenotypes derived from Electronic Health Records (EHR) with heterozygosity levels, adjusted for birth year, principal components of ancestry, and sex when traits occur in both sexes. We detected statistically significant evidence of protection from disorders of menstruation (P value=6E−6) and irregular menstrual cycles or bleeding (P value=4E−5). We detected suggestive evidence of protection from gynecological neoplasias, asthma, and atopic dermatitis (P values=[3E−4 to 1E−3]). We calculated the burden of diagnoses from each individual's EHR and detected a complex protective effect of increasing heterozygosity on risk of diagnoses with restricted cubic splines, adjusted for principal components, sex, and birth year (spline P values=[2E−9 to 0.001]). These findings from a clinical EHR dataset show evidence of heterosis where admixture confers reproductive advantage and health benefits to humans.

14 | UK Biobank GWAS Identifies over 100 Novel Variants Associated with Blood Pressure

Evangelos Evangelou[1,2], Helen Warren[3,4], Claudia Cabrera[3,4], He Gao[2], Ioanna Tzoulaki[2], Michael Barnes[3,4], Mark Caulfield[3,4], Paul Elliott[2], On Behalf of: UKB-CMC BP working group, International Consortium for Blood Pressure, CHARGE+ Consortium, T2D-GENES Consortium, GoT2DGenes Consortium, ExomeBP Consortium and CHD Exome+ Consortium

[1]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece; [2]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [3]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [4]NIHR Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

Genetic factors contribute approximately 30 to 50% of Blood Pressure (BP) variance. Large studies are needed to identify novel variants and unravel the complex genetic architecture of BP.

We assessed associations between genetic variants and systolic BP, diastolic BP and pulse pressure in a large population from UK-Biobank. We investigated ∼73M typed and imputed autosomal variants in a total of 140,866 participants. We selected variants to follow up in a large set of independent datasets (N>200k). Loci were considered validated if they replicated at p<0.01, with concordant direction of effect, and were genome-wide significant in the combined discovery and replication meta-analysis. Our analysis revealed ∼110 common and rare, independent signals at novel loci, and additional new variants in many known regions.

We performed an integrative bioinformatics analysis of novel loci, including evidence of known BP biology, functional impact, eQTLs, tissue expression and Hi-C 3D conformation. We found evidence of biological function potentially relevant to hypertension mechanisms for many novel genes. Moreover, gene set enrichment and pathway analyses identified enriched pathways previously associated to cardiovascular diseases. Furthermore, we assessed the validated variants for associations in untargeted metabolomics profiling of plasma and urine by proton nuclear magnetic resonance (^{1}H-NMR). Findings showed associations with a range of metabolic features and lipoprotein fractions derived from the NMR spectra.

Our work adds important new knowledge to the understanding of the genetic basis of BP and combined with bioinformatics analyses and integration of additional –omics data may reveal potential causal genes.

15 | Test to Identify Co-Localization of Genetic Association Signals Across Multiple Traits Using Summary Statistics

Christopher N. Foley[1], James R. Staley[1], Frank Dudbridge[1,2], Joanna M.M. Howson[1]

[1]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Genome-wide Analysis Studies (GWAS) have been very successful in unravelling genetic components of complex disease. The next challenge is to elucidate causal genes and mechanisms underlying pathophysiology. Given the evidence of widespread pleiotropy across common diseases and quantitative traits, specific genetic variants may be causative of multiple phenotypes. An approach to assess shared genetic

aetiology is to integrate information on gene regulation, gene expression, metabolic pathways and complex diseases under the assumption that: genetic association signals shared between phenotypes are supportive of a causal genomic region. This is known as genetic statistical co-localization.

Current methods to assess co-localization are limited by pairwise comparisons of traits. To overcome this limitation we have developed a two-stage Bayesian statistical method, using only summary statistics from genetic association studies, that can process many traits simultaneously by: (i) identifying shared genetic regions and, for traits that satisfy this condition, (ii) assessing evidence supporting a shared causal variant. The method returns information on candidate causal variants and, in the absence of identifying a putative causal variant, identifies shared genomic regions whose apparently shared signals are driven by a combination of linkage-disequilibrium and/or several causal-variants. We demonstrate the method by testing for overlapping genetic association signals, across the entire genome, with gene expression, cardiovascular risk factors and coronary artery disease simultaneously. In doing so, we provide information about candidate causal gene(s) and prioritise molecular pathways.

16 | Genome-Wide Association Study Identifies Novel Loci for Asthma in the UK Biobank

Audrey V. Grant[1], Markus Munter[1], Bing Ge[1], Toby Hocking[1], Florence Demenais[2], Miriam F. Moffat[3], William O. Cookson[3], Tomi Pastinen[1], Stephen J. Sawcer[4], Mark Lathrop[1]

[1] McGill University and Génome Québec Innovation Centre, Montréal, Canada; [2] Inserm, UMR-946, Paris, France; [3] National Heart and Lung Institute Imperial College London, London, United Kingdom; [4] Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom

Asthma is a complex disease implicating the interplay of environmental exposure, lifestyle and innate genetic factors, which explains differences in distribution across ethnicities. In addition, the heterogeneous genetic contributions to asthma etiology may be partly explained by differences in age of onset. The UK Biobank provides an unprecedented opportunity to identify genetic variation underlying asthma risk in an ethnically homogeneous population. Currently, the analyzed dataset consists of 120,286 genotyped individuals recruited at 40–69 years from the United Kingdom, including 13,985 with self-declared physician-diagnosed asthma. Imputation was performed using ~800,000 genotyped polymorphisms, yielding 9.8 million quality control filtered autosomal markers. These markers were tested for association with asthma using linear mixed models accounting for relatedness within the population, implemented in BOLT, which incorporates marker effect sizes as Bayesian priors to increase computational efficiency. In the MHC region, extended haplotype structure was used to infer high-resolution alleles for 5 HLA genes using European pre-fit classifiers (HIBAG software). We identified 34 regions harboring at least one variant displaying genome-wide significance ($p<5\times10^{-8}$) underlying asthma risk, as well as the MHC region, with some showing significant effects only in childhood onset disease. Relationships to autoimmune diseases and other phenotypes such as lung function and smoking have been explored using the extensive data available in the UK Biobank. The high resolution afforded by this study allowed for the identification of seven novel regions of association. In addition, multiple independent risk loci within previously known regions were identified through conditional regression analyses as well as L1 (LASSO) regularized logistic regression. Functional annotation of identified regions was performed thanks to the BLUEPRINT Epigenome Project which catalogued coordinated genetic effects on gene expression, methylation and histone variation in major human immune cell types on ~200 individuals.

This research has been conducted using the UK Biobank Resource.

17 | Genomics of Lipid Metabolism: Identifying Novel Causal Pathways and New Therapeutic Targets for Reducing risk of Coronary Heart Disease

Eric L. Harshfield[1], David Stacey[1], Dirk S. Paul[1], Albert Koulman[2], Angela M. Wood[1], Adam S. Butterworth[1], Eric Fauman[3], Julian L. Griffin[2], John Danesh[1], Danish Saleheen[4]

[1] Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom; [2] MRC Human Nutrition Research, University of Cambridge, Cambridge, United Kingdom; [3] Computational Sciences Center of Emphasis, Pfizer Worldwide Research and Development, Cambridge, Massachusetts, United States of America; [4] Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Coronary Heart Disease (CHD) is one of the leading causes of death worldwide; mortality rates are expected to continue to rise over the coming decades. Circulating lipids have been shown to be strongly and linearly associated with risk of CHD; however, despite considerable efforts to demonstrate causality, available evidence is conflicting and insufficient. Study of the underlying metabolic pathways implicated in the association between lipids and CHD would help to disentangle and elucidate these complex relationships. Direct infusion high-resolution mass spectrometry was performed on 5,551 participants from the Pakistan Risk of Myocardial Infarction Study; raw data were then processed, cleaned, and normalized to extract signals corresponding to 444 known lipid metabolites. Cross-correlations of lipid metabolites and their correlations with circulating lipids were examined, and the association of principal components of lipid metabolites with CHD risk factors was assessed. Genome-wide analyses were conducted to analyze the association of each lipid metabolite with 7.2 million genotyped and imputed SNPs. Following conditional analyses on the lead SNP within each loci, we identified genome-wide significant associations at 148 independent metabolic loci and 54 novel loci. We then used functional annotation to link the variants associated with each metabolite to the most probable causal genes, and two-sample Mendelian randomization to examine the causal effect of lipid metabolites on risk of CHD. Analyses of circulating lipid metabolites in large epidemiological studies could lead to enhanced understanding of mechanisms for CHD development and identification of novel causal pathways and new therapeutic targets.

18 | Population Differences in Burden of Fibroproliferative Risk-Increasing Alleles Support Selection as a Cause for Racial Disparities

Jacklyn N. Hellwege[1,2], Eric S. Torstenson[1], Shirley B. Russell[1], Digna R. Velez Edwards[1,2,3], Todd L. Edwards[1,2]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

Fibroproliferative Diseases (FD) are common complex traits which vary widely in presentation because scarring and overgrowth of connective tissue can affect many organ systems. Many FD are more prevalent in African-Derived Populations (ADP) than in those of European ancestry, leading to pronounced health disparities. It is hypothesized that the increased prevalence of these diseases in ADPs is due to selection for protection against helminth infections at loci that have pro-fibrotic consequences. We constructed a Genetic Risk Score (GRS) of FD risk-increasing alleles using 100 linkage disequilibrium-pruned variants identified through GWAS of eight FDs with large prevalence disparities. A comparison of the FD GRS between 1KGP continental populations detected a higher mean GRS in AFR than EUR (T-test *P* value=2.46×10^{-142}). To test whether differences in GRS burden are systematic and may be due to selection, we have developed a novel approach, the Trait-Based Selection Test (TBST). TBST measures the difference in the cumulative distribution functions of the GRS burden of risk- or trait-increasing alleles between two genetically differentiated populations while accounting for the background differences in allele frequencies. Large-scale forward-time genome-wide simulations with and without selection at 100 loci demonstrate good performance and controlled type I error, outperforming anti-conservative T-tests of the GRS and the conservative sign test, which uses summary data. Evaluation of this FD GRS using TBST indicates that the population differences in risk-increasing allele burdens at these FD SNPs are systematic and support a model featuring selective pressure (*P* value=0.034).

19 | A New Method for Genetic Region Association Testing with Massively Different Sequencing Depths of Coverage

Audrey E. Hendricks[1,2,3,4], Stephen Billups[1], Eleftheria Zeggini[4], Inês Barroso[4,5], Stephanie A. Santorico[1,2,3], Josée Dupuis[6]

[1]Mathematical and Statistical Sciences-University of Colorado-Denver, Denver, Colorado, United States of America; [2]Human Medical Genetics and Genomics Program-University of Colorado-Denver, Aurora, Colorado, United States of America; [3]Biostatistics and Informatics-Colorado School of Public Health, Aurora, Colorado, United States of America; [4]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [5]University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom; [6]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America

Large samples are needed to achieve adequate power in case-control studies of rare variants. To improve power, one might use whole-genome sequenced population controls. However, differences in sequencing depth between cases and controls cause severe bias with traditional statistics. An alternative is case-only analysis, which is not susceptible to case-control bias, where the rate of variants within a gene region of interest is compared to the genome-wide average. Case-only analysis can achieve higher power than case-control analysis but is highly sensitive to regional differences in variant frequencies not related to case-status (e.g., mutation rate, annotation accuracy, sequencing accessibility). To address this, we have developed a simple, closed form genetic region test statistic that combines case-only and case-control information enabling analysis in samples with extremely different sequencing depths (e.g., 10× versus 50×). We present results from a wide-variety of simulations and application to a UK10K dataset of 926 whole-exome sequenced cases (~80×) and 3,621 whole-genome sequenced controls (~7×). Our method achieves equivalent power to existing case-control and case-only methods while maintaining appropriate type I error in the context of region-level and/or case-control biases. Specifically, when we simulate cases to have 30% more rare variants compared to controls, the type I error increases to $>$ 20% for traditional case-control tests while our method maintains the expected type I error of 5%. As this method can combine datasets with drastically different sequencing depths, there is the potential to greatly increase the sample size and subsequently power by harnessing publicly available resources.

20 | Constrained Instrumental Variable Approach and its Application to Mendelian Randomization with Pleiotropy

Lai Jiang[1], Celia Greenwood[1,2], Karim Oualkacha[3]

[1] *Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada;* [2] *Departments of Epidemiology, Biostatistics & Occupational Health, Oncology, and Human Genetics, McGill University, Montreal, Canada;* [3] *Département de mathématiques, Université du Québec à Montréal, Montreal, Canada*

In Mendelian randomization, genetic variants (SNPs) are used to construct an instrumental variable to estimate the causal effect of a phenotype (or exposure) on a disease. However, the existence of pleiotropy violates the assumption that the instrument (SNPs) and the response (disease) are independent conditional on the phenotype of interest. As a result, the ordinary two-stage least squares estimator, using all desired SNPs as instrumental variables, overestimates the required causal effect.

We propose novel constrained methods to perform adjusted causal effect estimation, by finding a projection orthogonal to a set of possibly pleiotropic phenotypes that are not of interest. Assuming that there are sufficient potential genetic instruments, constrained quadratic optimization with a smoothed-L0 norm can correct the bias induced by pleiotropy and lead to sparse models and stable instrumental variables.

In simulations, we compared our approach to a naive method (using all SNPs), the limited information maximum likelihood method, Canonical Correlation Analysis (CCA/sparseCCA), and constrained stepwise selection methods. Results show our approach leads to causal effect estimators with the smallest bias and variance among those compared. We are currently analyzing human blood metabolite data from TwinsUK studies, where our goal is to obtain adjusted estimates of the causal effects of gene expression on metabolite levels, using SNPs as instrumental variables.

In conclusion, our approach finds a robust sparse model, enforces automatic feature selection, and leads to better estimate of causal effects even when pleiotropy is present.

21 | Why Real Biological Interactions are Usually not Detectable in Genetic Association Analyses

Nuri Kodaman[1], Scott M. Williams[1]

[1] *Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, United States of America*

Association studies assessing interactions between genetic variants and background factors (such as environment, sex, or other physiological traits) have failed to explain much additional genetic variance, despite strong evidence indicating that genetic effects are context-dependent at the organismal level. A possible explanation for this paradox is that the "main effect" term in statistical regression models is more sensitive to biological interactions than the interaction term. We developed population genetic models of context-dependent genetic effects to explore this possibility. First, we show that the conventional interaction term used in linear regression equations inadequately models context-dependent genetic effects, particularly when the interacting variable does not induce a change in the *direction* of the genetic variant's effect at the biological level. With this constraint imposed, we show that the ratio of the proportion of variance explained by the main effect vs. the interaction effect in a regression analysis can be expected to have a lower bound of $2/\pi$. We also modeled a continuous outcome such that the slope of the genetic effect increases at incremental quantiles of an interacting variable. According to this model, as the number of quantiles increases from 2 to ∞, the ratio of expected proportion of variance explained by the main effect vs. the interaction effect ranges from $9\pi/2$ to π, and does not vary with other parameters. Our models indicate that highly heterogeneous genetic architecture can account for the large number of small, additive effects observed in genetic association studies, but that relying solely on interaction P values will fail to reveal this complexity. Statistical methods that jointly consider main effects and interaction effects should be most powerful. To test such methods, we provide a flexible way to simulate realistically structured and biologically plausible interaction data.

22 | Bayesian Meta-Analysis for Cross-Phenotype Genetic Association Study

Arunabha Majumdar[1], Sourabh Bhattacharya[2], John S. Witte[1]

[1] *University of California, San Francisco, California, United States of America;* [2] *Indian Statistical Institute, Kolkata, India*

Simultaneous analysis of genetic association across multiple traits may reveal shared genetic susceptibility among traits (pleiotropy). Alongside measuring the overall pleiotropic association, it is crucial to identify the traits associated with a risk locus because only a subset of traits may have true genetic effects. We propose a Bayesian cross-phenotype meta-analysis using spike and slab prior that provides a posterior probability of association (Stephens and Balding, 2009) measuring global association and an optimal subset of associated traits. In spike and slab prior, the spike corresponding to null genetic effects can either be a positive mass at zero (Dirac spike), or a normal distribution with zero mean and a variance smaller than that of a zero mean normal representing the non-null effects (continuous spike). Gibbs samplers are designed for uncorrelated and correlated summary statistics across traits. Simulations show that a continuous spike is a better alternative than Dirac spike. This meta-analysis allows heterogeneity in the direction and magnitude of genetic effects. It is applicable to both cohort data and multiple studies of different traits. For strongly correlated summary statistics, the Gibbs sampler can mix poorly due to multi-modality issue; hence we propose a joint strategy combining the uncorrelated and correlated versions of the meta-analysis. Simulations show that it offers substantially better accuracy in selecting associated traits than a subset-based meta-analysis

ASSET (Bhattacharjee et al., 2012). We analyze 22 traits in the Kaiser cohort and identify several loci associated with at least two traits.

23 | Novel Locus Discovery Through Trans-Ethnic Association Analyses of Glycemic Traits Using Densely Imputed Genetic Data

Gaelle Marenne[1], On Behalf of the Meta-Analyses of Glucose and Insulin-Related Traits Consortium (MAGIC) Investigators

[1] Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Previous large-scale glycemic trait genetic association analyses have identified more than 120 loci associated with Fasting Glucose (FG), Fasting Insulin (FI), glycated Hemoglobin (HbA1c) and 2-hour Glucose (2hG). To aid additional locus discovery and fine-mapping, we conducted meta-analyses of 281,416 individuals without diabetes from five ancestries (71% Europeans, 13% East Asian, 7% Hispanic, 6% African American and 3% South Asian), imputed to 1000 Genomes reference panel.

Genetic association analyses results were combined by ancestry-specific meta-analyses, followed by trans-ethnic meta-analyses, which allows for heterogeneity between diverse ethnic groups. We used a $\log_{10}$ Bayes factor ($\log_{10}$BF) threshold of 6 to identify genome-wide significant signals. Lastly, GARFIELD was used to perform functional enrichment analysis.

Preliminary results identified 132 regions associated with FG, 93 new including those at *CDK14* and *GAD2* genes ($\log_{10}$BF=19.75 and 16.44). We also identified 78, 189 and 31 regions associated with FI, HbA1c and 2hG respectively (61, 138 and 23 novel). The most significant novel signals fall within *BCL2*, *HBB* and *DVL2* genes ($\log_{10}$BF=11.56, 25.28 and 11.10, respectively). Enrichment analysis showed that HbA1c signals were mostly enriched in open chromatin seen in blood cell type, whereas FG and 2hG signals were mostly enriched in liver cell type, and FI signals were mostly enriched in fetal lung cell type.

In conclusion, this large international effort has identified over a hundred novel loci, and suggested new hypotheses about the biology and the genetic architecture underlying glycemic traits.

24 | TreeLMM: Modelling Heterogeneity of Genetic Effects

Rachel Moore[1,2], Francesco Paolo Casale[2], Inês Barroso[1], Oliver Stegle[2]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom; [2] European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Genetic effects can vary between individuals. For example, the effect of a genetic variant may depend on environmental exposures such as smoking status, and other lifestyle covariates. However, widely used approaches to analyze genotype-phenotype associations commonly assume identical genetic effects across all samples within a cohort.

One approach to address these differences in genetic effects is to divide the samples into distinct groups and then either perform analyses for individual groups followed by meta-analysis, or conduct multi-trait modelling. This can result in a prohibitive numbers of groups, with small sample sizes, that are not fully independent of one another. Hence, independent analyses do not fully leverage the structure in the data.

We propose TreeLMM, an approach that explicitly accounts for heterogeneity in effect size and group structure; achieved by defining a prior on effect sizes that encodes the similarity between different groups of individuals. The method is efficient, facilitating the analysis of thousands of samples and arbitrary number of groups.

To illustrate our approach, we have performed eQTL mapping in 44 tissues from the GTEx project. We have leveraged expression profiles to estimate similarities and mapped *cis*-eQTLs across the cohort of 7,051 samples. Compared to independent modelling and methods that ignore heterogeneity in effect size, we identify a 2-fold, and 4-fold increase in the number of independent variants associated with gene expression changes. Additionally, we can estimate posterior probabilities of the effect of genetic variants in individual tissues enabling us to dissect tissue-specific effects as well as secondary association signals.

25 | Investigating DNA Methylation as a Marker for Historical Smoke Exposure and a Mediator of Disease Risk

Rebecca C. Richmond[1], Matthew Suderman[1], Philip Haycock[1], Gibran Hemani[1], Caroline L. Relton[1], George Davey Smith[1]

[1] MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

DNA methylation at some CpG sites related to in-utero smoke exposure shows persistence in childhood and adolescence. Of interest is whether signals persist into adulthood. We investigated associations between maternal smoking in pregnancy and methylation in peripheral blood among women in the Avon Longitudinal Study of Parents and Children. Among women with a mean age of 29 years (N=866), there was an inflated signal above that expected by chance at CpGs previously associated with in-utero smoke exposure in newborns (lambda=2.80 vs. 1.05 for all CpG sites on the Illumina Infinium HM450 array). In-utero smoke exposure was associated with hypermethylation at 3 CpGs in *MYO1G* ($p<1\times10^{-7}$). These signals remained when adjusted for own smoking and were observed in peripheral blood when the women were a mean age of 47 years (N=885). A persistent methylation signal was also identified in whole blood and LCLs in the 1958 Birth Cohort (mean age=45 years; *N*=40) and in saliva in the AFAST study (mean age=47 years; *N*=83). We next investigated whether persistent methylation at *MYO1G* plays a causal role in the aetiology of disease using a hypothesis-free Mendelian randomization approach. We identified cis-SNPs associated with methylation at *MYO1G* and performed an agnostic look-up of these SNPs in GWAS summary data. We found enrichment for a causal effect of *MYO1G* methylation on a range of metabolites. These findings suggest that DNA methylation may be used as a biosocial archive for historical exposure and indicate that methylation change related to in-utero smoke exposure may have health implications.

26 | Novel Genome-Wide Sequence Variants Influence Antibody Response to Epstein-Barr Virus in an African Population

Neneh Sallah[1,2], Tommy Carstensen[1,3], Katie Wakeham[4], Rachel Bagni[5], Nazzarena Labo[6], Martin O. Pollard[1,3], Deepti Gurdasani[1,3], Kenneth Ekoru[1,3], Segun Fatumo[1,3], Cristina Pomilla (1,3,), Elizabeth H. Young (1,3,) Gershim Asiki[4], Anatoli Kamali[4], Manjinder Sandhu[1,3], Paul Kellam[2], Denise Whitby[6], Robert Newton[4], Inês Barroso[1]

[1]Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom; [2]Virus Genomics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom; [3]Department of Medicine, University of Cambridge, Cambridge, United Kingdom; [4]MRC/Uganda Virus Research Institute, Uganda Research Unit on AIDS, Entebbe, Uganda; [5]Protein Expression Lab, Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America; [6]Viral Oncology Section, Aids and Cancer Program, Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America

Globally, 95% of the adult population are infected with Epstein-Barr Virus (EBV), a common human herpesvirus. While infection is lifelong and generally asymptomatic, EBV is associated with 200,000 new cases of cancer and more than 140,000 deaths annually. How host genetic variation influences infectious disease traits such as EBV is largely unknown, particularly in Africa. As Immunoglobulin G (IgG) antibody levels to EBV have been shown to be heritable and associated with developing malignancies, we use it as a proxy for infection and potential disease risk. We therefore explore the heritability and perform the first genome-wide association analysis of anti-EBV IgG traits in an African population, using a combined approach including array genotyping, whole-genome sequencing and imputation to a panel with African sequence data to extensively capture genetic variation and aid locus discovery. In 1562 Ugandans, we identify two novel African-specific loci associated with anti-VCA IgG responses, an intergenic variant on chromosome 7 (p=4.0×10^{-10}) and an intronic variant in *GALC* (p=6.8×10^{-10}). We also identify a variant in *HLA-DQA1* (p=2.6×10^{-17}) associated with anti-EBNA-1 responses. Trans-ancestry meta-analysis and fine-mapping with European-ancestry individuals suggest the presence of distinct *HLA* class II variants driving associations in Uganda. Our study reinforces the importance of studying diverse populations to uncover population specific variants, differences in effect sizes and gene-environment interactions which are known to vary significantly between European and non-European populations.

27 | Quantifying Treatment Benefit in Molecular Subgroups to Assess a Predictive Biomarker

Jaya M. Satagopan[1], Alexia Iasonos[1]

[1]Memorial Sloan Kettering Cancer Center, New York, New York, United States of America

We are now witnessing an increasing interest in finding predictive biomarkers that can guide treatment or preventive intervention options for mutation carriers and non-carriers. The statistical assessment of variation in Treatment Benefit (TB) according to the biomarker carrier status plays an important role in evaluating predictive biomarkers. For time to event endpoints, the Hazard Ratio (HR) for interaction between treatment and a biomarker from a Proportional Hazards regression model is commonly used as a measure of variation in treatment benefit. While this can be easily obtained using available statistical software packages, the interpretation of HR is not straightforward. We propose different and clinically interpretable summary measures of variation in TB on the scale of survival probabilities beyond a specific time point for evaluating a predictive biomarker. The proposed summary measures can be easily interpreted as quantifying variation in TB in terms of relative survival or excess absolute survival associated with treatment in carriers versus non-carriers. We illustrate the use and interpretation of our proposed measures using published data from completed clinical trials of: (1) tamoxifen treatment, progesterone receptor, and disease-free survival in breast cancer; (2) immunotherapy with panitumumab, *KRAS* mutation, and progression free survival in *EGFR*-positive metastatic colorectal cancer; and (3) combination immunotherapy with nivolumab and ipilimumab, *KRAS* mutation, and progression free survival in metastatic melanoma. We recommend interpreting variation in TB in terms of measures based on survival probabilities, particularly in terms of excess absolute survival, as opposed to HR.

28 | Calibration Testing for Survival Models at the Extremes of Risk: Implications of Unobserved Genetic Interactions

David Soave[1,2], Lisa J. Strug[1,2]

[1]University of Toronto, Toronto, Canada; [2]The Hospital for Sick Children, Toronto, Canada

Risk prediction models can translate genetic association findings to inform clinical decision-making. Most models are evaluated on their ability to discriminate, and the calibration of risk-prediction models is largely overlooked in applications. Models that demonstrate good discrimination in training datasets, if not properly calibrated to produce unbiased estimates of risk, will generally perform poorly in new patient populations. Poorly calibrated models arise due to missing covariates, such as genetic interactions that in many instances are unknown or not measured. We demonstrate that models omitting interaction effects can lead to increased bias in predicted risk for patients at the tails of the risk distribution; i.e. those patients most likely to be affected by clinical decision making. We propose a new calibration test for Cox risk-prediction models for time-to-event data that leverages power from bias in risk estimates at the extremes. Our test aggregates martingale residuals for subjects from extreme high and low risk groups with a test statistic maximum chosen by varying which risk groups are included in the extremes. We show how to estimate the empirical significance of our test statistic by simulating from a Gaussian distribution using the covariance matrix for the grouped sums of martingale residuals. A simulation study shows our new test maintains control of type 1 error and demonstrates improved power over conventional goodness-of-fit tests when risk prediction is poorest at the tails of the risk distribution. We apply our method in the development of a prediction model for onset of cystic fibrosis-related diabetes.

29 | A General Framework for Association Analysis of Microbial Community on a Taxonomic Tree

Zheng-Zheng Tang[1], Guanhua Chen[1], Hongzhe Li[2]

[1]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [2]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America

Biological and empirical evidence have suggested that microbiome plays an important role in human health. Recent advances in high-throughput sequencing technology have made it possible to obtain data on the composition of microbial communities and to study the effects of dysbiosis on the human host. We have developed a general framework to (a) perform robust association test at any taxonomic level for the microbiome data exhibiting arbitrary inter-taxa correlation; (b) localize the covariate-associated lineages on the taxonomic tree; (c) assess the overall association of the microbial community. Unlike the existing association analysis of the microbial community, our framework does not make any distributional assumption on the taxa count, allows for adjustment of the confounding variables, accommodates excessive zero observations, and incorporates the information of the taxonomic tree. We performed extensive simulation studies to evaluate the new methods under a wide-range of scenarios and demonstrated their advantages over existing methods. An application to a cutaneous microbiome study led to novel discoveries of microbial lineages associated with psoriasis.

30 | Hierarchical Bayesian Modeling of Mediation by High-Dimensional Omics Data

Duncan Thomas[1]

[1]University of Southern California, Los Angeles, California, United States of America

Various high-dimensional epigenetic, transcriptomic, proteomic, metabolomic, and other –omic data have become available to provide insight into the mediation of genetic and environmental influences on disease risk through the internal environment. For example, the "exposome" concept has been implemented using mass spectrometry metabolomic measurements to capture a broad spectrum of internal metabolites of exogenous exposures, but statistical methods for analyzing these and other -omic data are in their infancy. The "Meeting-in-the-Middle" principle aims to identify the subset of metabolites that are related to both exposure and disease. Here, we introduce a novel hierarchical Bayes framework for implementing this idea through simultaneous variable selection on exposure-metabolite and metabolite-disease associations, while incorporating external information such as the pathways in which the different metabolites are thought to act. The approach is validated by simulation and applied to data on hepatocellular carcinoma of the liver in relation to a panel of 125 metabolites and 7 established risk factors from a nested case-control study within the EPIC cohort. 15 of the metabolites yielded Bayes factors for mediation greater that 20 ("strong" evidence), the majority of these with multiple exposures. To explore this phenomenon further, we expanded the hierarchical model to include the pathways through which these metabolites act as prior covariates. The strongest associations with exposures were found for the class of lysophosphatidylcholines and the strongest with disease for biogenic amines and acylcarnitines. These approaches could be extended to study mediation through multiple types of –omic data.

31 | Kinship Estimation Based on Extremely Low-Coverage Sequencing Data

Jinzhuang Dou[1], Sonia Chothani[1], Xueling Sim[2], Jason D. Hughes[3], Dermot F. Reilly[3], E. Shyong Tai[2], Jianjun Liu[1], Chaolong Wang[1]

[1]Genome Institute of Singapore, Singapore, Singapore; [2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore; [3]Merck Research Laboratories, Kenilworth, New Jersey, United States of America

Estimation of kinship is important to genetic association studies, both for control of cryptic relatedness to avoid spurious associations and for estimation of trait heritability. However, estimation of kinship is challenging for target sequencing studies, where sequencing efforts focus on small regions of the genome. Existing methods often assume accurate genotypes at a large number of markers across the genome. We show that these methods, without accounting for the genotype uncertainty in shallow sequencing data, can yield a strong downward bias in kinship estimation. We develop a computationally efficient method called SEEKIN to estimate kinship for both homogeneous samples and samples with population structure and admixture. Our method explicitly models the genotype uncertainty and leverages linkage disequilibrium through imputation. We test SEEKIN on a whole exome sequencing dataset of Singapore, Chinese and Malays, including many admixed samples. We show that SEEKIN can accurately estimate kinship coefficient using off-target sequencing data down sampled to as low as 0.1× depth. In application to the full dataset without down sampling, SEEKIN also outperforms existing methods to estimate kinship by properly analyzing the off-target data. Our method enables control of cryptic relatedness in target sequencing studies without additional genotyping data.

32 | Testing of Parent-of-Origin Effect in eQTL Mapping Using RNA-seq Data

Feifei Xiao[1], Guoshuai Cai[2], Christopher I. Amos[3]

[1]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America; [2]Department of Molecular and Systems Biology, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America; [3]Department of Biomedical Data Science, Dartmouth College, Hanover, New Hampshire, United States of America

Genomic imprinting is an important epigenetic phenomenon where the expression of certain genes depends on their parent-of-origin. Many imprinting genes are known to play important roles in human complex diseases such as diabetes, breast cancer and obesity. In recent years, array based eQTL studies have identified many regulatory variants that show associations with gene expression level. Nowadays, the rapid arising next-generation-sequencing is often done for eQTL mapping. We believe that parent-of-origin effect can contribute to regulating gene expression along with the overall effect from the gene. However, multicollinearity occurs naturally when we are modelling multiple genetic components, such as additive, dominance and imprinting effects. Moreover, it has been repeatedly shown that RNA-seq data are overdispersed which brings challenge to the modelling of the gene expression profiling. To address these issues, we introduced a novel method to test the main allelic effects along with the parent-of-origin effect in detecting eQTL. We utilized an orthogonalization procedure, which allowed for efficient imprinting effect detection whereas maintained the power to detect the main allelic effect from eQTLs. We conducted extensive simulations to demonstrate the statistical behavior of the proposed method. We also applied the models to a large-scale breast cancer study and revealed potential imprinted regulatory elements that control gene expression.

33 | Type I Probes on the Illumina Methylation Array are Systematically Biased to more Extreme Methylation Values

Nora Zwingerman[1], Mathieu Lemire[2], Jessica Dennis[1], David-Alexandre Trégouët[3,4], France Gagnon[1]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2]Ontario Institute for Cancer Research, Toronto, Canada; [3]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé 116, Paris, France; [4]Institute for Cardiometabolism and Nutrition, Paris, France

The Illumina 450K Methylation Array (450M) is widely used in epigenomic association studies. The array utilizes two probe types (types I and II, each characterized by different features and design) to measure methylation. Normalization methods assume that type I probe distribution reflects true underlying biology, thus aim to normalize type II probe distribution to resemble that of type I.

The objective is to investigate whether probe features introduce a bias in intensity distributions and methylation levels.

Raw and normalized data from the 450M measured in 600 individuals were used to assess the effects of probes features on methylation intensity distributions and levels. The primary features investigated were the probe type and number of CpG sites within the probe.

The type I probes have a more extreme distribution compared to type II probes. These differences are not explained by different genomic contexts of the cytosine's interrogated (e.g., CpG Islands, gene body). For type I probes the density distributions reflect the underlying number of CpG sites in the probe, with greater number of CpG sites leading to more extreme distributions. Raw type II density distributions do not vary by the number of CpG sites in the probe. Normalization methods also result in changes to the correlation between CpGs.

Type I probes appear to have a systematic bias to more extreme values related to the number of CpGs in the probe sequence likely due to preferential hybridization with the probes. This has important implications on the interpretation of the 450M findings.

POSTER PRESENTATIONS

34 | Type 2 Diabetes Genes with Cross-Traits Relevance: Identifying Genetic Links of Common Diabetes with its Comorbidities

Karla V. Allebrandt[1], Hartmut Ruetten[1], Daniel Crowther[2], Francesca Frau[1]

[1]Department of Translational Medicine, R&D Diabetes Division, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany; [2]Structure, Design & Informatics Frankfurt, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany

The comorbidities associated with type 2 diabetes (T2D) suggest a common etiology for these phenotypes and complicate the management of the disease. In this study, we focused on the genetics underlying these relationships, using systems genomics to identify genetic variation associated with T2D and 12 other traits. T2D associated SNPs were used as a reference panel to identify corresponding association signals with other traits, based on empirical estimates of linkage disequilibrium between SNPs. GWAS summary statistics for pairwise comparisons were obtained for glycemic traits, obesity, coronary artery disease, and lipids from large consortia GWAS meta-analyses. We identified 38 T2D associated loci with pleiotropic effects in at least two other domains (glycemic, obesity, lipids or CAD). We could demonstrate the robustness of these findings leveraging published experimental evidence showing that many of the pleiotropic loci: (i) had functional studies supporting their relevance for the associated traits, (ii) 30 were cis eQTLs (expression quantitative trait loci) associated with transcript levels in different tissues, (iii) some mapped to genes that were differentially expressed in the islets of T2D patients versus healthy individuals, and (vi) there was an overconnectivity of pleiotropic genes. In this study, we demonstrated how systems genomics and network medicine approaches can shed light into T2D GWAS discoveries, translating findings into a more therapeutically relevant context. As a result, we identified a set of genetic variants with pleiotropic effects that point to a main network of genes that are relevant for T2D and its comorbidities.

35 | Genome-Wide Analyses of Survival Time in the Rare Disease, Idiopathic Pulmonary Fibrosis

Richard Allen[1], Martin Tobin[1], Louise Wain[1], Rebecca Braybrooke[2], Ian Hall[3], Ian Sayers[3], Gisli Jenkins[3], On Behalf of the United Kingdom ILD Consortium

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Division of Epidemiology and Public Health, University of Nottingham, United Kingdom; [3]Division of Respiratory Medicine, University of Nottingham, United Kingdom

Idiopathic Pulmonary Fibrosis (IPF) is a rare lung disease with poor prognosis (median survival time of 3 years) characterized by scarring of lung tissue. IPF has been linked with a number of environmental and genetic factors; the strongest genetic association being in the *MUC5B* gene. Despite this the pathogenesis of IPF is still unclear and more needs to be done to understand the genetic basis of IPF.

We are conducting analyses genome-wide to investigate survival time in 565 IPF cases. Previous evidence suggests variants associated with susceptibility to IPF may not be associated with survival time or may even have effects in the opposite direction. A genome-wide association case-control study was also conducted allowing variants associated with susceptibility to IPF and survival time to be directly compared.

This analysis has raised a number of methodological issues such as which survival models to fit, how well survival models fit variants of different allele frequencies, and how these influence power. The statistical power relates not only to the allele frequency but also to the number of events in each genotype group. In this study, we fitted a Cox proportional hazards model, which makes no assumption about the distribution of the underlying baseline hazard function, and compared findings with those from alternative models.

36 | Socioeconomic Status, Genetic Risk, and How their Interactions Affect Risk of Oral Clefts

Lynn M. Almli[1], Mary M. Jenkins[2], Camden P. Bay[3], Paul A. Romitti[3], Lina Moreno-Uribe[3], George L. Wehby[3], and the National Birth Defects Prevention Study

[1]Carter Consulting, Inc. Atlanta, Georgia, United States of America; [2]Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America; [3]University of Iowa, Iowa City, Iowa, United States of America

Using data from a population-based, case-control study (National Birth Defects Prevention Study [NBDPS]), we examined whether the association between socioeconomic status (SES) and risk of cleft lip with/without cleft palate (CL/P) was modified by genetic predisposition to CL/P. Our sample included 614 cases with isolated CL/P and 2213 controls without any major birth defect. SES was estimated by maternal education (≤high school [HS] vs >HS) or household income (1st and 2nd tertiles vs 3rd tertile). Genetic risk was estimated by a Genetic Risk Score (GRS) based on a set of genes and loci associated with CL/P in previous GWAS. Logistic regression analyses were used to estimate CL/P risk adjusting for maternal age, NBDPS site, and maternal race/ethnicity. The risk for CL/P was greater for mothers with ≤HS compared to mothers with >HS (OR=1.27, 95% CI 1.03-1.58) and for mothers in the lowest income tertile compared to those in the highest tertile (OR=1.35, 95%CI 1.01-1.80). A significant SES×GRS interaction was observed: the risk of CL/P associated with ≤HS was higher at 1 standard deviation (SD) below the GRS mean (i.e., a low genetic risk profile; OR=2.00, 95%CI 1.43-2.81) than at the mean GRS (OR=1.40, 95%CI 1.11-1.76). Similarly, comparing the lowest tertile of income to the highest, the association with CL/P was stronger at 1 SD below the GRS mean (OR=2.13, 95%CI 1.36-3.33) than at the mean GRS (OR=1.47, 95%CI 1.07-2.00). Our preliminary results suggest that SES is a more influential risk factor for CL/P among individuals with lower genetic risk scores.

37 | High-Resolution Analyses of Effects of Polygenic Risk Scores on Time-to-Event Outcomes in Longitudinal Studies

Konstantin G. Arbeev[1], Liubov S. Arbeeva[1], Ilya Y. Zhbannikov[1], Olivia Bagley[1], Igor Akushevich[1], Mikhail Kovtun[1], Alexander M. Kulminski[1], Irina V. Culminskaya[1], Svetlana V. Ukraintseva[1], Anatoliy I. Yashin[1]

[1]Biodemography of Aging Research Unit (BARU), Social Science Research Institute, Duke University, Durham, North Carolina, United States of America

Polygenic Risk Scores (PRS) are widely used to summarize genetic effects of multiple markers on various traits of interest. Longitudinal studies provide the opportunity to analyze PRS in relation to time-to-event outcomes. Traditional methods to estimate the effects of individual genetic markers or PRS on time-to-event outcomes (such as Cox regression model) can be complemented with approaches that utilize data on genotyped and non-genotyped individuals which can increase power compared to analyses of genotyped individuals alone. We developed a software tool for automated high-resolution analyses (i.e., with a large number of thresholds) of effects of PRS on time-to-event outcomes using such methods. We applied the tool to analyses of data from (GWAS) of Alzheimer's Disease (AD) in the subsample of white participants from the Cardiovascular Health Study (CHS) Candidate Gene Association Resource (CARe). The results of GWAS of AD were used to construct PRS in high-resolution PRS analyses of lifespan in CHS CARe data using the Cox regression model. The results show that the best-fitting score is constructed using the P value threshold 0.0088 (PRS_0088) and that the score has significant effect on lifespan (P=0.0077, HR=1.06, 95%CI: [1.01, 1.10]). We also observed a similar effect for dichotomized PRS_0088 (above/below median) in the approach combining mortality data from genotyped and non-genotyped participants. We will discuss practical aspects of implementation and further generalizations that involve models for joint analyses of time-to-event outcomes and repeated measurements of biomarkers.

38 | Linkage Analyses Reveal Significant Signals on Multiple Chromosomes for Familial Lung Cancer

Joan E. Bailey-Wilson[1], Anthony M. Musolf[1], Claire L. Simpson[1], Mariza de Andrade[2], Diptasri Mandal[3], Colette Gaba[4], Ping Yang[2], Ming You[5], Elena Y. Kupert[5], Marshall W. Anderson[5], Ann G. Schwartz[6], Susan M. Pinney[7], Christopher I. Amos[8]

[1]National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2]Mayo Clinic, Rochester, Minnesota, United States of America; [3]Louisiana State University Health Sciences Center, New Orleans, Louisiana, United States of America; [4]University of Toledo Dana Cancer Center, Toledo, Ohio, United States of America; [5]Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America; [6]Karmanos Cancer Institute, Wayne State University, Detroit, Michigan, United States of America; [7]University of Cincinnati College of Medicine, Cincinnati, Ohio, United States of America; [8]Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, United States of America

Lung cancer is the leading cancer killer in the U.S. and risk increases with environmental exposures such as cigarette smoking; there is also a substantial genetic risk. We analyzed ~240,000 SNPs on 190 individuals from 26 families with a family history of lung cancer (Genetic Epidemiology of Lung Cancer Consortium). An affected-only model, autosomal dominant, 80% penetrance in carriers and 1% in non-carriers was used for: two-point linkage using TwoPointLods; multipoint analyses using SIMWALK2 and regional-based linkage using SEQLinkage and MERLIN. SEQLinkage builds multiallelic regional haplotype-based markers using rare variants (here MAF < 0.15) within a gene or a portion of a gene. Two-point linkage was then performed on the regional markers using MERLIN.

We found 7 regional markers (genes) with heterogeneity LOD (HLOD) scores that were genome-wide significant (HLOD≥3.4) on chromosomes 2, 3, 6, 8, 16, 18, and 20. Four of the genes with highest HLODs have been implicated in cancer, three in lung cancer. The highest HLOD (3.97) was on chromosome 18 at the PTPRM gene, which has been reported to control methylation patterns in pulmonary tumors. Multipoint analyses showed a signal on chromosome 18 near PTPRM (not genome-wide significant; most likely due to pruning of SNPs to remove linkage disequilibrium, resulting in a sparse map). The regional approach has been shown to control Type I error while improving power when using rare variants. Because both locus and allelic heterogeneity are common in familial cancers, different families may contain different high risk variants.

39 | Modelling Complex Genetic Architectures: An Interaction between a Variant in *PNPLA3* and Multiple Metabolic Risk Factors for Liver Attenuation

Llilda Barata[1], Mary F. Feitosa[1], Lawrence F. Bielak[2], Brian Halligan[3], Abigail Baldridge[4], Jie Yao[5], Albert V. Smith[6,7], Xiuqing Guo[5], Laura J. Rasmussen-Torvik[4], Jeffrey R. O'Connell[8], Patricia A. Peyser[2], Ingrid Borecki[9], Elizabeth K. Speliotes[3], Michael Province[1]

[1]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, United States of America; [2]Department of Epidemiology, School of Public Health,

University of Michigan, Ann Arbor, Michigan United States of America; [3]Division of Gastroenterology, Department of Internal Medicine, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America; [4]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois United States of America; [5]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, LABioMed at Harbor-UCLA Medical Center, Torrance, California, United States of America; [6]Icelandic Heart Association, Kopavogur, Iceland; [7]Faculty of Medicine, University of Iceland, Reykjavik, Iceland; [8]Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, United States of America; [9]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, New York, United States of America

Robust detection and modelling of interactions for complex traits in humans has been challenging, due to the subtlety of individual genetic signals and the complexity of the biological networks. For Liver Attenuation (LA), a computed tomography quantitative measure of hepatic steatosis, multiple GWAS loci have been identified, and multiple lipid and metabolic traits are known risk factors, but the exact interplay between these variants and risk factors is unclear. Using data from multiple cohorts (FramHS, GENOA, Amish, AGES, MESA, CARDIA, and FamHS: N=14,055 individuals; 11,174 European ancestry; 2,881 African ancestry), we looked for statistical interactions between 5 validated GWAS LA variants and multiple metabolic risk factors, including insulin resistance (HOMA-IR), insulin and glucose levels, Body Mass Index (BMI), waist-hip-ratio adjusted for BMI (WHRadjBMI), triglycerides, HDL-cholesterol, and LDL-cholesterol. We found a robust interaction between *PNPLA3*-rs738409 and several metabolic traits in the association to LA in multiple ethnic populations.

We conducted meta-analyses on the parameters of mixed-model interaction models ($Y_{LA}=\alpha+\beta SNP+\beta Trait+\beta Trait*SNP$) fit to each study for LA. We adjusted for sex, age, alcohol consumption and population substructure, and excluded diabetics and statin users. We found statistically significant interactions of *PNPLA3*-rs738409-G with increased insulin, HOMA-IR, glucose, BMI, WHRadjBMI and triglycerides, after Bonferroni correction, in predicting LA (P values ranged from $1.20\times10^{-11}$to 1.44×10^{-2}). The *PNPLA3*-rs738409 high risk G-allele (frequency=0.23) increased the impact of these metabolic abnormalities on the amount of liver fat. Such interaction models can help dissect the complex interplay between genetic and measured factors on complex disease phenotypes.

40 | Quantifying the Contribution of Genetically Predicted Endophenotypes via Variance Components with Error in Variables Analysis

Alvaro Barbeira[1], Yang Li[2], Hae Kyung Im[1]

[1]The University of Chicago, Chicago, Illinois, United States of America; [2]Stanford University, Stanford, California, United States of America

GWAS have been tremendously successful in identifying genetic variants that are robustly linked to complex diseases and traits. These variants are mostly located in non-coding regions of the genome making the interpretation more difficult. It has been shown that these trait-associated variants are enriched in classes of variants that affect gene expression levels, RNA splicing, and chromatin structure among others. These enrichment studies give us important clues to the underlying biology. However many of these functional classes are defined based on ad-hoc P value threshold and do not take into account the effect size of the variants. Thus the level of uncertainty in classification confounds the assessment of relative contribution of different classes. Here we propose a method that uses variance component with error in variables analysis to quantify the contribution of each functional class to phenotypic variability. We show that ignoring the uncertainty in QTL classification and effect sizes can lead to underestimation of the contribution. Relative contribution of eQTL versus splicing QTLs will be shown for a number of complex phenotypes.

41 | Estimation of Treatment Effects in Genotype Subgroups Following an Allelic Association in Randomized Controlled Trials

Amina Barhdadi[1], Marie-Pierre Dubé[1,2]

[1]Beaulieu-Saucier Université de Montréal Pharmacogenomics Centre, Montreal Heart Institute, Montreal, Canada; [2]Faculty of Medicine, Université de Montréal, Canada

In the conduct of genome-wide studies using data from Randomized Controlled Trials (RCT), it is typical to assess genetic association with study outcomes without informing directly as to whether the genetic marker can classify patients into subgroups with statistically significant clinical differences. In order to provide insight into the clinical utility of a pharmacogenomics association, here, we relate the allelic effect size to the comparative treatment effect within genotype subgroups.

We first assume that the outcome of interest is binary (0/1) and that the genetic marker has three genotypes, for a RCT with an overall treatment effect (OR_T), an overall allelic effect (OR_A), an allelic effect in each study arm (OR_{A1}) and (OR_{A2}) and a set Minor Allele Frequency (MAF).

The comparative treatment effects within each genotype subgroups OR_{T1}, OR_{T2} and OR_{T3} can be estimated for an appropriately designed RCT following an allelic association study performed in each treatment arm separately.

Here, we present the mathematical basis for estimating OR_{T1}, OR_{T2} and OR_{T3}. Our method is based on logistic regression which models the probability of events in each study arm as a function of MAF, OR_{A1}, OR_{A2}, and the risk in the absence of a genetic effect.

We have applied this method to a pharmacogenomics study where the OR of an associated SNP was equal to 0.63 (p=4.1×10^{-8}) in the treatment arm and to 0.91(p=0.25) in the placebo arm. The comparative treatment effects were calculated to be OR_{T1}=0.60, OR_{T2}=0.93 and OR_{T3}=1.29.

42 | From One Family to Replication in Five Data Sets: Two Loci Associated with Age-at-Onset of Familial and Sporadic Alzheimer Disease

Elizabeth E. Blue[1], Ellen M. Wijsman[1,2,3], Thomas D. Bird[4,5,6], Chang-En Yu[7], Timothy Thornton[2]

[1]Division of Medical Genetics, University of Washington, Seattle, Washington, United States of America; [2]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [3]Department of Genome Science, University of Washington, Seattle, Washington, United States of America; [4]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [5]Department Neurology, University of Washington, Seattle, Washington, United States of America; [6]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, United States of America; [7]Division of Gerontology and Geriatric Medicine, University of Washington, Seattle, Washington, United States of America

We fine-mapped a linkage region in one family, evaluated sequence variants within the region, and replicated association at 2 loci in 5 data sets. Beginning with strong linkage signals for Age-At-Onset (AAO) modifiers in a German from Russia (GFR) family with early-onset Alzheimer Disease (AD) caused by the *PSEN2* N141I variant, we combined a Measured Genotype (MG) approach with identity-by-state and -descent estimates to shrink the linkage region to one tenth the original size. We combined exome and genome sequences, applied bioinformatics and identity-by-state filters, and the MG approach to derive candidate variants. We tested for association between Cox proportional hazards residuals with AAO of AD adjusted for both *APOE* and *PSEN2* and the candidate variants in subjects with GFR ancestry ascertained from late-onset AD and *PSEN2* families, finding 13 variants nominally significant in the discovery family and in the larger GFR sample. Similar association tests at these loci (±25kb) in exome variants observed in non-Hispanic European subjects from the Alzheimer's Disease Sequencing Project found highly significant ($p<10^{-9}$) association surrounding the GFR variants in a promoter of *NCSTN* (chr1q23.2) and *KDM6B* (chr17p13.1). Association testing in 5 SNP-based data sets including subjects with early-and late-onset AD, with and without family history, and European or Hispanic ancestry found nominally significant evidence for association between AAO of AD and SNPs surrounding *NCSTN* and *KDM6B*. We therefore show that loci associated with phenotypic variation in a Mendelian early-onset form of AD have similar effects in cases with sporadic or familial AD across populations.

43 | Correlation of Psychosis, Mania and Depression Symptom Dimensions with Polygenic Scores in the Eastern Quebec Kindred Study

Sébastien Boies[1,2], Chantal Mérette[1,4], Yvon C. Chagnon[1,4], Michel Maziade[1,4], Alexandre Bureau[1,3]

[1]Centre de recherche de l'Institut universitaire en santé mentale de Québec du Centre intégré universitaire en santé et services sociaux de Québec, Québec, Canada; [2]Département de mathématiques et de statistique, Université Laval, Québec, Canada; [3]Département de médecine sociale et préventive, Université Laval, Québec, Canada; [4]Département de psychiatrie et neurosciences, Université Laval, Québec, Canada

The Psychiatric Genomics Consortium (PGC) results on SNPs association with Schizophrenia (SZ) and Bipolar Disorder (BD) can be used to calculate polygenic scores. There is an interest in studying the possible relation of polygenic scores with psychosis, mania and depression symptom dimensions for subjects affected by SZ, BD and Schizoaffective Disorder (SZA).

Our objective was to study the correlation between polygenic scores and the symptom dimensions in subjects affected by SZ, SZA and BD from the Eastern Quebec kindred study.

We used a sample of 333 subjects, including 153 affected (57 SZ, 13 SZA and 83 BD), from 17 SZ and BD families in the Eastern Quebec population. Polygenic scores were calculated based on the PGC SNPs association with SZ using P value cut-offs of 0.05 and 0.1. Symptoms were rated according to the Comprehensive Assessment of Symptoms and History (CASH) instrument on a scale from 0 to 5 (0: None, 5: Severe). We estimated heritability of normalized symptom dimensions explained by polygenic scores using linear mixed models with random additive genetic effects.

Affected subjects had higher SZ polygenic scores than non-affected subjects (p=0.049, Nagelkerke R^2= 0.016). There was no statistical difference between SZ and BD. We found significant association between SZ polygenic scores and thought disorder in stabilized and acute states (P values of 0.010 and 0.012 and proportion of explained heritability of 30% and 65% respectively). No other symptoms were associated with SZ polygenic scores.

SZ polygenic scores seem to be more specifically correlated with a symptom dimension associated to SZ.

44 | Prenatal Exposure to Disinfection by-Products, Cytochrome P450 Gene Polymorphisms and Risk of Intra-Uterine Growth Restriction

Samuella G. Bonou[1], Patrick Levallois[1,2], Yves Giguère[3], Manuel Rodriguez[4], Alexandre Bureau[1,5]

[1]Département de médecine sociale et préventive, faculté de médecine, Université Laval, Québec City, Canada; [2]Direction de la santé environnement et de toxicologie, Institut National de Santé Publique du Québec and Axe Santé des populations et pratiques optimales en santé,

Centre de recherche du CHU de Québec, Québec City, Canada; [3]Département de biologie moléculaire, de biochimie médicale et de pathologie, Faculté de médecine, Université Laval, Québec City, Canada; [4]École supérieure d'aménagement du territoire and Chaire d'eau potable, Université Laval, Québec City, Canada; [5]Axe de recherche neurosciences cliniques et cognitives, Institut universitaire en santé mentale de Québec, Centre intégré de santé et de services sociaux de la Capitale-Nationale, Québec City, Canada

Prenatal exposure to Chlorination By-Products (CBPs) in drinking water has been suggested as a possible etiologic factor for Intra-Uterine Growth Restriction (IUGR). Few epidemiological studies investigated the role of genetic variability in modulating CBP effects on fetal growth. To date attention was particularly given to cytochrome P450 (*CYP2E1*) and Glutathione-S-Transferases (*GSTM1, GSTT1*) xenobiotic-metabolizing genes. However, *CYP1A2*, *CYP2A6*, *CYP2D6* and *CYP17* genes, coding for cytochrome P450 enzymes are potential candidates for gene-environment interactions, but have never been studied in regard to CBP exposure and IUGR risk.

This study evaluated the interaction between prenatal exposure to Trihalomethanes (THMs) or Haloacetic Acids (HAAs) measured by gas chromatography/mass spectrometry or electron capture and genetic polymorphisms of mothers and foetuses on Small for Gestational Age (SGA) neonates by investigating SNPs in the *CYP1A2, CYP2A6, CYP2D6* and *CYP17* genes.

A population-based case-control study of 1438 mothers-children pairs was conducted. DNA was extracted from blood and saliva cells and Sequenom Technology was used for SNPs genotyping. Statistical interaction was evaluated by unconditional logistic regression with control of potential confounders.

Positive interactions were observed between chloroform, dichloroacetic and trichloroacetic acids and neonates *CYP17* rs4919687 A or rs743572 G alleles whereas chloroform interaction with mothers *CYP17* rs4919687 A allele was found protective. However after correction for multiple testing, reported interactions became non-statistically significant. No interactions were found with *CYP1A2, CYP2A6* and *CYP2D6* genotyped SNPs.

This study suggests that mother/child *CYP17* rs4919687 and rs743572 (SNPs) interaction with DPB exposure may influence the risk of IUGR.

45 | Genetic Analysis of the Telomere Interactome Pinpoints new Candidate Genes for Melanoma Risk

Myriam Brossard[1,2], Amaury Vaysse[1,2], Hamida Mohamdi[1,2], Yuanlong Liu[1,2], Eve Maubec[1,2,3], Pilar Galan[4], Marie-Françoise Avril[5], Mark Lathrop[6], Florence Demenais[1,2]

[1]INSERM, UMR-946, Paris, France; [2]Université Paris Diderot, Paris, France; [3]AP-HP, Hôpital Bichat, Service de Dermatologie, Paris, France; [4]INSERM, UMR U557; Conservatoire national des arts et métiers, Centre de Recherche en Nutrition Humaine, Ile de France, Bobigny, France; [5]AP-HP, Hôpital Cochin et Université Paris Descartes, Paris, France; [6]McGill University, Montreal, Canada

There is increasing evidence that telomere biology plays a key role in the development of Cutaneous Melanoma (CM). Common variants and rare mutations in eight genes related to telomere maintenance have been associated with CM risk. Four of these genes (*ACD*, *POT1*, *TERF1*, *TERF2IP*) encode proteins of the Shelterin Complex (SC) that includes two additional proteins (TERF2, TINF2). SC is crucial for the maintenance of the telomere structure and its signaling functions. To identify new candidate genes for CM, we investigated the effects and pairwise interactions of 467 genes of the telomere interactome, that is made of the six SC genes and their interacting partners retrieved from BioGRID. The gene-level analysis was carried out using VEGAS2 and the cross gene SNP-SNP interaction analysis was based on logistic regression models. This analysis was conducted in the MELARISK study (3,976 subjects) that had 1000-Genomes imputed SNP across the genome. Gene-based analysis identified three genes meeting the multiple testing corrected threshold ($P_{Gene}\leq 10^{-4}$):2 in known loci for CM risk (*TERT*, *TUBB3)* and a new candidate gene, *ANKMY2* involved in the hedgehog signaling pathway. Epistasis analysis identified 97 SNP pairs showing suggestive interaction ($7\times10^{-6}\leq P_{inter}\leq 3\times10^{-5}$), although not meeting the multiple testing corrected threshold. These pairs involved *TERF1* with two genes (*HOXA3, RYR2*) and *TERF2* with three genes (*COTL1*, *RGMA*, *S100P*). Replications of these results in another large dataset are ongoing.

46 | Genetic Correlation of Lung Function with Anthropometric Measures in the Busselton Health Study

Gemma Cadby[1], Philip E Melton[1], Nina S McCarthy[1], Jennie Hui[2,3], John Beilby[2,3], AW (Bill) Musk[2,4,6], Alan L James[2,5,6], Joseph Hung[6,7], John Blangero[8], Eric K Moses[1]

[1]Centre for Genetic Origins of Health and Disease, Curtin University and The University of Western Australia, Crawley, Australia; [2]Busselton Population Medical Research Institute Inc., Perth, Australia; [3]PathWest Laboratory Medicine WA, J Block, QEII Medical Centre, Nedlands, Australia; [4]Department of Respiratory Medicine, Sir Charles Gairdner Hospital, Nedlands, Australia; [5]Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands, Australia; [6]School of Medicine and Pharmacology, The University of Western Australia, Crawley, Australia; [7]Department of Cardiovascular Medicine, Sir Charles Gairdner Hospital, Nedlands, Australia; [8]South Texas Diabetes and Obesity Institute, The University of Texas Rio Grande Valley, Brownsville, Texas, United States of America

Anthropometric measures genetically correlated with lung function may provide insight into lung function pathways and aetiology. We employed empirically derived Identity By Descent (IBD) measures to estimate the genetic correlation between lung function (Forced Expiratory Volume in one second (FEV_1) and Forced Vital Capacity (FVC)) and anthropometric measures (BMI, waist-hip ratio, weight and circumferences of the abdomen, upper arm, thigh, calf and neck). Subjects (n=4671) were taken from the Busselton Health Study. IBD estimates were derived from genome-wide data using LDAK and genetic correlations (r_g) were calculated in GCTA. All traits were adjusted by age and sex, with lung function also adjusted by smoking and height. For significantly associated genetic correlations, anthropometric measures were also adjusted by BMI to investigate whether these associations were independent of obesity. The false discovery rate was used to correct for multiple testing. The sample was 56% female, with an average age of 50.8 years (SD=17.3). Mean FEV_1 and FVC were 3.01L (SD=0.97) and 3.89L (SD=1.17), respectively. FEV_1 and FVC were both heritable (FEV_1=0.41 and FVC=0.40; $P<1.0\times10^{-5}$). FEV_1 was genetically correlated with abdominal circumference (r_g=−0.25; q=0.03). FVC was genetically correlated with circumferences of the abdomen (r_g=−0.34; q=0.001), thigh (r_g=−0.49; *q*=0.002), and calf (r_g=−0.32; *q*=0.001). These associations all remained statistically significant after adjusting for BMI. This study reveals that the genetic correlations between lung function and circumferences of the abdomen, thigh and calf cannot be explained by obesity alone and may be due to pleiotropic genetic effects.

47 | Single Nucleotide Polymorphisms (SNPs) Associated with Fasting Blood Glucose Trajectory and Type 2 Diabetes Incidence: A Joint Modelling Approach

Mickaël Canouil[1], Philippe Froguel[1,2], Ghislain Rocheleau[1]

[1] *Université Lille, CNRS, CHU Lille, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France;* [2] *Department of Genomics of Common Disease, Imperial College London, London, United Kingdom*

In observational cohorts, longitudinal data are collected with repeated measurements at predetermined time points for many biomarkers, along with other covariates measured at baseline. In these cohorts, time until a certain event of interest occurs is commonly reported and very often, a relationship will be observed between a biomarker repeatedly measured over time and that event. Joint models were designed to efficiently estimate statistical parameters by combining a mixed model for the longitudinal biomarker trajectory and a survival model for the event risk, using a set of random effects to account for the link between the two types of data.

First, using genotypes assayed with the MetaboChip DNA arrays (Illumina) from close to 4,500 subjects recruited in the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance), we assessed the feasibility of implementing the joint modelling approach in a real high-throughput genomic dataset. Second, we checked model consistency based on different simulation scenarios, varying sample size, minor allele frequency, number of repeated measurements and missing data patterns. In our study, the event of interest was onset of type 2 diabetes (T2D), and the longitudinal biomarker repeatedly measured over time was fasting plasma glucose level.

To the best of our knowledge, joint models have never been applied into a genetic epidemiology context and could help identify novel loci sharing effects on both glycaemic traits and T2D.

48 | Mendelian Randomization (MR) Predicts a Causal Role for Serum ACE, APOC-1, APOE, Clusterin, and GDF-15 in Alzheimer Disease (AD)

Deanna Alexis Carere[1,2,3], Jennifer Sjaarda[1,2,4], Sibylle Hess[5], Hertzel Gerstein[1,6,7], Guillaume Paré[1,2,3]

[1] *Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada;* [2] *Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada;* [3] *Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada;* [4] *Department of Medical Sciences, McMaster University, Hamilton, Canada;* [5] *Sanofi Aventis Deutschland GmbH R&D Division Diabetes, Frankfurt, Germany;* [6] *Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada;* [7] *Department of Medicine, McMaster University, Hamilton, Canada*

Dozens of blood-based biomarkers are associated with Alzheimer Disease (AD), but their causal relationship to disease (if any) is unknown. We used Mendelian Randomization (MR) to identify blood serum biomarkers with a causal role in AD. Genotypes and baseline serum biomarker measurements were available from 5,078 Outcome Reduction with Initial Glargine Intervention (ORIGIN) trial participants. Summary statistics from the largest GWAS of AD were obtained from the International Genomics of Alzheimer's Project (IGAP). MR analyses were restricted to biomarkers directly encoded by an autosomal gene and SNPs within 300Kb of a biomarker's corresponding gene. 1,731 non-redundant SNPs associated with 203 biomarkers were available for MR. After Bonferroni correction (α=0.05/203), MR predicted a deleterious effect of apolipoprotein C-1 (APOC-1; OR=1.89 (1.65-2.16)) and growth differentiation factor 15 (GDF-15; OR=1.24 (1.14-1.35)); and a protective effect of angiotensin-converting enzyme (ACE; OR=0.91 (0.87-0.95)), apolipoprotein E (APOE; OR=0.33 (0.32-0.35)), and clusterin (OR=0.67 (0.56-0.80)). The independent effects of APOE and APOC-1 (encoded by proximal genes) persisted in a joint analysis, and when the *APOE* SNPs which determine epsilon allele status (rs7412, rs429358, and all SNPs in LD with these) were removed from analysis. Mendelian Randomization (MR) suggests a causal role for serum ACE, APOC-1, APOE, clusterin, and GDF-15 concentration in AD. The risk effects of increased APOC-1 and decreased APOE concentration appear to be independent of each other, and of *APOE* epsilon allele status. Our study represents the first investigation of multiple blood biomarkers for a causal role in AD.

49 | RVMMAT: Rare-Variant Mixed Model Association Tests for Binary Traits in Structured and Related Samples

Han Chen[1], Xihong Lin[1]

[1] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America*

We develop the computationally efficient Rare-Variant Mixed Model Association Tests (RVMMAT) for binary traits in structured and related samples. With the advance in next-generation sequencing technology, statistical methods for testing genetic association with rare variants have been proposed and widely applied to unrelated samples. These methods are also known as gene-based or SNP-set tests, because rare variants are often grouped by genes or genomic regions in the analysis. The burden test and Sequence Kernel Association Test (SKAT) are two widely applied rare variant tests. Here we propose and implement the burden test, SKAT and a novel combined test in RVMMAT. All tests share the same null generalized linear mixed model, which only needs to be fitted once in a whole-genome analysis. We show in simulation studies that the proposed tests control correct type I error rates in the presence of population stratification and cryptic relatedness, in both single-cohort studies and meta-analysis. We compare the power of these tests in various scenarios and illustrate how they can be used to test a broad class of different scientific hypotheses in large-scale sequencing studies. We also apply RVMMAT to a real data whole-exome analysis.

50 | Bayesian Model Averaging Approach for the X-Inactivation Dilemma in Genetic Association Studies

Bo Chen[1], Radu V. Craiu[1], Lei Sun[1]

[1] *Department of Statistical Sciences, University of Toronto, Toronto, Canada*

Due to many analytical challenges, the X-chromosome is often excluded from the whole-genome genetic association studies. One of these challenges is the dosage compensation issue where one of the two female X-chromosomes may or may not be inactivated. In the absence of evidence in favor of one specific model, we consider a Bayesian model averaging framework that provides a principled way to account for model uncertainty. Given an established association, we then use data-based evidence for model selection via Bayes factors. We examine the inferential properties of the proposed methods and compare them with frequentist solutions in both simulation and application studies.

51 | Compositional Epistasis Detection using a Few Prototype Disease Models

Lu Cheng[1], Mu Zhu[1]

[1] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada*

A limited amount of work is devoted to detecting complete compositional epistasis that bares the original meaning of "masking effect" when the term “epistasis” was coined. It is straightforward to model such type of epistasis by the 512 two-locus, two-allele, two-phenotype and complete-penetrance disease models (Li et al 1999). To achieve high detection power and testing efficiency, it is imperative to determine proper disease models for given SNPs so that the right compositional epistasis is captured.

Observing that the disease models are similar to each other, we define a novel “distance” metric to measure how different two disease models are and then use it to group disease models into a few clusters. We find that the 512 disease models form 6 clusters most of the time, and a prototype disease model selected from each cluster serves as a good representative model that can be used for epistasis loci detection. It is worth mentioning clustering epistasis models this way is not only beneficial to the computational and multiple-testing problems, but also allows us to better understand and characterize different disease models for future research.

By carrying out simulation studies on some popular epistasis mechanism, we observe that our approach provides satisfying power when compared with two other most relevant methods, i.e., MDR and the complete compositional epistasis detection approach by Wan et al^{2013}. Motivated by the differences of these methods, we also propose more refined ways to determine prototype disease models so that detection power is further improved.

52 | Identification of Genetic Modifiers that Protect Memory in Puerto Rican PSEN1 Mutation Carriers

Rong Cheng[1], Badri Vardarajan[1,2], Rafael Lantigua[3], Dolly Reyes-Dumeyer[1], Angel Piriz[1], Martin Medrano[4], Ivonne Z Jimenez-Velazquez[5], Richard Mayeux[1,2], Joseph H Lee[1]

[1] *Sergievsky Center/Taub Institute, CUMC, New York, New York, United States of America;* [2] *Department of Neurology, CUMC, New York, New York, United States of America;* [3] *Department of Medicine, CUMC, New York, New York, United States of America;* [4] *School of Medicine, Pontificia Universidad Catolica Madre y Maestra, Santiago, Dominican Republic;* [5] *Department of Internal Medicine, University of Puerto Rico School of Medicine, San Juan, Puerto Rico*

To identify protective genetic variants that may influence memory performance, we studied 45 families with the G206A founder mutation in PSEN1. This study extends our earlier study which identified genetic modifiers that may delay or

hasten the Age At Onset (AAO) of Alzheimer Disease (AD) in mutation carriers. We reasoned that the variants that delay the AAO would slow down the decline of memory in carriers of the PSEN1 mutation.

We conducted a joint Whole Genome Sequencing (WGS) and GWAS. All members had GWAS data, and, for each family, 1 to 10 members had WGS data. For individuals with GWAS data, we imputed WGS data into GWAS data using SHAPEIT2 and IMPUTE2. We then performed family-based allelic association analysis using a linear mixed model to assess how variants in candidate genes were associated with memory.

For this purpose, we focused on the genes that were associated with the AAO of AD in our earlier study. These genes include MLF1IP, SNX25, PDLIM3, and SORBS2 on 4q35 and SH3RF3 and NPHP1 on 2q13. For these 6 genes, 29 variants were associated memory at p<0.001. Two most significant variants were rs4861681 (p=1.5E−4) on SORBS2 and Chr2:109832517 (p=7.4E−5) on SH3RF3. rs71593655 was associated with better memory performance, while Chr2:109832517 was associated with lower memory performance.

This study shows that the variants that were associated with delayed or hastened AAO were clinically associated with enhanced or lowered memory performance in mutation carriers, suggesting these genes may be potential targets for therapeutics.

53 | Working Together: Genetic Risk and Environmental Contributors to Adolescent Adiposity

Hao Yu Chen[1,2], Marie-Pierre Sylvestre[3,5], Erika Dugas[3], Igor Karp[4,5], Line Dufresne[2], Katia Desbiens[2], George Thanassoulis[1,2], Jennifer O'Loughlin[3,5], James Engert[1,2]

[1]Division of Experimental Medicine, McGill University, Montreal, Canada; [2]Research Institute of the McGill University Health Centre, Montreal, Canada; [3]Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montreal, Canada; [4]Department of Epidemiology and Biostatistics, Western University, London, Canada; [5]Department of Social and Preventative Medicine, University of Montreal, Montreal, Canada

GWAS in adults have identified many loci associated with BMI, although relatively few studies have investigated these associations despite adolescent obesity predicting adult cardiovascular risk. We and others have shown that BMI is associated with blood pressure in adolescence, underscoring the importance of increasing our understanding of the determinants of adiposity in adolescents. Using DNA samples from the Nicotine Dependence in Teens (NDIT) cohort (n=674), we genotyped 40 GWAS-significant SNPs known to be associated with BMI in adults, and constructed a Genetic Risk Score (GRS). The association between the GRS and BMI (kg/m^2) was analyzed cross-sectionally in four age groups (mean ages of 12.7, 15.1, 16.9, and 24.0 years) and longitudinally in multivariable models that included age, sex, Physical Activity (PA) and diet. Cross-sectionally but not longitudinally, higher PA was associated with lower BMI in the youngest age group only. Consumption of unhealthy foods was associated with a higher BMI in the youngest age group cross-sectionally and longitudinally (β=0.81, p=0.020). The GRS was associated with BMI in all age groups, even after omitting the fat mass and obesity associated (*FTO*) locus SNP. Longitudinally, the GRS was associated with an increased BMI (0.15 units per risk allele, p=1.35×10^{-5}). This is greater than the 0.10 BMI units per allele observed in the 2015 GIANT Consortium study. Our findings in conjunction with those of GIANT suggest that the genetic contribution to BMI may decline across the lifespan.

54 | Improvements in Genotype Imputation by using a Population Specific Reference Panel in Africa

Ji Chen[1], Meng Sun[2], Tommy Carstensen[1], Andrew Morris[2], Fraser Pirie[3], Ayesha Motala[3], Manj Sandhu[1], Mark McCarthy[2], InêsBarroso[1], Eleanor Wheeler[1], Anubha Mahajan[2],

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom; [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [3]University of KwaZulu-Natal, Nelson R Mandela School of Medicine, 419 Umbilo Road, Congella, Durban 4001, South Africa

Genotype imputation is an important step to boost power in GWAS. As discovery efforts in non-European ancestry populations increase, it has become popular to supplement existing imputation panels with population specific reference panels. In our study, we conducted a GWAS of Type 2 Diabetes (T2D) in African subjects of Zulu descent (1,602 cases/976 controls), and compared the imputation quality of two different imputation schemes. In scheme 1, we utilized the 1,000 genomes phase 3 panel (Panel 1) with uniformly distributed 1Mb chunks. Panel 1 was merged with 2,298 African samples from the African Genome Variation Project (AGV) and the Uganda 2,000 Genomes Project (UG2G) into 1000Gp3+AGV+UG2G panel (Panel 2). In scheme 2, we used panel 2 with chunks containing 200 more variants in both the genotyped data and panel 2, avoiding large gaps and shorter than 5Mb. The imputation quality was evaluated in three aspects: number of variants imputed, imputation accuracy and imputation information score. Imputation accuracy was assessed using 97 Zulu samples with both genotype and sequence data. The sequence data was treated as the truth set and four metrics (genotype concordance, non-reference sensitivity, non-reference genotype concordance and precision) were applied to evaluate the similarity between the genotype data and the sequence data. As expected, in our data, scheme 2 both increased the number of imputed variants (20,262,808 African specific bi-allelic variants were only imputed in scheme 2) and improved the imputation accuracy and information score in all chromosomes.

55 | A New Statistical Method for Polygenic Risk Modelling to Incorporate LD and Functional Information of SNPs using GWAS Summary-Level Data

Ting-Huei Chen[1], Jianxin Shi[2], Nilanjan Chatterjee[3]

[1]Department mathematics and statistics, Laval University, Quebec, Canada; [2]Biostatistics Branch, NCI/DCEG, Bethesda, Maryland, United States of America; [3]Departments Biostatistics and Medicine at the Johns Hopkins University Bloomberg School of Public Health and Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

Polygenic Risk Score (PRS) has been a popular tool for genetic risk prediction of complex diseases. It can be constructed based on the summary statistics from GWAS. Several methods have been proposed to improve the performances of PRS. The standard version is based on a set of independent SNPs with *P* values less than a predefined significance level. Moreover, winner's-curse adjustments and external functional/annotation on a set of independent SNPs have been proposed to improve the performances of PRS. Another type of modification is to incorporate LD information based on Bayesian modelling technique, but without the usage of the functional knowledge. We propose a new method to incorporate both LD and functional annotation based on a penalized estimation approach. We applied our methods to GWAS summary-level data of 6 traits including height, BMI and complex diseases. Overall, the incorporation of LD enhances the prediction efficiency as 10–75% augmentation in prediction R squares compared to the standard PRS. For instance, the improvements of BMI and type 2 diabetes are from 4.3% to 6.6% and 2.1% to 3.6% respectively. The additional information of functional annotation of SNPs has only benefited certain traits such as type 2 diabetes to 4.2%. Extensive simulation studies have been conducted to illustrate the performances of the proposed method. In addition, we provide the prediction R squares obtained from other competing methods in both real data and simulation studies and the results demonstrate that our proposed method has significant advantages.

56 | A comparison of Genetic Risk Prediction Approaches in the Presence of Heterogeneity

Wenan Chen[1], Gregory Jenkins[2], Joanna M. Biernacka[2]

[1]St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America; [2]Mayo Clinic, Rochester, Minnesota, United States of America

Many complex diseases, particularly psychiatric traits, are highly heterogeneous, and it is believed that diagnostic criteria used to define these diseases aggregate numerous genetically distinct subtypes. This aggregation of subtypes results in increased polygenicity, attenuated effect sizes of individual genetic variants, and consequently diminished power to detect genetic associations. This study investigated the effects of genetic heterogeneity on the performance of risk prediction based on genetic data. We considered the widely used additive polygenic risk score approach, two penalized regression methods that incorporate variable selection, LASSO and elastic net, and the machine learning methods random forests and gradient boosting machines. Simulations were performed by generating training and test datasets with binary phenotypes derived from one or more polygenic liabilities; the training datasets were used to generate predictive models, while the test datasets were used to assess the predictive performance of the resulting models using the area under the receiver operating characteristic curve. As expected, the simulations demonstrated that performance of all approaches declines with increasing genetic heterogeneity and increasing total number of SNPs, and improves with increasing sample size. Overall, elastic net performed best across the simulated scenarios, without the need to pre-filter SNPs for inclusion in the model building, as the approach incorporates variable selection. This study demonstrates the rapid decline in predictive performance of risk prediction methods when the underlying causal model departs from simple additive SNP effects – new methods are needed to optimally model risk of complex genetic traits.

57 | Association Tests of Multiple Genetic Variants for Time-to-Event Traits

Yen-Feng Chiu[1], Li-Chu Chien[2], Donald W. Bowden[3,4,5]

[1]Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan, ROC; [2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC; [3]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America; [4]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America; [5]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America

Identification of functional variants is one of the crucial steps to dissect the genetic mechanism of complex diseases. Family-based designs enriched with affected subjects and disease associated variants can increase statistical power for identifying functional rare variants. However, few rare variant analysis approaches are available for time-to-event traits in family designs. We developed novel pedigree-based burden and kernel association tests for time-to-event outcomes with right censoring for pedigree data, referred to FamRATS (Family-based Rare variant Association Tests for Survival traits). Cox proportional hazard models were employed to relate a time-to-event trait with rare variants (MAF<0.05) with flexibility to encompass all ranges and collapsing of multiple variants. Robustness of violating proportional hazard assumptions was investigated for the proposed and four current existing tests including the conventional population-based Cox proportional model and the burden, kernel and sum of squares statistic (SSQ) tests for family data. The proposed tests can be applied to large-scale whole-genome sequencing data. They are appropriate for the practical use under a wide range of misspecified Cox models, as well as for population-based, pedigree-based or hybrid designs. In our extensive simulation study and data example, we showed that the proposed kernel test is the most powerful and robust choice among the proposed burden test and the existing four rare variant survival association tests. When applied to the Diabetes Heart Study, the proposed tests identified exome variants of the *JAK1* gene on chromosome 1 to be associated with age at onset of type 2 diabetes ($P<4.82\times10^{-5}$).

58 | Dynamic Prediction of Colorectal Cancer Risks in Lynch Syndrome (LS) Families Accounting for Screening Information and Family History

Yun-hee Choi[1], Helene Jacqmin-Gadda[2,3], Agnieszka Krol[2], Laurent Briollais[4,5], Virginie Rondeau[2,3]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Canada; [2]Equipe de biostatistique, ISPED, Université de Bordeaux, Bordeaux, France; [3]INSERM, Centre INSERM U897-Epidémiologie-Biostatisque, Bordeaux, France; [4]Lunenfeld-tanenbaum research institute, Mt. Sinai Hospital, Toronto, Canada; [5]Biostatistics, Dalla Lana School of Public Health, University of Toronto, Canada

Lynch Syndrome (LS) is a familial genetic disorder caused by mutations in DNA mismatch repair genes–MLH1, MSH2, MSH6. The screening recommendation for mutation carriers is to perform colonoscopies every 1–2 years for early detection and removal of polyps thus decrease the risk of subsequent Colorectal Cancers (CRCs).

A complex issue in genetic counselling of LS individuals is predicting the risk of developing CRC within a time interval given individual's disease, screening and family history. Another important challenge is to recommend optimal screening intervals for those LS individuals to reduce substantially CRC risks.

To address these problems, we propose a joint nested frailty model for the screening visit and cancer processes by accounting for two sources of dependence arising at family and individual levels. The association between the two processes is explained by shared family- and individual-specific frailties. In the joint modelling framework, we provide dynamic prediction of cancer risks based on the familial disease and screening history information.

We apply this new joint model to two series of LS families: one from Newfoundland and the other from Toronto. We provide estimates for the effects of covariates (i.e., age, gender, mutation type, polyp/cancer detection, proband's age) on the two processes and their association and predict personalized cancer risks of LS family members given their family history of screening and disease. We also evaluate the effectiveness of colonoscopy to decrease CRC risks in LS families and determine the optimal length and frequency of screening intervals via dynamic predictions and simulation studies.

59 | Using LASSO Regression to Identify Gene-Gene and Gene-Environment Interactions Influencing Cognitive Function in those with Increased Alzheimer's Risk

Burcu F. Darst[1], Murat Bilgel[2,3], Rebecca L. Koscik[4], Bruce P. Hermann[4,5,6], Bruno M. Jedynak[7], Sterling C. Johnson[4,5,8], Corinne D. Engelman[1,4,8]

[1]Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [2]Department of Biomedical Engineering, Johns Hopkins University School of Engineering, Baltimore, Maryland, United States of America; [3]Laboratory of Behavioral Neuroscience, National Institute on Aging, NIH, Baltimore, Maryland, United States of America; [4]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [5]Geriatric Research Education and Clinical Center, William. S. Middleton Memorial Veterans Administration Hospital, Madison, Wisconsin, United States of America; [6]Department of Neurology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [7]Department of Mathematics and Statistics, Portland State University, Portland, Oregon, United States of America; [8]Alzheimer's Diseases Research Center, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America

Although the heritability of Alzheimer's Disease (AD) is estimated to be 58%, most known genetic variants only account for 8% of the phenotypic variance of AD. Expanding investigations to include complex nonlinear relationships between genetic and environmental factors may explain some of this missing heritability. In this analysis, we explore gene-gene and gene-environment interactions that may be influencing cognitive function using participants from the Wisconsin Registry for Alzheimer's Prevention, a longitudinal family-based cohort of initially cognitively healthy adults (N=1,293) enriched for a parental history of AD. This analysis included 19 genetic variants associated with AD by the International Genomics of Alzheimer's Project, *APOE* $\varepsilon2$ and $\varepsilon4$, and eight environmental factors associated with AD: age, gender, education, physical activity, waist-hip ratio, total cholesterol, insulin resistance, and inflammation as measured by interleukin 6 (IL6). We performed variable selection using least absolute shrinkage and selection operator (LASSO) regression, inputting all possible pairwise interaction terms. One interaction identified by LASSO was statistically significant in a mixed linear model and evidenced in three of six other independent cognitive measures (all with $P<=0.004$). This interaction between rs6656401 (*CR1*) and rs9271192 (*HLA-DRB1*) suggested that being a carrier of either of these variants was associated with poorer cognitive function, whereas being a carrier of both of these variants was associated with better cognitive function, displaying a collective protective effect. Both of these genes are in the immune system pathway and associated with several autoimmune diseases. These findings support the early involvement of the immune system in AD.

60 | Multivariate Analysis in the Genomic Era: Back to the Future

Mariza de Andrade[1], Nubia E. Duarte[2], Julia M.P. Soler[2]

[1]Mayo Clinic, Rochester, Minnesota, United States of America; [2]University of São Paulo, São Paulo, Brazil

Multivariate analysis has been found in scientific literature for more than a century, where it has mostly been used to reduce a large number of variables to a smaller set of factors. In genomic settings the goal is to identify a pleiotropic effect which is a common genetic effect affecting more than one outcome. Several statistical methods have been proposed in the literature to identify genetic variants with pleiotropic effects in humans, animals and plants. In the era of linkage analysis, there was a surge of methods for human familial studies using mixed effect models with various estimation methods. In animal studies, family structure was ignored and different approaches were proposed to identify pleiotropic effects, the most common being regression models. During the Genome-Wide Association Studies (GWAS) era, the field has now moved back to the future, i.e., the overall goal is to identify pleiotropic effects because few causal variants have been identified for a single trait. So, today, the methods developed during the linkage analysis era are being used on GWAS data. Here, we will provide a description of the pleiotropic methods available in the literature, and as an application we will use two data sets using unrelated subjects from the GENOA study and related subjects from the Baependi Heart study, both with genotype data from Affymetrix 6.0 SNP chip.

61 | X-Inclusion: Analyzing X-Chromosome in Whole Genome Association Studies of Variance Heterogeneity

Wei Q. Deng[1], Lei Sun[1,2]

[1]Department of Statistical Sciences, University of Toronto, Toronto, Canada; [2]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Genetic variants associated with variance heterogeneity of quantitative traits are considered promising candidates to follow-up. Similar to genetic association studies of main effects, this approach to discover potentially interacting SNPs has so far been restricted to autosomes. Extension to X-chromosomal SNPs faces several analytical challenges including dosage compensation via X-inactivation and allelic heterogeneity. Focusing on Levene's test to detect phenotypic variance differences associated with the genotype groups of an X-chromosome SNP, we first note that sex can be an inherent confounder; sex necessarily correlates with the number of minor allele of the SNP, and it could also correlate, through both mean and variance, with the distribution of quantitative traits that exhibit sexual dimorphism, such as height and BMI. We then show theoretically that the naïve 3-group analytical strategy combining the genotypes of females and males could lead to spurious variance signals through confounding. We also provide results from extensive simulation and application studies to show that 1) the extent of confounding depends on the discrepancy between the two sex-stratified trait distributions, 2) a 5-group analytical approach accounting for the inherent sex-specific mean and variance effects is the most robust, and 3) there is much need for novel broad-sense heritability estimation methods that incorporate X-chromosome loci.

62 | Harnessing Electronic Medical Records Linked to DNA Biobanks in the Search for Biomarkers of Neuropsychiatric Disorders

Jessica Dennis[1], Donald Hucks[1], Guanhua Chen[2], Nancy Cox[1], Lea Davis[1]

[1]Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

The Vanderbilt University Medical Center biobank, BioVU, includes electronic medical records linked to genetic data in over 20,000 people, offering unprecedented opportunities to study relationships between genetic variants, biomarkers, and disease. Neuropsychiatric Disorders (NPD, e.g., anxiety disorders, schizophrenia) are notoriously intractable, and biomarkers for NPD could improve diagnostics, early interventions, and therapeutics. Multiple genetic variants across the genome contribute to the etiology of NPD, and our objective is to leverage the polygenicity of NPD to identify novel NPD biomarkers from among the ~200 clinical laboratory values commonly measured in BioVU (e.g., platelet volume, lipids). We will calculate Polygenic Risk Scores (PRS) for each genotyped individual in BioVU across each NPD, where the PRS is the sum of the number of risk alleles weighted by the effect size of each risk allele from a discovery GWAS (e.g., Psychiatric Genomics Consortium GWAS'). NPD-specific biomarkers will be the laboratory values that are correlated with an NPD PRS. In an initial proof-of-principle experiment, we showed that a PRS for coronary artery disease derived from the CARDIoGRAMplusC4D meta-analysis (7086 SNPs with $p<0.11$ in the discovery GWAS) was significantly associated with mean triglyceride values in 4593 BioVU subjects genotyped on the Illumina Metabochip platform ($p=4.7\times10^{-12}$) and accounted

for 1% of mean triglyceride variance while adjusting for sex, age, and ancestry. Future analyses will explore approaches for modelling repeated lab values, medication use, and disease status. Our results will be immediately relevant to NPD, and the approaches we develop will be easily extendible to any polygenic phenotype.

63 | Interactive Effect Between ATPase-Related Genes and Early-Life Tobacco Smoke Exposure on Bronchial Hyper-Responsiveness Detected in Asthmatic Families

Marie-Hélène Dizier[1,2], Patricia Margaritte-Jeannin[1,2], Lucile Pain[3], Marie-Claude Babron[1,2], Chloé Sarnowski[1,2], Myriam Brossard[1,2], Hamida Mohamdi[1,2], Nowlenn Lavielle[1,2], Jocelyne Just[4], Mark Lathrop[5], Emmanuelle Bouzigon[1,2], Catherine Laprise[3], Florence Demenais[1,2], Rachel Nadif,[6,7]

[1]INSERM, UMR-946, Paris, France; [2]Université Paris Diderot, Paris, France; [3]Université du Québec à Chicoutimi, Chicoutimi, Canada; [4]Centre de l'Asthme et des Allergies – INSERM, UMR-S 1136, Equipe EPAR, Paris, France; [5]Mc Gill University, Montreal, Canada; [6]INSERM, UMR-S 1168, Villejuif, France; [7]Université Versailles Saint-Quentin en Yvelines, Versailles, France

In a previous positional cloning study of the 17p11 region, we identified genetic variants interacting with tobacco smoke (ETS) exposure in early life for Bronchial Hyper-Responsiveness (BHR) in the French Epidemiological study on the Genetics and Environment of Asthma (EGEA). These variants were located in *DNAH9* (Dizier et al., 2016), a gene having a key role in motile cilia function.

Our objective was to identify other genetic variants interacting with ETS for BHR by investigating genes involved in 'ATPase binding' and 'ATPase activity' pathways. They both include *DNAH9*, are target of cigarette smoke and are implicated in the movement of respiratory cilia.

Family-Based Association Test (FBAT) analyses were first conducted in 388 EGEA families. We applied FBAT-homogeneity test between exposed vs. unexposed siblings to detect SNPxETS interaction for BHR. Replication was performed in 253 families from the Saguenay-Lac-Saint-Jean (SLSJ) asthma collection.

In EGEA families, 25 SNPs showed interaction signals ($P \leq 5 \times 10^{-3}$) with ETS among BHR siblings. One SNP on 4p13 reached the threshold ($P=2.10^{-5}$) for a significant interaction with ETS when correcting for multiple testing. This result did not quite reach the nominal level for replication in SLSJ families ($P=0.13$) but there was improvement of evidence for interaction in the meta-analysis of the two samples ($P=10^{-5}$). Another SNP on 9q31 showed a stronger interaction signal for replication in SLSJ families ($P=0.003$), and a suggestive interaction by meta-analysis ($P=6.10^{-5}$).

Further analyses based on log linear modeling and further replication in additional samples will be conducted to confirm these findings.

64 | Efficient High-Dimensional Disease Outcome Prediction in Heterogeneous Populations

Frank Dondelinger[1], Sach Mukherjee[2]

[1]Lancaster Medical School, Lancaster University, Lancaster, United Kingdom; [2]German Centre for Neurodegenerative Diseases, Bonn, Germany

The increasing availability of high-dimensional molecular data in biomedicine has led to a renewed interest in predicting disease outcomes based on a patient's genetic profile. This requires new statistical and machine learning methodology to build high-dimensional predictive models that can deal with the inherent heterogeneity of many diseases. For example, in the neurodegenerative disease Amyotrophic Lateral Sclerosis (ALS), it is well known that some patients exhibit rapid progression, while others progress very slowly. However, the causes underlying these differences are unknown.

We present a principled method for disease outcome prediction in heterogeneous settings, based on information sharing across penalized linear regression models. Our method is able to reflect the heterogeneity of the population, while at the same time leveraging the commonalities between groups of homogeneous samples. We present two related approaches, using l1 and l2 fusion penalties on the model-specific parameters, and show in extensive simulation studies that they outperform both naive pooling and complete separation of models. We then apply our method to three datasets:

1. gene expression and drug sensitivity data from The Cancer Cell Line Encyclopedia (Barretina et al. 2012),
2. clinical trial data from ProACT, a database of ALS patients, and
3. GWAS mutation data from the ADNI Alzheimer's database.

Using these datasets, we demonstrate that our method improves on other popular prediction approaches, as well as being highly computationally efficient. Additionally, we show that our model can be used to identify biomarkers for disease progression, and to identify common features between otherwise disparate population groups.

65 | Adaptive Bayesian Whole Genome Regression for Predicting Responsiveness to Treatment in Randomized Clinical Trials (RCTs)

Bahar Erar[1], George D. Papandonatos[1]

[1]Center for Statistical Sciences, Brown University School of Public Health, Providence, Rhode Island, United States of America

A critical aspect of personalized medicine is the development of methods that evaluate genetic predisposition to respond to a treatment. Studies have shown that a small number of genes that can be identified by one-at-a-time testing methods are shown to be inadequate for prediction modelling. Whole Genome Prediction (WGP) methods have been shown to improve predictive accuracy in complex traits. However, the methodology is not tailored for outcome prediction in human Randomized Clinical Trials (RCTs). We propose an adaptive Bayesian WGP (BWGP) method that accounts for the underlying genetic heterogeneity present in populations often targeted in RCTs using a mixed model approach. Under this model, small effects regulated by the treatment that would normally go undetected and disregarded are captured by the unconstrained covariance structure of the genetic random effects. Computational challenges that arise in the implementation of this multi-group model are addressed by employing an efficient estimation approach that allows application to large datasets at a reasonable computational cost. Predictive accuracy of adaptive BWGP is compared to other methods using simulated phenotypes generated from real genotypes under various scenarios. Results demonstrate that adaptive BWGP performs better or at least as well as stratified methods, such as BayesC, Bayesian Ridge Regression and Bayesian LASSO. The gain in prediction accuracy is highest for moderately and highly heritable traits under realistic effect regulation scenarios.

66 | Improvements in Efficiency and Power Associated with Joint Trait-Dependent and SNP-Dependent Sampling in Two-Phase Designs

Osvaldo Espin-Garcia[1,2], Radu V Craiu[3], Shelley B Bull[1,2]

[1] *Dalla Lana School of Public Health, University of Toronto, Toronto, Canada;* [2] *Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada;* [3] *Department of Statistical Sciences, University of Toronto, Toronto, Canada*

We evaluate two-phase designs to follow up GWAS findings when high sequencing costs preclude regional sequencing of an entire cohort. We propose a semiparametric maximum likelihood formulation where an identified GWAS SNP is an auxiliary covariate in inferring association between a sequence variant and a normally-distributed Quantitative Trait (QT). We apply an EM algorithm for estimation tailored to post-GWAS scenarios by considering SNP data available for all individuals in phase one, while inference on missing-by-design sequence variants is conducted using phase one and two data.

We perform simulations to quantify improvements of joint QT-SNP sampling compared to QT/SNP marginal sampling under alternative sample allocations, according to estimation efficiency, test validity and power. We observe improved efficiency and power of joint sampling with increasing Linkage Disequilibrium (LD) between GWAS-SNP and sequence variants across values of Minor Allele Frequencies (MAF). For phase two sample size=2500/5000 (overall sampling fraction=50%), SNP/sequence MAFs=0.3/0.2, LD r=0.75(D′=0.98), under combined allocation −balanced on GWAS-SNP genotype and extreme on QT value, joint sampling achieves 92% score test power (versus 97% in complete data) compared to 89% under QT- or SNP-based marginal sampling. Contrastingly, under proportional allocation, joint sampling achieves 86% power compared to 86% (84%) with marginal QT (or SNP) sampling. Our method doubles, on average, the empirical power obtained when analyzing phase-two data only under the same allocation.

Our approach can be extended to generalized linear models and can be applied in other instances where it is desirable to apply an expensive technology for a large existing cohort.

67 | Blood Lipoprotein Levels and Thrombin Generation Potential: Associations and Epigenetic Mediation

Karl Everett[1], Nora Zwingerman[1], France Gagnon[1]

[1] *Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*

Motivated by evidence that lipoproteins can influence pro- and anti-coagulant reactions, and the recent epigenome-wide association study that showed lipoprotein levels to influence blood DNA methylation (DNAm), we hypothesized that lipoprotein levels are associated with Endogenous Thrombin Potential (ETP – a venous thrombosis endophenotype), and that DNAm mediates these effects.

In 240 individuals from 5 pedigrees ascertained on single probands with venous thrombosis, we first tested for association between ETP and levels of low- and high-density lipoproteins (LDL and HDL). We then tested for replication the blood CpG sites recently found associated with LDL and HDL (3 and 4 sites respectively). The CpG sites were further tested for associations with ETP. Linear mixed effect models were adjusted for key covariates, and accounted for cellular heterogeneity and family structure.

A one Standard Deviation (SD) increase in LDL levels was associated with a 0.3 SD rise in ETP (95% CI: 0.16 − 0.46), while HDL was associated with a 0.2 SD reduction (95% CI: −0.35, −0.07). Recently-reported associations between HDL and DNAm at two CpG sites were replicated in our study sample (p=$10^{-7}$and 10^{-2}), while the LDL-CpG sites were not. DNAm levels at both HDL-associated CpG sites were associated with ETP. These ETP-associated CpG sites are located in introns of the cholesterol transport *ABCG1* gene.

Lipoprotein levels are associated with ETP, and preliminary findings suggest that these associations are mediated by DNAm. Replication and systematic Mendelian randomization analyses are underway to establish the directions of these relationships and infer causal effects.

68 | *SNAP-25* (rs1051312) Gene Might be Associated with Placebo Response in Children with ADHD

Weam Fageera[1,2], Natalie Grizenko[1,3], Sarojini M. Sengupta[1], Ridha Joober[1,2,3]

[1] *Douglas Mental Health University Institute, Montreal, Canada;*
[2] *Department of Human Genetics, McGill University, Montreal, Canada;*
[3] *Department of Psychiatry, McGill University, Montreal, Canada*

The idea that Placebo Response (PR) is not necessarily in an individual's head, but rather genes, has recently received more attention in the scientific community. Different neurotransmitters have been associated with PR and genetic variations in the brain's neurotransmitter pathways could modify placebo effects. Synaptosomal-associated protein 25 (SNAP25) is an essential component for synaptic vesicle mediated release of neurotransmitters. Therefore, genetic variations that might affect SNAP25 function have been suspected in the pathophysiology of ADHD. Making SNAP25 a good candidate for study its effect on PR. Here, we tested the association of rs1051312, a polymorphism in the 3′ UTR of SNAP25, with response to placebo in children with ADHD. 378 children with ADHD (6-12 years) participated in a randomized, double-blind, placebo-controlled crossover trial and genotyped for rs1051312 polymorphism. PR was calculated as the difference in Conners' (parents and teachers) scores at baseline and during placebo week. Repeated measures analysis of variance revealed a significant interaction between rs1051312 and response to placebo (p=0.022) according to teachers' assessments. Patients with C/C genotype were significantly more responsive to placebo. Patients with C/T and T/T genotypes barely showed any improvement after giving placebo. However, after administering MPH, all genotype groups tend to similarly respond to active medication. These findings provide evidence for the implication SNAP25 in placebo response. To the best of our knowledge, this is the first study to report an association between SNAP25 and placebo response.

69 | The Association Between *COMT* (Val158Met) and *DRD3* (Ser-9_Gly) GENOTYPES and METHYLPHENIDATE Side Effects

Weam Fageera[1,2], Natalie Grizenko[1,3], Sarojini M. Sengupta[1], Ridha Joober[1,2,3]

[1] *Douglas Mental Health University Institute, Montreal, Canada;*
[2] *Department of Human Genetics, McGill University, Montreal, Canada;*
[3] *Department of Psychiatry, McGill University, Montreal, Canada*

This study aims to investigate the potential role of two markers in *COMT* and *DRD3* genes on side effects related to stimulant medication. 285 children with ADHD (6-12 years) participated in a randomized, double-blind, placebo-controlled crossover trial and genotyped for *DRD3* (ser-9-gly) and *COMT* (Val158Met) polymorphisms. Parents were asked to complete the Stimulants Side Effect Scale (SERS) after the week of placebo and Methylphenidate (MPH). Principal Components Analysis (PCA) was performed to reduce the number of variables. PCA of SERS resulted in four factors: Emotionality, Somatic Complaints, Sleep Problems, and Socially Withdrawn. ANCOVAs analysis revealed significant differences between COMT*DRD3 genotypes on Somatic Complaints (i.e. 3-way interaction) (p=0.012). Two-way interactions were also observed (COMT* Somatic Complaints=0.002) (DRD3* Somatic Complaints=0.013). Children with the Val/Val-C/C genotypes displayed more sever placebo side-effect rating than all other groups. Interestingly, on MPH, this genotype group has the least sever stimulant side effects. This preliminary analysis found an association between two SNP markers (ser-9-gly & Val158Met) and the occurrence of Somatic Complaints (a group composed of stomachache, dizziness, and headache) as a side effect of short-acting methylphenidate. Replicating this analysis on a larger scale could potentially identify individuals with higher risk for MPH side effects.

70 | Using Behavioral Dynamic Approaches to Test for GENE-BY-GENE Interaction in Modulating ADHD Behaviors

Weam Fageera[1,2], Natalie Grizenko[1,3], Sarojini M. Sengupta[1,3], Ridha Joober[1,2,3]

[1] *Douglas Mental Health University Institute, Montreal, Canada;*
[2] *Department of Human Genetics, McGill University, Montreal, Canada;*
[3] *Department of Psychiatry, McGill University, Montreal, Canada*

Dopamine (DA) plays an important role in the pathogenesis of ADHD. Genes that regulate DA at consecutive junctures of its synaptic activity, such as *COMT* and *DRD3* may interact to modulate ADHD behaviors. Using pharmacological probes that affect synaptic DA concentration (such as methylphenidate and placebo) may increase our capacity to reveal the effect of these genes on ADHD behaviors.

We aim to test for *DRD3* (ser-9-gly) by *COMT* (Val158Met) gene/gene interaction in modulating ADHD behaviors using a pharmacological challenge with methylphenidate and placebo. 391 children with ADHD were included in this study. Parents and teachers were asked to evaluate child's behavior at baseline, placebo, and MPH weeks using the Conners' scale. Association between genotypes and ADHD behavior was tested using repeated measure ANOVA, the two genes were the between-subject factors and the behaviors under the three Experimental Conditions (EC), were the within-subject factor. A 3-way interaction (*DRD3***COMT**EC) was revealed with teachers assessment (p=0.001). *COMT**EC two-way interaction (but not *DRD3**EC) was also

significant; with the Met/Met genotype group having lower scores at baseline and on placebo, but the difference between groups disappeared on MPH. Remarkably, when children where stratified according to their *COMT* genotypes, statistically significant and biologically meaningful effects of the *DRD3* genotypes were detected. In conclusion, using a combination of methodological tools (pharmacological probes, large sample size, and selecting genes consecutively implicated in synaptic DA activity) might be essential to better understand the role of candidate genes in complex behaviors.

71 | A Genome-Wide Association Study of Pathological Inflammatory Responses in Leprosy

Vinicius M. Fava[1,2], Aurélie Cobat[1,2], Marianna Orlova[1], Jeremy Manry[1,2], Nguyen Van Thuc[3], Nguyen Ngoc Ba[3], Vu Hong Thai[3], Milton Moraes[4], Laurent Abel[5,6,7], Alexandre Alcaïs[5,6,7,8,9], Erwin Schurr[1,2]

[1]The McGill International TB Centre, The Research Institute of the McGill University Health Centre, Montreal, Canada; [2]Departments of Human Genetics and Medicine, McGill University, Montreal, Canada; [3]Hospital for Dermato-Venereology, Ho Chi Minh City, Vietnam; [4]Laboratory of Leprosy Research, Oswaldo Cruz foundation, Rio de Janeiro, Brazil; [5]Laboratoire de Génétique des Maladies Infectieuses, Institut National de la Santé et de la Recherche Médicale, U980, Paris, France; [6]Université Paris René Descartes, Faculté Médicine Necker, Paris, France; [7]Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, New York, United States of America; [8]URC-CIC, Hôpital Tarnier, Paris, France

Leprosy Type-1 Reactions (T1Rs) are pathological inflammatory responses characterized by sudden onset of cell-mediated immunity that result in host peripheral nerve damage. The etiology of T1R is uncertain but strong evidence supports a role of genetic susceptibility in T1R. We employed a family-based GWAS approach followed by stepwise replication in independent populations to identify novel genetic variants associated with T1R. In the discovery phase, a set of T1R-affected and a set of T1R-free families were compared. The GWAS followed by imputation resulted in approximate 6.3 million variant tested for association in both sets. Of those, only SNPs significant for T1R-affected families with significant evidence of heterogeneity relative to T1R-free families were considered T1R-specific. In the T1R-affected set we identified a region on chromosome 10p21.2 preferentially associated with T1R (p=8.2×10^{-7}). SNPs associated with T1R could be grouped in seven bins (r^2>0.9) and a tag SNP for each bin was selected for stepwise replication. A SNP bin tagged by rs1875147 was replicated and validated for association with T1R in two independent populations from Vietnam and Brazil, respectively (p<0.01 for both). Combined analysis of rs1875147 in all samples resulted in a stronger evidence of association with T1R (p=7.3×10^{-9}; OR=1.50, 95% CI=1.30-1.71). The SNPs associated with T1R located to a genome interval between two recombination hot spots were a single lncRNA was found. Our findings support an important role of lncRNAs in human inflammatory disorders.

72 | A Multi-Tissue Transcriptome-Wide Association Study of Breast Cancer Identifies 29 Novel Breast Cancer Susceptibility Loci

Helian Feng[1], Alexander Gusev[1], Bogdan Pasaniuc[2], Lang Wu[3], Jirong Long[3], Roger Milne[4], Doug Easton[5], Georgia Chenevix-Trench[6], Wei Zheng[3], Peter Kraft[1], On Behalf of the Breast Cancer Association Consortium

[1]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]University of California at Los Angeles, Los Angeles, United States of America; [3]Vanderbilt University School of Medicine, Nashville Tennessee, United States of America; [4]Cancer Council of Victoria, Melbourne Australia; [5]University of Cambridge, Cambridge United Kingdom; [6]QIMR Berghofer Medical Research Institute, Brisbane Australia

GWAS have identified over 180 loci associated with Breast Cancer (BrCa). However, these explain only 18% of the familial relative risk for BrCa, and the causal mechanisms at these loci remain largely unknown. To identify additional BrCa risk loci and identify likely causal genes at known loci, we conducted a Transcriptome-Wide Association Study (TWAS) of BrCa. We constructed genetic predictors of gene expression in over 40 tissues using Bayesian sparse linear mixed models and data from GTex, two large blood eQTL studies and one large adipose eQTL study (n=76 to 3,832). We then tested the association between predicted expression and BrCa risk using summary statistics from a GWAS meta-analysis of 119,000 cases and 101,000 controls. We identified 619 statistically significant associations between BrCa risk and predicted tissue-specific expression levels after adjusting for the number of tests conducted (p<0.05/35,026=1.4e-6). These associations involved 134 distinct transcripts, including 100 protein-coding genes and 19 lncRNAs. Twenty-nine (23%) of these transcripts were located more than 500kb of a known BrCa GWAS hit, representing potential novel risk loci. Eighteen of 683 tested transcripts in GTEx mammary tissue were associated with breast cancer risk, including the apoptosis and autophagy genes *CASP8* and *ATG10*. TWAS-significant genes were enriched for mutated genes in TCGA (p=0.0003), including *CASP8*, *AKT1* and *MAP3K1*. These analyses integrating GWAS summary data and eQTL information identified novel candidate breast-cancer genes and shed additional light on the causal mechanisms at known loci.

73 | *NOS1AP* Variant rs7539120 is Associated with Appropriate Implantable Cardioverter Defibrillators (ICD) Shock

Luisa Foco[1], Claudia B. Volpato[1], Massimiliano Manfrin[2], Anna Cima[3], Ilaria Bozzolan[1], Alessandro De Grandi[1], Peter P.

Pramstaller[1,4,5], Cosetta Minelli[6], Massimiliano Marini[7], Roberto Cemin[2]

[1] *Center for Biomedicine, European Academy Bozen/Bolzano (EURAC), affiliated to the University of Lübeck, Bolzano, Italy;* [2] *Department of Cardiology, San Maurizio Regional Hospital of Bolzano, Bolzano, Italy;* [3] *Department of Cardiology, S. Maria del Carmine, Rovereto, Italy;* [4] *Department of Neurology, General Central Hospital, Bolzano, Italy;* [5] *Department of Neurology, University of Lübeck, Lübeck, Germany;* [6] *Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College, London, UK;* [7] *Department of Cardiology, Santa Chiara Hospital, Trento, Italy*

Implantable Cardioverter Defibrillators (ICD) are used to prevent Sudden Cardiac Death (SCD) in patients with heart failure, but currently only 30% of them experience an appropriate ICD activation during their life. Implantation guidelines need to be improved and the inclusion of genetic factors represents a possibility. Multiple independent studies have shown that variants in the *NOS1AP* region are associated with SCD and with QT length, a predictor of SCD. Recently, a functional variant located in a cardiac enhancer upstream *NOS1AP*, rs7539120, has been shown to modulate the QT interval through altered cellular electrophysiology in cardiomyocytes. We studied the association of rs7539120 with life-threatening arrhythmias, as identified by the appropriate ICD shock and/or anti-tachycardia pacing therapy.

We analyzed 286 patients affected by ischemic or non-ischemic hypokinetic cardiomyopathy, enrolled at the hospitals of Bolzano and Trento (Italy), who received an ICD for primary prevention of SCD. The cohort was followed-up for a median time of 3.04 years (IQR: 1.71-5.68) after implantation, during which 79 (28%) appropriate ICD activations occurred. We analyzed the data using multiple Cox regression models adjusted for age, gender and hospital. In the 128 non-ischemic cardiomyopathy patients, rs7539120 was associated with appropriate ICD shock (HR=2.58, 95% CI: 1.30-5.11, p=0.006), while we did not find any evidence of an effect in the 158 ischemic patients (HR=1.03, 95% CI: 0.62-1.73, p=0.904). Our findings are consistent with the hypothesis that a stronger genetic risk component underlies non-ischemic cardiomyopathy.

74 | Structural Brain Imaging (MRI) Case-Control Study of Cortical Thickness and Surface area in Children Affected with Attention Deficit Hyperactivity Disorder (ADHD)

Nellie Fotopoulos[1,2], Gabriel A. Devenyi[1], Mallar M. Chakravarty[1,3,4], Sherif Karama[3,5], Sarojini M. Segunpta[1,3], Natalie Grizenko[1,3], Ridha Joober[1,2,3]

[1] *Douglas Institute, Montreal, Montreal, Canada;* [2] *Department Human Genetics, McGill University, Montreal, Canada;* [3] *Department of Psychiatry, McGill University, Montreal, Canada;* [4] *Department of Biomedical Engineering, McGill University, Montreal, Canada;* [5] *Montreal Neurological Institute, Montréal, Canada*

MRI studies have attempted to elucidate the structural brain differences between children with and without ADHD. Findings in the literature are conflicting; some studies report morphological differences although others do not. Evidence has emerged that motion artifacts present on images can underestimate measurements. Hyperactivity is one of the main symptoms of ADHD; therefore, positive results must be interpreted with caution. Moreover, an unexplored venue of imaging is studying effects of environmental risk factors on brain morphology in children with ADHD.

Our primary objective was to conduct a structural MRI study that compares the cortical thickness and surface area in children with ADHD and controls. A second objective was to explore the effect of maternal smoking during pregnancy on brain morphology in children with ADHD. A third objective is to investigate the effect of polygenic risk scores of candidate ADHD genes on brain morphology.

ADHD cases were recruited from a double-blind placebo control clinical trial from the ADHD clinic. All subjects were scanned with a 3T MRI scanner at the CIC at the Douglas Institute. General linear modeling was performed using CIVET 1.1.12. Cortical thickness and surface area were used as key outcome measures.

Analysis showed that age and gender had significant effects on cortical thickness and surface area in both groups (n=103). Secondary analysis revealed no statistically significant difference in cortical thickness or surface area between children with ADHD and controls. However, a significant increase in surface area was observed in children with ADHD who were exposed to maternal smoking during pregnancy.

75 | On Model Selection for Genetic Effects on Response Measures in the Presence of Correlation with Baseline Values

René Fouodjio[1], Yassamin Feroz Zada[1], Marie-Pierre Dubé[1,2]

[1] *Beaulieu-Saucier Université de Montréal Pharmacogenomics Centre, Montreal Heart Institute, Montreal, Canada;* [2] *Faculty of Medicine, Université de Montréal, Montreal, Canada*

In pharmacogenomics studies, it is often of interest to assess the effect of a genetic variant on a continuous outcome after drug treatment. Previous studies have demonstrated that a regression model which adjusts the Post-Treatment (PT) score by the Baseline (BL) score (model 1) was more powerful than the analysis of change or percentage scores (model 2). However, the question remains for a situation where there would also be an association of the genetic variant with the baseline measure in addition to having an effect on the response to treatment. We have performed simulations to assess this question. We assumed a genetic additive model modeling Low-Density Lipoprotein cholesterol (LDL) measured at BL and PT with a genetic exposure G. We compared Type I and Type-II error rates for model 1 and model 2 for situations with and without correlation between post-treatment and baseline scores, and correlations between G and baseline scores. We performed Monte Carlo simulations in SAS under the null hypothesis that varying conditions of corr(PT, BL) and corr(BL, G) do not affect the genetic association with PT scores and change scores.

Our results suggest that under situations with low correlations of BL-PT and G-BL, model 1 and model 2 perform differently, with a type I error rate for model 2 significantly higher than that of model 1. In other situations, model selection should be guided by the research questions, whether they are intended to be limited to genetic effect on drug response or any genetic effect.

76 | A Genome-Wide two-Component Mixture Model Expectation-Maximization Algorithm for Time to Event Data

Ben Francis[1], Peng Yin[1], James Cook[1], Andrea Jorgensen[1], Jane Hutton[2], Andrew Morris[1]

[1]University of Liverpool, Liverpool, United Kingdom; [2]University of Warwick, Coventry, United Kingdom

Traditional survival analysis of Time To Event (TTE) data assumes that all individuals will experience the Event Of Interest (EOI). In pharmacogenetics studies, there are patients who will not experience the therapeutic effect of a drug regardless of dosage or duration of prescription. Those who are unable to experience the EOI are deemed to be the part of the “cure fraction”.

Modelling TTE data consisting of those susceptible to the EOI and the cure fraction requires a “two-component” approach; enabling estimation of the effect of covariates on both susceptibility and the time to the occurrence of an EOI. One widely-used method incorporates an accelerated failure time model and expectation-maximization algorithm but is too computationally intensive to be applied genome-wide. To circumvent this problem, we obtained survival and susceptibility residuals from a model including clinical covariates only, to then use as phenotypes in a linear regression model and a multivariate “reverse regression” analysis.

To assess the performance of these approaches, we performed detailed simulations incorporating a range of models of SNP effect on survival and susceptibility. Under a null model of no association of a SNP with survival or susceptibility, the type I error rates of all analytical approaches were maintained. Over the range of association models considered, the multivariate reverse regression approach was more powerful than linear regression for survival and susceptibility and no less powerful than the full two-component model. In conclusion, we have developed a novel “approximate” computationally efficient approach to enable genome-wide analysis of two-component TTE data.

77 | Genetic Variability in Both the Adaptive and Innate Immune Systems Contribute to Alzheimer’s and Parkinson’s Disease Risk

Sarah A. Gagliano[1,2,3], Jennie G. Pouget[2,3], John Hardy[4], Michael R. Barnes[5], Jo Knight[6], Mina Ryten[4], Michael E. Weale[1]

[1]Department of Medical & Molecular Genetics, King’s College London, London, United Kingdom; [2]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada; [3]Institute of Medical Science, University of Toronto, Toronto, Canada; [4]Institute of Neurology, University College London, London, United Kingdom; [5]William Harvey Research Institute Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [6]Faculty of Health and Medicine and Data Science Institute, Lancaster University, Lancaster, United Kingdom

Neurodegenerative disorders are personally devastating and a burden on health-care systems worldwide. Studies have demonstrated enrichment of disease-associated genetic variants with functional annotations. These annotations vary depending on cell/tissue-type. Given the many ways in which complex disorders arise, and for human brain disorders, the well-recognized cellular heterogeneity of the brain, pinpointing cell-types of interest is important to further understand pathogenicity.

We obtained publically available GWAS summary statistics from Parkinson’s Disease (PD), Alzheimer’s Disease (AD) and Amyotrophic Lateral Sclerosis (ALS). We applied stratified LD-score regression to determine if functional categories (e.g. histone modifications) from various tissues are enriched for heritability. We also looked at the enrichment of sets of brain and immune genes.

We found little enrichment for heritability in brain annotations. Annotations in immune cells from outside the brain, from both the adaptive and innate immune systems, were significantly enriched for AD and to a lesser degree for PD.

Our gene list analysis provided complimentary results. The immune gene list was enriched for heritability in AD (5.19, p=4.8×10^{-4}), and the effects in PD and ALS were suggestive but did not survive multiple-testing correction (4.48, p=0.02 and 2.46, p=0.03, respectively). The brain gene list was not significantly enriched in any of the disorders (enrichment range: 0.87 - 1.85, p>0.04 for all three disorders).

From multiple lines of evidence we show that there is a significant contribution of variants exhibiting functional marks in immune cells to the heritability of two neurodegenerative disorders, namely AD and PD.

78 | Evaluation of Serum Level Parameters and *C677T* Polymorphism of *MTHFR* Gene in Preeclampsia Patients

Payam Ghasemi-Dehkordi[1], Azar Jaafari[2], Shahrbanuo Parchami-Barjui[1], Somayeh Reiisi[3],
Morteza Hashemzadeh-Chaleshtori[1], Sepideh Miraj[2]

[1]Cellular and Molecular Research Center, Shahrekord University of Medical Sciences, Shahrekord, Iran; [2]Department of Obstetrics and Gynecology, Shahrekord University of Medical Sciences, Shahrekord, Iran; [3]National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

The aim of this study was to evaluate the correlation between *C677T* polymorphism of Methylenetetrahydrofolate reductase (*MTHFR)* gene, acid folic, and homocysteine serum levels in Iranian pregnant women with preeclampsia. This study was performed in 129 preeclamptic pregnant women and 125 control individuals and *MTHFR* gene (*C677T* polymorphism) was determined by Polymerase Chain Reaction-Restriction Fragment Length Polymorphism (PCR-RFLP) method and the plasma levels of homocysteine and acid folic was measured by ELISA method. The *CC*, *CT*, and *TT* genotypes were not significantly different in patients compared to control (p=0.614). Low mean levels of homocysteine and acid folic in the preeclamptic cases were observed compared to control group. Also, the levels of BMI, gestational age, and neonatal weight statistically different in two groups and other variables revealed no significant differences between these groups. These findings showed that there was not any correlation between the *C677T* polymorphism of *MTHFR* gene and preeclampsia but the *TT* genotype of *C677T* polymorphism seems to be a protective factor for preeclampsia. It is also concluded that in our study homocysteine and acid folic serum levels and BMI are significantly affected on patients with preeclampsia compared to controls and can increase the risk of developing sever side effect to mothers and neonates.

79 | Mapping Multivariate Phenotypes in the Presence of Missing Data

Saurabh Ghosh[1]

[1]Human Genetics Unit Indian Statistical Institute, Kolkata, India

Clinical end-point traits are often characterized by quantitative and/or qualitative precursors and it has been argued that it may be statistically a more powerful strategy to analyze a multivariate phenotype comprising these precursor traits to decipher the genetic architecture of the underlying complex end-point trait. We (Majumdar et al., 2015) recently developed a Binomial Regression framework that models the conditional distribution of the allelic count at a SNP given a vector of phenotypes. The model does not require a priori assumptions on the probability distributions of the phenotypes. Moreover, it provides the flexibility of incorporating both quantitative and qualitative phenotypes simultaneously. However, it may often arise in practice that data may not be available on all phenotypes for a particular individual. In this study, we explore methodologies to estimate missing phenotypes conditioned on the available ones and carry out the Binomial Regression based test for association on the "complete" data. We partition the vector of phenotypes into three subsets: continuous, count and categorical phenotypes. For each missing continuous phenotype, the trait value is estimated using a conditional normal model. For each missing count phenotypes, the trait value is estimated using a conditional Poisson model. For each missing categorical phenotype, the risk of the phenotype status is estimated using a conditional logistic model. We carry out simulations under a wide spectrum of multivariate phenotype models and assess the effect of the proposed imputation strategy on the power of the association test vis-à-vis the ideal situation with no missing data.

80 | RVS: An R Package to Integrate Next Generation Sequencing (NGS) Data Across Cohorts for Association Analysis

Jiafen Gong[1], Zeynep Baskurt[1], Andriy Derkach[2], Angelina Pesevski[1], Lisa J. Strug[1,3,4]

[1]Program in Genetics and Genome Biology, Research Institute, The Hospital for Sick Children, Toronto, Canada; [2]Biostatistics Branch, National Cancer Institute, 9609 Medical Center Drive, Rockville, Maryland, United States of America; [3]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Canada; [4]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada

A growing number of studies are using NGS technologies. Integrating NGS data across cohorts for association analysis is challenging. Different studies may use different sequencing platforms, parameters (e.g. read depth), and alignment and variant calling algorithms, that can result in spurious findings unless statistical methodology implemented explicitly accounts for them. Previously, we developed a workflow to conduct case-control association analysis for common and rare variants with 'out-of-study' controls when aligned reads in BAM format are available. The proposed method, the Robust Variance Score (RVS), eliminates read depth bias, controls type I error and has comparable power to studies when true genotypes are known. Here we develop a user-friendly R package to make the RVS accessible, and extend its utility to integration of NGS data for multiple control samples, and NGS data across consortia for quantitative trait analysis. The '*RVS'* package contains three modules: (1) variant association analysis using RVS, which takes a multi-sample variant call format file (VCF, for example, from Genome Analysis Tool Kit (GATK)) created from all samples as input, obtains the genotype probabilities for the reported variants, and uses them in the RVS statistic for association analysis; (2) conventional association using variant calls from plink format files as input; and (3) a simulation module to simulate sequence data for different read depths and error rates. The RVS package has been tested using simulated data, data downloaded from 1000 genomes project website and NGS data from our studies of childhood-onset epilepsy. The RVS package is available at www.tcag.ca.

81 | Combining Evidence After Selection: A Powerful Framework for Testing the Global Null Hypothesis

Emery T. Goossens[1], Lei Sun[2]

[1]Purdue University, West Lafayette, Indiana, United States of America; [2]University of Toronto, Toronto, Canada

Statistical analyses of high dimensional data, including genetic association studies, are often interested in combining evidence across multiple sources. The most common setting is meta-analysis, in which P values (or other types of summary statistics) are combined across all studies to test the global null hypothesis that there is no association. This agnostic (either fixed-effect or random-effect) approach, however, may not be powerful when a proportion of the studies have truly null effects. Selective inference (appropriately testing parameters after some selection algorithm) has the potential to address this issue, but requires explicitly accounting for inherent selection bias. We propose a novel ordered-subset approach that simultaneously performs signal selection and adjusts for selection bias. We derive the approximate distribution for the test statistic under the null, and show the proposed method is accurate in finite samples. The performance of this method depends on a combination of factors: k, the total number of variables to be combined; k1 (k0), the number of true (null) variables; and the corresponding signal strength. Compared with traditional meta-analysis, our method improves power in the presence of heterogeneity when k0/k1 is large and each signal is relatively weak. In the absence of a mixture of null and non-null signals, we show the loss of power is limited via a simulation study. We also employ our method in the regression setting, illustrating its suitability for joint analysis of multiple SNPs in gene-based association studies and multiple genes in pathway analyses.

82 | Clustering Phenotype Trajectories with Genotype Covariates

Keelin Greenlaw[1], Eva Unternaehrer[2], Celia Greenwood[1], Antonio Ciampi[1]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Douglas Mental Health University Institute, McGill University, Montreal, Canada

In neurodevelopmental research, assessed phenotypes are expected to change over time, for example, mother-child interactions and bonding as the child develops. By including several of these longitudinal phenotypes, there is potential to examine and classify different types of phenotype trajectories, and it is of interest to identify genotypes that affect phenotype trajectory classification. We propose the use of latent class regression models, where we include predictors that influence the observable phenotype measurements, and covariates, such as genotypes, that influence the latent variable. To investigate this method, a dataset was simulated according to a model of mother-child interaction, which is modified by genotype. For a simple demonstration, we assume a gene variant of interest with known risk allele frequency. We then imagine a scenario where three variables are measured over time: child anxiety, child-mother attachment, and maternal anxiety. These three variables, together, describe three latent behavioral types, and we assume the genotype influences the probability of falling into each of the latent types. We show that the latent class regression model correctly identifies clusters of phenotype trajectories and the genotype as being significantly associated with latent class membership.

83 | Agreement in DNA Methylation Levels from the Illumina 450K Array, Across Batches, Tissues and Time

Celia M.T. Greenwood[1,2,6], Marie Forest[2,6], Gregory Voisin[2,6], Hélène Gaudreau[3,6], Michael Kobor[7], Julia MacIsaac[7], Lisa M. McEwen[7], Marinus H. Van IJzendoorn[4], Michael J. Meaney[3,5,6], Kieran J. O'Donnell[3,6]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Departments of Oncology, Epidemiology, Biostatistics & Occupational Health, and Human Genetics, McGill University, Montreal, Canada; [3]Douglas Mental Health University Institute, McGill University, Montreal, Canada; [4]Centre for Child and Family Studies, Leiden University, Leiden, The Netherlands; [5]Departments of Psychiatry and Neurology & Neurosurgery, McGill University, Montreal, Canada; [6]Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Canada; [7]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Vancouver, Canada

Although agreement has often been examined between technical replicates of DNA methylation levels on the Illumina 450K array, few studies have looked simultaneously at the effects of batch, tissue and time. Two different studies (Canadian, Dutch) with different designs included a mixture of technical replicates and biological replicates with different lengths of time between the replicates, in buccal swabs, whole blood and blood spots. Quality control steps included cell type mixture adjustment with surrogate variable analysis and normalization using funtooNorm. Agreement between replicates was assessed using Intraclass Correlations (ICCs) and examination of F-statistics from partitioned analyses of variance. Redundancy Analysis (RDA) was used to identify outliers and trends with covariates. Analysis of variance contrasts separating technical and biological replicates clearly showed better agreement between (i) technical replicates versus repeated samplings, (ii) buccal cells versus blood or blood spots, (iii) within a linear row on a chip versus between chips in a batch. ICCs demonstrated that a majority of probes have inter-individual variability of similar magnitude to within sample variability; however as the inter-individual variability increased, so did ICC. RDA analysis using the first two projections identified gender, tissue and batch effects. However, among technical replicates in the Dutch data set, we were also able to detect subtle differences in the agreement between females, one male individual, and one pregnant individual with RDA. Careful study design and large sample sizes are essential for DNA methylation studies, because the signals are susceptible to many types of bias.

84 | Penalized Estimation of Sparse Concentration Matrices Based on Prior Knowledge with Applications to Placenta Metal Data

Jiang Gui[1,4], Jai Woo Lee[1], Tracy Punshon[2], Margaret Karagas[3]

[1]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, United States of America; [2]Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire, United States of America; [3]Department of Epidemiology, Geisel School of Medicine, Lebanon, New Hampshire, United States of America; [4]Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, New Hampshire, United States of America

Essential metals (K, Ca, Fe, Cu, P, Mg, Zn, Co, S, Se, Mn) in placenta play critical cellular roles as structural components of bimolecular, signaling molecules, catalytic cofactors and regulators of protein expression. Their concentrations are tightly regulated via complex homeostatic networks, and altered metal homeostasis is characteristic of disease. The chemical architecture for those metals may be complex and advanced biostatical methods are demanded to infer the dependency structures of the metals in placenta. In this talk, we introduce a weighted sparse Gaussian graphical model that can incorporate prior knowledge to infer the structure of the network. We present a L1 penalized regularization procedure for estimating the sparse precision matrix in the setting of Gaussian graphical models. Simulation results indicate that the proposed method yields a better estimate of the precision matrix than the procedures that fail to account for the prior knowledge of the network structure. We also applied this method to a New Hampshire Birth Cohort Study for inference of the chemical network in placenta blood.

85 | Exploring the Genetic and Environment basis of Smoking in Association with Depression and Schizophrenia in the Scottish Population

Lynsey S Hall[1,2], Yan-Ni Zeng[1], Jude Gibson[1], Ella Wigmore[1], Ana-Maria Fernandez Pujals[1], Mark J Adams[1], Caroline Hayward[3], Chris S Haley[3,4], Andrew M McIntosh[1,5]

[1]Division of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom; [2]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; [3]Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [4]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [5]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom

Individuals with a diagnosed mental health problem have an increased likelihood of smoking relative to the general population. Data from Generation Scotland: The Scottish Family Health Study (n=19,994) was used to estimate the additive genetic and shared environmental contributions to smoking, the genetic correlation between smoking and a diagnosis of depression, and the association between smoking and Polygenic Risk Scores (PGRS) for depression and schizophrenia.

Mixed Linear Model Analysis (MLMA) was used to examine the contribution of additive genetics, and shared household, sibling and spouse environments to smoking, and to estimate the genetic and phenotypic correlations between smoking and a diagnosis of depression (2659 cases; 17327 controls). Data from two independent GWAS were used to test whether smoking was associated with an increased loading for depression- and schizophrenia-associated alleles.

Additive genetics accounted for 0.34 (0.02) of the phenotypic variance. Shared spouse environment accounted for 0.42 (0.02), sibling for 0.03 (0.01) and shared household for 0.03 (0.02). Smoking had a moderate and significant genetic correlation (rg (se)=0.42 (0.12), chi2=78.9, p<1E−16) and phenotypic correlation (rp (se)=0.11 (0.007), p=0.007) with depression. Smoking did not associate with depression PGRS (maximum beta (se)=0.011 (0.008), R2=0.01%, p=0.20), however it had a significant association with schizophrenia PGRS (maximum beta (se)=0.048 (0.009), R2=0.19%, p=5.68E−08).

Limitations of the current study include the possibility that spouse effects may be due to assortative mating and the relatively small polygenic risk score effect sizes.

86 | Impact of Genotyping Errors and Missingness on Phasing and Imputation in a Population Isolate

Anthony Francis Herzig[1,2], Teresa Nutile[3], Marie-Claude Babron[1,2], Marina Ciullo[3], Celine Bellenguez[4,5,6], Anne-Louise Leutenegger[1,2]

[1]Université Paris-Diderot, Sorbonne Paris Cité, UMR946, F-75010 Paris, France; [2]Inserm, UMR 946, Genetic Variation and Human Diseases, F-75010 Paris, France; [3]Institute of Genetics and Biophysics A. Buzzati-Traverso – CNR, Naples, Italy; [4]Inserm, U1167, F-59000 Lille, France; [5]Institut Pasteur de Lille, F-59000 Lille, France; [6]Université Lille, U1167, F-59000 Lille, France

In the search for genetic associations with complex traits, population isolates offer the advantage of reduced genetic and environmental heterogeneity. Proposed cost-efficient next-generation association approaches in such populations - where only a subset of study individuals is sequenced - require high quality genetic imputation and preliminary phasing. To identify an effective study-design for such genetic association studies in population isolates we compare a range of phasing and imputation methods by simulations, either recently developed for population isolates or previously established for general outbred populations. Genotyping errors and missing genotypes are simulated to observe their effects on the performance of each algorithm. We also compare different sequencing strategies by assessing the benefits of obtaining either a genome or exome sequenced local reference panel specific for the population isolate.

We simulated full sequence data on chromosome 10 over the large complex pedigree recorded in Campora, a village within the established population isolate of Cilento in southern Italy. Genotypes for Campora individuals with available DNA are simulated via gene-dropping along the pedigree with founder haplotypes sourced from the UK10K reference panel. We first assess the phasing performance of SHAPEIT2, EAGLE and SLRP software by comparing switch-error-rates. For imputation we compare IMPUTE2, BEAGLE and PRIMAL by computing correlations between imputed genotypes and true simulated genotypes. Furthermore, we examine the impact of genotyping errors and missing genotypes on each algorithm by considering the quality of phasing and imputation in the neighbourhoods of variants with simulated errors or missingness.

87 | Identifying Molecular Elements that Underlie Eye Disorders and Vision Loss Using Predicted Gene Expression

Jibril Hirbo[1], Eric Gamazon[1], Patrick Evans[1], Milam Brantley[2], Nancy Cox[1]

[1] Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2] Vanderbilt Eye Institute, Vanderbilt University Medical Center, Nashville Tennessee, United States of America

Eye disorders and vision loss pose a huge economic burden among Americans, estimated at over $150 billion in 2013. One of the main limiting factors in developing effective therapies for eye disorders is lack of clear understanding of the molecular mechanism in disease etiologies. Genome-Wide Association Studies (GWAS) have successfully identified numerous variants associated with some common eye diseases like glaucoma, age dependent macular degeneration, cataract etc., but the biological mechanism underlying these associations are yet to be elucidated. Gene expression and how it correlates with disease can provide a powerful method to prioritize genes involved in the etiology of the eye phenotypes. We used recently developed gene based method that provide a framework for directly correlating imputed gene expression data with individual's genetic profile. The imputed gene expression is generated using a transcriptome reference panel. This method was applied to over 18,000 samples in BioVU, Vanderbilt University Medical Center's biorepository of DNA extracted from discarded blood linked to electronic medical records collected during routine clinical testing. We identified both novel genes and those previously implicated in eye phenotypes. We found that nearly 40% of genes identified were previously implicated in Mendelian and congenital conditions of the nervous system. Some of the genes are those that were associated with complex neurological, eye disorders and metabolic disorders based on GWAS. Finally, ten novel genes, three of which are from olfactory gene family showed significant association with eye phenotypes.

88 | Polygenic Model does not Explain Very Low Odds Ratios (ORs)

Susan E. Hodge[1,2], David A. Greenberg[1,2]

[1] Battelle Center for Mathematical Medicine, The Research Institute, Nationwide Children's Hospital, Columbus, Ohio, United States of America; [2] Department of Pediatrics, The Ohio State University, Columbus, Ohio, United States of America

It is widely believed that polygenic inheritance – i.e., many genes, each with only small effect, working together additively – can explain the very low ORs, e.g., $OR<1.2$, typical of statistically significant case-control SNP differences in GWAS. However, we have shown that this is not true, over a broad range of polygenic models. Let n=number of disease loci, p=frequency of susceptibility allele at each disease locus (assumed same for all loci), and k=threshold (i.e., individuals are affected if and only if they have $\geq k$ susceptibility alleles). We examined models with n=20, 50, 100 loci, allele frequencies of p=0.05, 0.10, 0.20, 0.30…, and k set so that population prevalence is between 1–10%. With prevalence fixed at ∼5%, the *minimum* possible ORs are approximately 2.24, 1.68, 1.44, 1.31 for n=20, 50, 100, 200, respectively. ORs were lower only when prevalence increased to more than 20% or when both n and p increased; e.g., n=400, $p = 0.4$, and k=343 yielded OR=1.20. The "threshold ratio" $H = \frac{k}{2np}$, helps to understand this phenomenon. Thus, the polygenic model seems unlikely to explain the usual GWAS results, or diseases such as schizophrenia or diabetes, and more attention should be directed to considering genetic heterogeneity.

89 | A Novel Phenotype Permutation Method to Optimize Threshold selection in Random Forests

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

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

Standard analysis methods for GWAS are not robust to complex disease models, such as interactions between variables with small main effects. These effects very likely contribute to the genetic etiology of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One currently unresolved problem with RF or any other machine learning scheme is that there is no standardized method of selecting a threshold that identifies the correct variables while filtering out false positives. The optimal threshold is highly dependent on the underlying genetic model, which is typically unknown in biological data.

To address the long-standing threshold problem, we have developed an RF-based variable selection method called r2VIM. Among other novel features, this method incorporates a permutation scheme to optimize threshold selection. Briefly, we permute the phenotype and run r2VIM on the permuted data. This provides us with a null distribution of importance scores while maintaining the relationships between the predictor variables. A threshold is then selected based on the false positive rate in the permuted data. This threshold is then applied to the non-permuted r2VIM analysis. We refer to this permutation scheme as "phenoPerm."

Our results show that phenoPerm provides an ideal balance between power and false positive selection when compared to other selection methods across different complex genetic risk models, including gene-gene interactions, both with and without main effects. We also assess the relationship between the false positive rate in the permuted data to the false positive rate and power in the non-permuted data, and show how it can be used as a guide to perform threshold selection according to individual analysis requirements.

90 | Joint Metabolomic and Epigenomic Study of Cigarette Smoking

Yunfeng Huang[1], Qin Hui[1], Douglas Walker[2], Jack Goldberg[3], Dean Jones[2], Viola Vaccarino[1], Yan V. Sun[1]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America; [2]Department of Medicine, School of Medicine, Emory University, Atlanta, Georgia, United States of America; [3]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, United States of America

The chemicals and metabolites induced by cigarette smoking cause a wide range of molecular and cellular changes in the human body, and lead to many diseases. DNA methylation plays an important role in the pathways of smoking and smoking-induced diseases, and previous Epigenome-Wide Association Studies (EWAS) have identified a large number of smoking-associated CpG sites, but specific chemicals and pathways involved in the modification of smoking-induced DNA methylation remain unclear. Elucidation of these pathways may shed light into causal mechanisms. We conducted EWAS with multiple smoking-related metabolites identified from a metabolome-wide association study using a sample of 180 middle-aged male twins. We identified and annotated eleven smoking-related metabolites for separate EWAS. After adjusting for age, BMI, relatedness, batch effect and cell type heterogeneity, we identified 379, 352, 116, 89, 84, 72 and one CpG sites significantly associated (false discovery rate corrected $p<0.05$) with 13-C cotinine, N-acetylpyrrolidine, hydroxycotinine, 8-oxoguanine, cotinine, nicotine and caffeine, respectively. No CpG site was significantly associated with norcotinine, hexyl glucoside, hydroxypyridine or ornithine. Conditional on current smoking status, most identified associations diminished with a proportion remaining marginally significant. The epigenetic association with caffeine were not affected by the adjustment of smoking and remained epigenome-wide significant. The trans-omic analysis between smoking-related metabolites and smoking-associated CpG sites showed common and unique metabolites and pathways induced by smoking. Our joint metabolomic and epigenomic study provided a more accurate and comprehensive understanding of metabolic and epigenetic changes, as well as complementary pathways, influenced by cigarette smoking.

91 | Tissue-Specificity, Timing and Sexual Dimorphism in the Expression of Puberty Related Genes Implicated in Disease and GWAS Studies

Huayun Hou[1,2], Liis Uusküla-Reimand[1], Maisam Makarem[1,2], Christina Corre[1], Ariane Metcalf[1], Anna Goldenberg[1,2], Michael D. Wilson[1,2], Mark R. Palmert[1,2]

[1]Hospital for Sick Children, Toronto, Toronto, Canada; [2]University of Toronto, Toronto, Toronto, Canada

The timing of puberty varies among the general population, is sexually dimorphic, and associates with adverse health outcomes, including risks for breast cancer and cardiovascular disease. Rare mutations in patients with pubertal disorders have identified critical components of the Hypothalamic-Pituitary-Gonadal (HPG) axis, and recent GWAS studies have identified more than 100 loci associated with Age At Menarche (AAM). Little is known about how the GWAS-associated loci regulate the timing of puberty because only a few of them fall within or near genes previously implicated in regulation of the HPG axis.

To gain insights into how these newly identified genes influence pubertal timing, it is important to characterize the tissue specificity, developmental timing, and sexual dimorphisms in gene expression patterns for these genes.

To achieve this, we asked whether puberty disorder and GWAS-associated genes show tissue-specific expression using transcriptome data from 77 normal human tissues. We also investigated the dynamic regulation of these genes using a sensitive and high throughput microfluidic real time qPCR strategy at five time points that span the pubertal transition.

The expression of the GWAS-associated genes was significantly enriched for 10 tissues including hypothalamus, pituitary gland, and interestingly, the pineal gland. Using qPCR, we validated pineal-enriched expression of many GWAS genes, including *Bsx, Impg1, Rxrg, Wscd1, Rdh8, Tenm2* and *Olfm2*. In all tissues, the expression of puberty-associated genes changes most between postnatal day 12 and day 22, which is prior to the physical onset of mouse puberty. The largest number of genes with sex-biased expression is found in pituitary while the least in hypothalamus. Pituitary sex-biased genes are significantly enriched for pubertal disorders and nuclear receptor pathway and also include genes whose pituitary functions are unknown, such as *Dlk1* and *Retn*.

In conclusion, expression profiling of genes implicated in disease and GWAS studies revealed new potential mechanisms involved in regulating puberty timing.

92 | BRAVOE: A Bayesian Framework to Estimate Effect Sizes of Rare Genetic Variants in Case Control Studies

Hao Hu[1], Yao Yu[1], Chad Huff[1]

[1] *Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Genes harboring rare susceptibility variants are a major component of the genetic architecture of many human diseases. Accurate estimates of rare variant effect sizes are inherently difficult to obtain due to the limited number of rare allele observations. One commonly used approach is to group variants with similar properties, for example predicted functional severity, but this approach requires predefined variant groups and cannot model the effect size heterogeneity within each group. We present the Bayesian Rare Variant Odds-ratio Estimator (BRAVOE), which uses a hierarchical Bayesian framework to estimate ORs for rare variants. BRAVOE uses the allele count and variant annotation information to first identify the best hierarchical model and then calculate point estimates and OR Credible Intervals (CIs) for each variant. Variant annotations are modeled as covariates and can include categorical (e.g. nonsense vs. missense) or continuous (e.g. PolyPhen-2 scores) data. From simulations, we show that OR estimates from BRAVOE are substantially more accurate than logistic regression models with the same covariate information, with 43% to 56% lower Mean Squared Error (MSE). In a real-world breast-cancer case-control sequencing dataset on the *ATM* gene, BRAVOE identified two statistically significant predictors of missense variant ORs: whether the variant was located in the FAT, kinase, or FATC domains, and whether the variant had an AlignGVGD score of C65. Variants that satisfy both criteria had the highest estimated odds ratio (95% CI lower bound: 4.74). Our results demonstrate that BRAVOE provides an improved framework for rare variant effect size estimation in case-control sequencing studies.

94 | Brain-Derived Neurotrophic Factor Val66Met Variants (RS6265) are Associated with Coronary Heart Disease (CHD) Outcomes in a Patient Sample

Rong Jiang[1], Michael A. Babyak[1], Beverly H. Brummett[1], Elizabeth R. Hauser[2], Abanish Singh[1], Ilene C. Siegler[1], Svati H. Shah[2,3], Carol Haynes[2], Megan Chryst-Ladd[2], Redford B. Williams[1]

[1] *Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America;* [2] *Department of Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, United States of America;* [3] *Division of Cardiology, Department of Medicine, Duke University Medical Center, Durham, North Carolina, United States of America*

Exposure to psychosocial stress has been associated with an increased risk and adverse clinical course of Coronary Heart Disease (CHD). The *BDNF* Val66Met genotype (rs6265) moderates the impact of psychosocial stress on HPA axis function and may affect CHD outcomes. In this study, we examined the association among stress, Val66Met and clinical prognosis in a sample of 5510 cardiac catheterization patients (CATHGEN). The Val/Val genotype was associated with more severe coronary heart disease (number of diseased vessels –NDV; left ventricular ejection fraction – LVEF, P<0.05), and increased risk of adverse clinical events (death and myocardial infarction, P=0.034, HR=1.12). Decreased LVEF was associated with higher stress (constructed from financial strains and marital problems) in the Val/Val group, but not among Met carriers (P=0.011). The association between Val66Met and CHD events was partially mediated by NDV (P=0.02, estimated mediated proportion=13%, 95% CI=[3.4%, 69%]), and LVEF (P=0.05, estimated mediation proportion=13%, 95% CI=[0.4%, 72%]). Further, NDV and LVEF also partially mediated the stress effect on events within the Val/Val group (P<0.001), but not in the Met group (P>0.16). These findings in a large CHD patient sample support our hypothesis that Val/Val genotype may enhance the impact of psychosocial stress on the pathway through NDV and LVEF to CHD. If these findings are confirmed in further research, intervention studies in clinical groups with the Val/Val genotype could be undertaken to prevent disease and improve prognosis.

95 | A Targeted Genome Association Study Examining Transient Receptor Potential Ion Channels, Acetylcholine Receptors, and Adrenergic Receptors in Chronic Fatigue Syndrome

Samantha C. Johnston[1,2], Donald R. Staines[2], Sonya M Marshall-Gradisnik[1,2]

[1] School of Medical Science, Griffith University, Gold Coast, Australia; [2] The National Centre for Neuroimmunology and Emerging Diseases, Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia

Chronic Fatigue Syndrome, also known as Myalgic Encephalomyelitis (CFS/ME) is a debilitating condition of unknown aetiology. It is characterized by a range of physiological effects including neurological, sensory and motor disturbances. This study examined candidate genes for the above clinical manifestations to identify Single Nucleotide Polymorphism (SNP) alleles associated with CFS/ME compared with healthy controls.

DNA was extracted and whole genome genotyping was performed using the HumanOmniExpress BeadChip array. Gene families for transient receptor potential ion channels, acetylcholine receptors, and adrenergic receptors, and acetylcholinesterase were targeted. The frequency of each SNP and their association between CFS/ME and healthy controls was examined using Fisher's exact test, and multiple test corrections was applied with Bonferroni post-hoc analysis ($p<0.05$).

This study included 172 participants, consisting of 95 Fukuda defined CFS/ME patients (45.8 ± 8.9; 69% female) and 77 healthy controls (42.3 ± 10.3; 63% female). A total of 950 SNPs were analysed. 60 significant SNPs were associated with CFS/ME compared with healthy controls. After applying stringent Bonferroni correction, SNP *rs2322333* in adrenergic receptor $\alpha 1$ (*ADRA1A*) was higher in CFS/ME compared with healthy controls (45.3% vs. 23.4%; $p=0.058$). The genotype class that was homozygous minor (AA) was much lower in CFS/ME compared with healthy controls (4.2% vs. 24.7%).

We report for the first time the identification of *ADRA1A* and a possible association between CFS/ME and genotype classes. Further examination of the functional role of this class of adrenergic receptors may elucidate the cause of particular clinical manifestations observed in CFS/ME.

96 | Whole Exome Sequencing in Multigenerational Mixed Cancer Families Identifies a Putative Risk Variant in the *PDIA2* Gene

Rachel M. Jones[1], Phillip E. Melton[1], Alexander Rea[1], Evan Ingley[2], Mandy L. Ballinger[3], David J. Wood[4], David M. Thomas[3], Eric K. Moses[1]

[1] The Curtin UWA Centre for Genetic Origins of Health and Disease, Faculty of Health Sciences, Curtin University and Faculty of Medicine, Dentistry & Health Sciences, The University of Western Australia, Perth, Australia; [2] Harry Perkins Institute of Medical Research, Western Australia, Perth, Australia; [3] The Kinghorn Cancer Centre, Garvan Institute of Medical Research, New South Wales, Australia; [4] School of Surgery, University of Western Australia, Perth, Australia

The use of Whole Exome Sequencing (WES) in families represents an optimal study design for the identification of rare genetic variants involved in the risk of cancer. We investigated 3 multi-generational cancer-cluster families to identify putative deleterious exonic variants that predispose individuals for sarcoma and other cancers. We performed WES using Ion Ampliseq Exome RDY at 100× on genomic DNA extracted from peripheral blood samples from 19 patients and 2 sarcoma tumor samples. To reduce multiple testing, variants from WES were filtered to prioritize those detected in 101 known cancer genes (1,216 variants). These variants were analyzed in SOLAR with measured genotypes using a polygenic model. Our analysis revealed an association between a variant (rs45614840, c.1464C>G / p.T119R) in *PDIA2* at 16p13.3 and onset of all cancer patients in a single family, ($p=0.00028$). This variant is only present in 4 family members with cancer (1 sarcoma, 1 prostate, 2 melanoma) and not found in any unaffected family members or members of the other families and is predicted to be highly deleterious. A comparison between the matched tumor normal samples for the sarcoma proband shows an increase in the G allele from 49% in the germline to 60% in the tumor, indicating a possible loss of heterozygosity. A total of 49 variants in *PDIA2* have been previously reported and rs45614840 has been previously reported in two lymphoma cases. Our findings support the hypothesis that familial clustering of cancer is caused by pleiotropic variants that can influence multiple cancer types.

97 | Genome-Wide Interaction Study of Red-Blood Cell Fatty Acids on Inflammatory Biomarkers in the Framingham Heart Study

Anya Kalsbeek[1], Jenna Veenstra[1], Jason Westra[1], Craig Disselkoen[1], Caren Smith[2], Nathan Tintle[1]

[1] Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, Iowa, United States of America; [2] Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University, Boston, Massachusetts, United States of America

Numerous genetic loci have been identified which associate with fatty acids levels and/or inflammatory biomarkers (e.g., interleukin, c-reactive protein, tumor necrosis factor, etc.). Recently, using RBC fatty acid data from the Framingham Offspring Study, we conducted a Genome-Wide Association Study (GWAS) of over 2.5 million single nucleotide polymorphisms (SNPs) and 22 RBC FAs (and associated ratios) after adjusting for dietary covariates. Our analyses identified numerous causal loci. We now present the results of analyses which use a genome-wide approach to test for possible gene-FA interactions on a variety of inflammatory biomarkers to begin to evaluate the extent to which genetic variability modifies the relationship between circulating fatty acids and inflammatory biomarkers.

98 | Tight Clustering for Large Datasets with an Application to Microarray Data

Bikram Karmakar[2], Sarmistha Das[1], Sohom Bhattacharya[1], Rohan Sarkar[1], Indranil Mukhopadhyay[1]

[1] Human Genetics Unit, Indian Statistical Institute, Kolkata, India; [2] Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

This article is aimed to propose a practical and scalable version of tight clustering algorithm, originally introduced by Tseng and Wong (2005). This algorithm provides tight and stable relevant clusters as output leaving a set of points as noise or scattered points that would not go into any cluster. However the computational limitation to achieve this precise target of tight clusters prohibits it from being used for large microarray gene expression data or any other large datasets, which are common nowadays. Thus it is important to modify this method so as to make it applicable to large datasets.

We propose a modified and scalable version of tight clustering method keeping computational time and accuracy at a higher level. Our method is applicable to a dataset of very large size. With extensive simulation study and a real gene expression data analysis we present the validity of our proposed algorithm. For simulation data, the accuracy of our method is established through high value of Rand index, as we know the true clusters. On the other hand, for real dataset, we study the biological functions of the obtained clusters that seem to be meaningful. Thus our proposed method is applicable to large datasets maintaining a certain level of accuracy and removes the limitation of the existing tight clustering method.

99 | Meta Genome-wide Association for Total Cholesterol and High Density Lipoprotein Cholesterol in Type 1 Diabetes

Sareh Keshavarzi[1], Andrew Paterson[1], DCCT Research Group, CACTI Research Group, and WESDR Research Group

[1] Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada

Coronary heart disease is the major cause of death in type 1 diabetes (T1D). Plasma lipid levels have been shown to be a heritable, modifiable risk factor of cardiovascular disease in adults with T1D. This study aimed to identify genetic factors association with lipids, in people with T1D.

To identify new loci and refine known loci influencing lipid levels, the associations with Total Cholesterol (TC) and High Density Lipoprotein (HDL) were studied using up to 2,404 people with T1D. Findings from Genome-Wide Association Studies (GWAS) in 3 cohorts including Diabetes Control and Complications Trial (DCCT), Coronary Artery Calcification in Type 1 Diabetes (CACTI), and Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) were combined using inverse-variance meta-analysis. In each cohort a linear regression was fitted for HDL and TC to evaluate the additive effect of genotyped and imputed Single Nucleotide Polymorphisms (SNPs).

After genomic control correction, 58 SNPs in CETP region and 17 SNPs in APOE region reached to genome-wide association ($P<5\times10^{-8}$) for HDL and TC, respectively. Conditional analyses identified secondary signals for HDL in CETP region and for TC in APOE and LDLR regions. Of the associated loci with HDL and TC from the literature in non-diabetic individuals, rs3764261 in CETP and rs4420638 in APOE regions were associated with HDL and TC in diabetic individuals, respectively. These effects were in the same direction as the effect on HDL and TC in non-diabetic individuals. The magnitude of the effect of APOE was greater in the diabetic individuals.

100 | Allele-Based N-Test in Linkage Analysis

Sajjad Ahmad Khan[1, 2]

[1] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [2] Department of Statistics, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa, Pakistan

There are many tests of inheritance based upon sibling information for diseases that have late onset. The *N*-test (Green et al., 1983) is one of these tests, which utilizes information from affected siblings. The *N*-test is the count in affected siblings of the most frequently occurring haplotype from the father plus the analogous count from the mother. When applied to haplotypes, the *N*-test excludes recombinant families from the analysis. In this study, we modified the *N*-test to be based on alleles instead of haplotypes. This modified allele-based *N*-test can include all families (recombinant as well as non-recombinant). We carried out a simulation study to find the thresholds and powers.

101 | Adaptive Testing for Multiple Traits with Applications to Detect SNP-Brain Network Associations

Junghi Kim[1], Wei Pan[1]

[1] Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America

There has been increasing interest in developing more powerful and flexible statistical tests to detect genetic associations with multiple traits, as arising from neuroimaging genetic studies. Most of existing methods treat a single trait or multiple traits as response while treating an SNP as a predictor coded under an additive inheritance mode. In this paper we follow a reverse-regression approach in treating an SNP as an ordinal response while treating traits as predictors in a Proportional Odds Model (POM). In this way, it is not only easier to handle mixed types of multiple traits, e.g., some quantitative and some binary, but it is also more robust to

the common assumption of an additive inheritance mode. More importantly, we develop an adaptive test in a POM so that it can maintain high power across many possible situations. Contrary to the existing methods treating multiple traits as responses, e.g., in a Generalized Estimating Equation (GEE) framework, the proposed method can be applied to a high dimensional setting where the number of phenotypes (p) can be larger than the sample size (n), in addition to a usual small p setting. The promising performance of the proposed method was demonstrated with applications to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, in which either structural MRI driven phenotypes or resting-state functional MRI (rs-fMRI) derived brain functional connectivity were used as multiple phenotypes. The applications led to the identification of several top SNPs of biological interest. Furthermore, a simulation study showed competitive performance of the new method compared to several existing methods, including potential power gain of the new method in cases with a dominant inheritance mode.

102 | Powerful and Adaptive Testing for Multi-Trait and Multi-SNP Associations with GWAS and Sequencing Data

Junghi Kim[1], Yiwei Zhang[1], Wei Pan[1]

[1]*Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America*

There is accumulating evidence showing that some complex neurodegenerative and psychiatric diseases like Alzheimer's are due to disrupted brain networks, for which it would be natural to identify genetic variants associated with a disrupted brain network represented as multiple traits. In spite of its promise, testing for multivariate trait associations is challenging: if not appropriately used, its power can be much lower than univariate testing. Differing from most existing, we consider SNP set-based association testing to decipher complicated joint effects of multiple SNPs on multiple traits. We propose a highly adaptive test at both the SNP and trait levels, giving higher weights to those likely associated SNPs and traits, to yield high power across a wide spectrum of situations. We illuminate on relationships among the proposed and some existing tests, showing that the proposed test covers several existing tests as special cases. We compare the performance of the new test with several existing tests using both simulated and real data. The methods were applied to structural MRI data to identify genes associated with grey matter atrophy in the human brain default mode network. For GWAS, genes AMOTL1 on chromosome 11 was identified by the new test, while it was missed by single SNP-based analyses. The proposed method is also applicable to rare variants in sequencing data and extended to pathway analysis.

103 | Estimating and Testing Direct Genetic Effects in Directed Acyclic Graphs with Multiple Phenotypes using Estimating Equations

Stefan Konigorski[1,2], Yuan Wang[3], Candemir Cigsar[2], Yildiz E. Yilmaz[2,4,5]

[1]*Molecular Epidemiology Research Group, Max Delbrück Center (MDC) for Molecular Medicine in the Helmholtz Association, Berlin, Germany;* [2]*Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Canada;* [3]*Scotiabank Lethbridge Northside, Lethbridge, Canada;* [4]*Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada;* [5]*Discipline of Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada*

In genetic association studies, approaches to increase the power of association tests by jointly modeling multiple phenotypes are gaining popularity. However, any directional effects between genetic variants, intermediate phenotypes, and the primary phenotype have to be considered in the analysis. For example, eQTLs can predict gene expression, and both might further affect circulating protein levels and complex traits such as blood pressure. In addition, there might be unmeasured factors that mediate or confound the genetic effects on phenotypes. Hence, robust statistical methods are needed to accurately estimate the direct effect of genetic variants on the primary phenotype. In this study, we propose a new method for this purpose under the directed acyclic graph setting based on estimating equations with robust standard errors. We illustrate the method under generalized linear models and Accelerated Failure Time (AFT) models when the primary phenotype is subject to censoring. Simulation studies demonstrate the validity of the new method, and show that traditional multiple regression and structural equation models lead to invalid inferences under some scenarios. Furthermore, we demonstrate that the previously proposed G-estimation method for AFT models does not work properly when the primary phenotype is subject to censoring. Finally, we estimate and test genetic effects on blood pressure accounting for intermediate gene expression phenotypes in an application to the Genetic Analysis Workshop 19 dataset, using the proposed and traditional methods, and discuss the differences. The proposed method can identify genetic variants that would be missed by traditional methods and also prevent false positive findings.

104 | Blood Pressure Gene x Alcohol Exposure Interactions Capture Association Pleiotropy

Aldi T. Kraja[1], Daniel I. Chasman[2,3], Yun J. Sung[4], Hugues Aschard[3], Alisa K. Manning[5,6], Mary F. Feitosa[1], Thomas Winkler[7], Xiaofeng Zhu[8], Michael A. Province[1]; In the Name of CHARGE Gene-Lifestyle Interactions Working Group

[1]*Division of Statistical Genomics, Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University School of*

Medicine, St. Louis, Missouri, United States of America; [2]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America; [3]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America; [4]Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, United States of America; [5]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, United States of America; [6]Center for Genetics Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [7]Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; [8]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, United States of America

Associations of Systolic, Diastolic, Mean Arterial and Pulse Pressure with 1000G imputed SNPs in interaction with alcohol exposure, CurrD (Current Drinker vs. non-drinker) and QuD (up to 7 alcoholic drinks a week vs. more), may identify pleiotropic variants for hypertension. Interaction-meta-analyses of 23 studies, were performed in two groups with N>91,000 and N~64,000 individuals of European ancestry. Mega-meta of each group meta-results, correcting for correlated genetic association meta-scans followed. A mega-meta threshold -$\log_{10}$p > 8 identified 897 variants for CurrD and 442 for QuD. Sixty-four and sixty percent respectively had at least one trait-meta > 8 {3}, thus considered known Blood Pressure (BP) candidates; 6% and 10% had all four traits with 4 < meta-log10p < 8 {1}; the remaining 30% had combinations of traits by 2 and by 3 with 3< meta-log10p < 8 {2}. They were assigned respectively to 62 and 29 genes, where *NPR3*, *LOC105379235*, *RP1L1*, *MIR4286*, *LINC01150*, *NUCB2* and *TNKS*, *CATSPER2* for set {1}, 21 and 11 genes for set {2}, including set {3} to a total of 63 unique genes. Fifty-one BP pleiotropic genes exhibited **P**rotein-**P**rotein **I**nteractions with a network of 709 additional genes. Pleiotropic BP genes (connectivity degree) *COX5A* (44), *POC1B* (29), *INSR* (100), *BRAP* (54), *KIAA2013* (26), *NOS3* (54), *ATXN2* (58), *TNKS* (34), *FTO* (15) and *SH2B3* (15) ranked highest for PPI "PageRank". From the new set {1}, *TNKS* has been previously assigned to significant associations with HDLC, LDLC, CRP, calcium-binding proteins, alkaline phosphatase, MetS and asthma. We expand pleiotropy analysis to trans-regulatory elements.

105 | Transmission Based Association Test for Multivariate Phenotype using Quasi-Likelihood

Hemant Kulkarni[1], Saurabh Ghosh[1]

[1]Indian Statistical Institute, Kolkata, India

The classical Transmission Disequilibrium Test (TDT) [Spielman et al. 1993] based on the trio design is an alternative to the population based case-control design to detect genetic association as it protects against population stratification. Since the manifestation of most diseases are governed by multiple precursor traits, it is important to study the multivariate phenotype comprising these precursors to improve the statistical powers of the test procedure. We modify the classical transmission disequilibrium test for quantitative traits based on logistic regression [Waldman et al., 1999; Haldar and Ghosh, 2015] to incorporate multivariate phenotypes. We adopt a quasi-likelihood approach [Wedderbur, 1974] based on Generalized Linear Regression [McCullagh and Neldar, 1989] to develop a test of association for multivariate phenotypes. Since the Generalized Estimating Equation (GEE) approach [Gourieroux, Monfort, and Trognon, 1984; Liang and Zeger, 1986] used for solving the quasi-likelihood equation is highly influenced by outliers, we use a modified Resistance Generalized Estimating Equation approach (RGEE) [Preisser and Qaqish, 1999; Preisser and Qaqish, 1996; Hall, Zeger, and Bandeen-Roche, 1996] to down weight the outliers. We also explore a modified model that includes information on allelic transmission from both parents. We perform extensive simulations under a wide spectrum of genetic models and different correlation structures between the phenotypes. We compare our method with the FBAT test procedure [Lake et al. 2002] as well as the univariate approach. We find that the proposed method that incorporates information on both parents is more powerful than FBAT and the univariate approach.

106 | Gallbladder Cancer: Genetic Variants of Gemcitabine Metabolism Pathway Genes and Treatment Outcome

Ashok Kumar[1], Annapurna Gupta[2], Aarti Sharma[2], Sushma Agrawal[3], Neeraj Rastogi[3], Balraj Mittal[2]

[1]Department of Surgical Gastroenterology, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India; [2]Department of Genetic, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India; [3]Department of Radiotherapy, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India

Gallbladder carcinoma is one of the most aggressive biliary tract malignancies which continue to present many challenges including its treatment management. The aim of this study is to evaluate the effect of genetic variants involved in gemcitabine metabolism pathway genes. Total 222 GBC study subjects who were given gemcitabine based chemotherapy were recruited in this study. The five SNPs involved in Gemcitabine metabolic pathways studied were –$\text{CDA}_{\text{rs2072671}}$, $\text{DCK}_{\text{rs4694362}}$, $\text{RRM1}_{\text{rs9937}}$, $\text{hCNT2}_{\text{rs11854484}}$, and $\text{hENT1}_{\text{rs760370}}$. Treatment response was recorded by RECIST criteria and the toxicity profile by CTCAE version 3.0. Genotyping was performed by allelic discrimination assays. Statistical analysis was done by SPSS ver. 16. Higher order gene-gene interaction analysis was done by GMDR. Kaplan–Meier and Cox Regression tests were performed for Survival analysis. Data analysis revealed $\text{CDA}_{\text{rs2072671}}$ as significant modulator of treatment response, toxicity and survival in advanced stage GBC patients. Variant allele of $\text{RRM1}_{\text{rs9937}}$, $\text{hENT1}_{\text{rs760370}}$ was associated with drug related toxicities. In GMDR analysis, combination of $DCK_{rs4694362}$, $hCNT2_{rs11854484}$, $RRM1_{rs9937}$ was found to be significantly involved model in modulation of treatment response whereas interaction of $CDA_{rs2072671}$ $hCNT2_{rs11854484}$ $RRM1_{rs9937}$, emerged as best significant model for thrombocytopenia and neutropenia respectively. In this study, polymorphism of $CDA_{rs2072671}$ significantly influenced the treatment outcomes of GBC patients undergoing gemcitabine based treatment. Further a large and multicenter study is required to validate this finding.

107 | Gene- and Pathway-Based Association Tests for Multiple Traits with GWAS Summary Statistics

Il-Youp Kwak[1], Wei Pan[1]

[1] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America

To identify novel genetic variants associated with complex traits and to shed new insights on underlying biology, in addition to the most popular single SNP-single trait association analysis, it would be useful to explore multiple correlated (intermediate) traits at the gene- or pathway-level by mining existing single GWAS or meta-analyzed GWAS data. For this purpose, we present an adaptive gene-based test and a pathway-based test for association analysis of multiple traits with GWAS summary statistics. The proposed tests are adaptive at both the SNP- and trait-levels; that is, they account for possibly varying association patterns (e.g. signal sparsity levels) across SNPs and traits, thus maintaining high power across a wide range of situations. Furthermore, the proposed methods are general: they can be applied to mixed types of traits, and to Z-statistics or *P* values as summary statistics obtained from either a single GWAS or a meta-analysis of multiple GWAS. Our numerical studies with simulated and real data demonstrated the promising performance of the proposed methods. The methods are implemented in R package *aSPU*, freely and publicly available at CRAN.

108 | A New World of Biobanking: CARTaGENE, Created to Accelerate Breakthroughs in Disease Detection, Targeted Prevention and Personalized Medicine

Catherine Labbé[1], Joseph Tcherkezian[1], Catherine Boileau[1], Geneviève David[1], Sébastien Jacquemont[2], Anne Monique Nuyt[2]

[1] Centre de Recherche du Centre Hospitalier Universitaire (CHU) Sainte-Justine, Montreal Canada; [2] Department of Pediatrics, Faculty of Medicine, Université de Montréal, CHU Sainte-Justine, Montreal, Canada

Despite all the advancements in detection and treatment of several chronic diseases such as cancer, chronic disease-related deaths worldwide continue to increase at an alarming rate. Therefore, it is not too surprising that prevention and personalized medicine are gaining in popularity when it comes to disease control. Chronic diseases are caused by a combination of an individual's genetic predisposition and their exposure to certain environmental risk factors. CARTaGENE was created to find solutions to these very complex diseases. In fact, we are conducting the largest prospective health study on men and women in Québec. Since 2007, the project has recruited 43,000 individuals, aged 40–69 years, representing six metropolitan regions of the province and collected detailed lifestyle, health and medical data on these individuals. The program also gathered detailed physical measurements, clinical and biochemical measures at baseline. The fundamental goal of this unique public-funded project is to support the scientific community in identifying the determinants of chronic diseases of environmental and/or genetic origin. It was also created to accelerate the process of translational medicine through the identification of biomarkers for early diagnosis, disease treatment and prevention. Epigenomic, genotyping, and exome sequencing data have been collected in subsets of CARTaGENE participants and have resulted in both preliminary data and high profile studies, while supporting over 40 research programs. These ongoing partnerships, our interoperability with other national biobanks, and our plan to develop other sectors of activity support our mission to provide meaningful high-quality data and biospecimens for years to come.

109 | Exploring Gene×Environment Interactions through Pathway Analysis

Elin Larsen[1,2], Thérèse Truong[3], Marina Evangelou[4]

[1] Université Claude Bernard Lyon 1,Villeurbanne, France; [2] Ecole Centrale de Lyon, Ecully, France; [3] INSERM, Paris, France; [4] Department of Mathematics, Imperial College London, London, United Kingdom

Pathway analysis incorporates the available biological knowledge of SNPs (and genes) for association with disease. It has been used very successfully in recent years, identifying both pathways and SNPs that are associated with diseases of interest. The self-contained (association) null hypothesis of interest in the setting of pathway analysis states that the pathway SNPs are not associated with the phenotype.

Frequentist approaches proposed for pathway analysis combine the individual single-SNP analysis *P* values for testing the self-contained null hypothesis. As discussed in the literature, two of the most powerful methods are the Fisher's Product Method (FM) and the Adaptive Rank Truncated Product method (ARTP). Both approaches are usually combined with a permutation approach for estimating the association *P* value of each pathway. In our conducted work, we have adapted both of these methods to test for Gene with Environment (G×E) interactions within pathways and have also combined them with the bootstrap approach proposed by Buzkova et al.[2011] for testing the association of each pathway. Further, we have modified the Bayesian hierarchical framework proposed by Evangelou et al.[2012], which outperforms the Fisher's approach. The Bayesian hierarchical framework does not rely on the results on single-SNP analysis and it can simultaneously perform variable selection for both the additive and interaction effects of the pathways.

The results of the conducted simulation study will be presented as well as the application of the methods on breast cancer data.

110 | Genetic Modifiers Delay the Age at Onset of Alzheimer Disease in Carriers of the G206A Founder Mutation in PSEN1

Joseph H. Lee[1], Rong Cheng[1], Badri Vardarajan[1, 2], Rafael Lantigua[3], Dolly Reyes-Dumeyer[1], Angel Piriz[1], Martin Medrano[4], vonne Z. Jimenez-Velazquez[5], Richard Mayeux[1,2]

[1]Sergievsky Center/Taub Institute, CUMC, New York, New York, United States of America; [2]Department of Neurology, CUMC, New York, New York, United States of America; [3]Department of Medicine, CUMC, New York, New York, United States of America; [4]School of Medicine, Pontificia Universidad Catolica Madre y Maestra, Santiago, Dominican Republic; [5]Department of Internal Medicine, University of Puerto Rico School of Medicine, San Juan, Puerto Rico

To identify genetic variants that may protect individuals from developing Alzheimer Disease (AD) by delaying the Age At Onset (AAO) of AD, we studied 45 families with the G206A founder mutation in PSEN1. This study was motivated by the fact that the AAO of AD varies widely among carriers of the PSEN1 mutation, and environmental factors had little influence on the variability of AAO.

We performed a Whole Genome Sequencing (WGS) study, combining WGS data and genome-wide SNP (GWAS) data to identify variants that segregate with the AAO. Here, all examined family members had GWAS data, and 1 to 10 individuals per family had WGS data. We then imputed WGS data into GWAS data using SHAPEIT2 and IMPUTE2. To prioritize candidate regions, we first performed linkage analysis, then conducted family-based gene-wise association analysis using FAMSKAT. For the genes that were significant, we performed a SNP-wise association analysis using PSEUDOMARKER and linear mixed models.

We confirmed our earlier linkage peaks at 4q23 (lod=3.35) and 2q13 (lod=2.09). For 4q35, the gene-wise analysis identified 11 genes, but the strongest support was observed for MLF1IP, SORBS2, and LOC285441. For 2q13, SH3RF3 and AS1 were associated with AAO. In particular, SORBS2 had multiple variants that are associated with variable AAO, and some of the variants affected transcription factor binding.

This study confirmed and further identified variants that are significantly associated with delayed AAO in mutation carriers. Further investigation of these modifiers may provide insight into the pathobiology of AD and potential therapeutic measures.

111 | Family-Based Rare Variant Association Study of Familial Myopia in Amish and Ashkenazi Jewish families

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

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America, [2]Genetics, Genomics and Informatics, University of Tennessee Health Sciences Center, Memphis, Tennessee, United States of America; [3]University of Texas at El Paso, El Paso Texas, United States of America; [4]Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

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 because 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) that were suggestively associated ($p<1.4\times10^{-4}$) 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.

112 | Rewiring of Enhancer-Gene Interactions Drives PLAU Overexpression in the Pathogenesis of Quebec Platelet Disorder

Minggao Liang[1,5], Asim Soomro[7], Subia Tasneem[7], John S. Waye[3,7], Andrew D. Paterson[1,2], Georges E. Rivard[4], Catherine P.M. Hayward[3,6,7]Michael D. Wilson[1,5]

[1]Department of Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada; [2]The Dalla Lana School of Public Health and Institute of Medical Sciences, University of Toronto, Toronto, Canada; [3]Hamilton Regional Laboratory Medicine Program, Hamilton, Canada;

[4]Hematology/Oncology, Centre Hospitalier Universitaire Sainte-Justine, Montreal, Canada; [5]Department of Molecular Genetics, University of Toronto, Toronto, Canada; [6]Department of Medicine, McMaster University, Hamilton, Canada; [7]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada

Determining the molecular mechanism of disease-linked genetic variation represents an important yet challenging step in bridging epidemiological discoveries with clinical intervention. Our study applies high-throughput genomics strategies to gain insight into the molecular mechanism of Quebec Platelet Disorder (QPD), a rare inherited bleeding disorder for which the causal mutation has recently been described. The hallmark of QPD is a marked overexpression of urokinase-type plasminogen activator (PLAU) in platelets and megakaryocytes, resulting in a unique gain-of-function defect in fibrinolysis. The genetic cause of QPD is a heterozygous tandem duplication of a 77 kb region of 10q22 that includes PLAU and its putative regulatory elements. However, the markedly increased (>100 fold) and cell-type specific expression of PLAU in QPD is not fully explained by a copy-number gain. To establish the molecular mechanism of PLAU overexpression in QPD, we performed transcriptomic (mRNA-seq) and epigenomic (ChIP-seq for histone modifications) profiling of blood cells (granulocytes and cultured megakaryocytes) from QPD patients and controls. We identified a putative enhancer ∼40 kb downstream of PLAU that is highly enriched for the histone modification H3K27ac in megakaryocytes compared to granulocytes. Chromosome conformation capture experiments support that this enhancer interacts with the promoter of the downstream vinculin (VCL) gene. VCL is upregulated during megakaryocyte differentiation and the QPD duplication places one copy of PLAU downstream of this enhancer. We propose that the positioning of PLAU downstream this enhancer results in aberrant enhancer-gene interactions that in turn drive PLAU overexpression in QPD megakaryocytes.

113 | Joint Analysis of Multiple Phenotypes in Association Studies Using K-Means Clustering Approach

Xiaoyu Liang[1], Qiuying Sha[1], Shuanglin Zhang[1]

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

In the study of complex diseases, several correlated phenotypes are usually measured for a disorder or its risk factor. There is also an increasing evidence showing that testing pleiotropic association between the Single Nucleotide Polymorphism (SNP) and multiple-dependent phenotypes jointly is often more powerful than analyzing only one phenotype at a time. Therefore, developing statistical methods to test for genetic association with multiple traits has become increasingly important. In this paper, we develop a reversed-discriminant analysis by using a K-means clustering approach for joint analysis of multiple phenotypes in association studies. In the K-means clustering method, we consider a genetic variant of interest as a responding variable with three classes, and the correlated phenotypes as predictors. The proposed K-means clustering method uses a data driven approach to select phenotypes that are associated with the genetic variant and uses these phenotypes to predict the corresponding class of genotypes by using the K-means clustering approach. The test statistic of the K-means clustering method is the prediction accuracy evaluated by the generalized cross-validation. We perform extensive simulation studies to evaluate the performance of the K-means clustering method and compare the power of K-means clustering method with the powers of Trait-based Association Test that uses Extended Simes procedure (TATES), SUM-SCORE based on univariate score test statistics, and the standard MANOVA.

114 | Modified Random Forest (RF) for Trio Data with Alternative Splitting Criterion to Allow for Missing Genotypes

Qing Li[1], Emily Holzinger[1], Mary L. Marazita[2], Terri H. Beaty[3], Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Baltimore, Maryland, United States of America; [2]Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America

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

115 | Multivariate Genetic Risk Scores Can Increase Risk Prediction Accuracy for a wide Range of Traits

Robert Maier[1], Matt Robinson[2], Naomi Wray[1,2], Peter Visscher[1,2]

[1]The Queensland Brain Institute, The University of Queensland, St Lucia, Australia; [2]Institute for Molecular Bioscience, The University of Queensland, St Lucia, Australia

Polygenic risk prediction has emerged as a powerful research tool in human genetics in the past few years. Although prediction accuracy has an upper bound that is set by the heritability captured by genetic markers, current polygenic risk scores achieve far from this value. Recent studies show that pleiotropy is widespread, with evidence for genetic correlations between even seemingly unrelated diseases. This provides an opportunity to improve the accuracy of polygenic risk predictors by appropriately weighting information from independent but correlated data sets.

We have previously shown that the joint analysis of five genetically correlated psychiatric traits can lead to an effective increase in sample size and thus to increased accuracy of a genetic predictor. The method we developed required full individual-level genotype data for all traits. Here, we present a novel computationally inexpensive method for multivariate risk prediction, which only requires GWAS summary statistics.

We show through theory, simulation and application to a wide range of traits traits that our new method can yield similar increases in prediction accuracy as an analysis where full individual-level data are available.

Our results indicate that prediction accuracy of a multi-trait predictor can be higher than that of a single-trait predictor, especially if the additional traits in the multi-trait predictor have high heritability, large sample size and a high genetic correlation with the predicted trait.

116 | Integration of Whole Genome Sequence and Epigenomic Data Highlights Regulatory Activity in the Genomic Architecture of Glycemic Traits

Alisa K. Manning[1], Samantha Lent[2], Michelle Jones[3]; On Behalf of CHARGE, GOT2D/T2D-GENES

[1]Center for Human Genetics Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [2]School of Public Health, Boston University, Boston, Massachusetts, United States of America; [3]Cedars-Sinai Medical Center, Los Angeles, California, United States of America

Studies of common genetic variants have identified many loci associated with Type 2 Diabetes (T2D) and glycemic traits that fall outside of protein-coding regions, implicating gene regulation as the causal mechanism. We hypothesize that by integrating diabetes physiology in association tests for Whole Genome Sequence (WGS) data, we will reveal novel molecular mechanisms that lead T2D and insulin resistance. We analyzed 69 previously described loci with WGS data from 2,854 European, non-diabetic individuals from 3 CHARGE cohorts, with targeted replication in 1400 non-diabetic individuals from GoT2D. We used inverse-normalized covariate-adjusted Fasting Glucose (FG) and log-transformed Fasting Insulin (FI) levels, also accounting for possible interaction by obesity. We aimed to (1) fine-map common variant associations, (2) discover independent common and rare genetic associations, and (3) test for rare variant associations within active regulatory elements in pancreas, liver, adipose and skeletal muscle tissue. We found 267 loci with nominal ($P<0.0001$) single variant associations with FG and/or FI levels, within which 34 independent variant associations were observed (14 common, 6 low-frequency and 14 rare). Only 2 of the common variants replicated with $P<0.01$ in GoT2D. In rare variant tests within marks of active transcription, 25 loci showed nominal ($P<0.001$) associations with either the same trait as previously reported for the locus or both FG and FI. Replication in GoT2D showed rare variant associations of non-coding variants (SKAT $P<0.01$) at the *SLC10A66*, *ADCY5* and *PCSK1* loci (with epigenomic peaks observed in multiple tissues.) This study increased understanding of the genomic architecture of glycemic traits.

117 | Missense Single Nucleotide Polymorphisms on Exon 1 of the PKD1L2 Gene from Isolated B Lymphocytes in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis Patients

Sonya Marshall-Gradisnik[1,2], Samantha Johnston[1,2], Lavinia Gordon[3], Donald Staines[2]

[1]School of Medical Science, Griffith University, Gold Coast, Australia; [2]The National Centre for Neuroimmunology and Emerging Diseases, Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia; [3]Australian Genome Research Facility Ltd, The Walter and Eliza Hall Institute, Parkville, Australia

The underlying pathomechanism of Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (CFS/ME) has been associated with immunological dysfunction, including changes in B cell phenotypes and calcium signaling. We previously found significant Single Nucleotide Polymorphisms (SNPs) for Transient Receptor Potential ion channels (TRPs) from Peripheral Blood Mononuclear Cells (PBMCs), isolated B cells and NK cells from CFS/ME patients. This study aimed to determine the presence of exome SNPs in TRPs from isolated B cells from CFS/ME patients.

Flow cytometric protocols were used to determine B cell purity, followed by exome analysis from 21 mammalian TRP ion channel genes using Illumina HiSeq platform and PLINK analysis software for exome SNP association. CFS/ME patients (age=45.89±5.88 years) defined according to the Fukuda criteria and healthy controls (age=45.89±5.88 years) were included. 237 SNPs were identified for TRP ion channel. Two missense SNPs, located on exome 1 for the subunit TRPP respectively were significantly associated with CFS/ME patients compared with healthy controls. A synonymous SNP for TRPC3, located on exome 1 was also found to be significantly associated with CFS/ME patients compared with healthy controls. Three genotypes were identified from SNPs that were reported significant for TRPC3, and two for PKD1L2 (ORs 10.5–17.5) for CFS/ME patients compared with healthy controls.

This preliminary investigation identified SNPs and genotypes for TRP ion channels from isolated B cells in patients with CFS/ME that may be involved in changes in B cell function and warrant further examination.

118 | Genome-Wide Association Study of Duloxetine and Placebo Response in Major Depressive Disorder

Victoria S. Marshe[1,2], Malgorzata Maciukiewicz[1], Arun K. Tiwari[1,3], Trehani M. Fonseka[1,4,5], Natalie Freeman[1], Susan Rotzinger[3,4], Jane A. Foster[4,6], James L. Kennedy[1,2,3], Sidney H. Kennedy[3,4,5], Daniel J. Mueller[1,2,3]

[1] Pharmacogenetics Research Clinic, Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada; [2] Institute of Medical Science, Faculty of Medicine, University of Toronto, Toronto, Canada; [3] Department of Psychiatry, University of Toronto, Toronto, Canada; [4] University Health Network, Toronto, Canada; [5] Department of Psychiatry, St. Michael's Hospital, Toronto, Canada; [6] Department of Psychiatry and Behavioral Neurosciences, McMaster University, Hamilton, Canada

Major depressive disorder is a prevalent psychiatric disorder treated with antidepressants, such as duloxetine. Given the genetic contribution to antidepressant response and possibly placebo, identifying a 'genetic signature' for response will help to implement personalized treatment.

We performed a Genome-Wide Association Study (GWAS) in depressed patients treated with either duloxetine (N=215) or placebo (N=235) for up to 8 weeks. Individuals were genotyped using the Illumina PsychChip, and genetic data was imputed to >2 million variants per individual. We conducted standard quality controls for individuals and markers. Treatment response was operationalized as MADRS score change (%) from baseline and was the main outcome variable in an ANCOVA model, including baseline depression severity, length of treatment and cohort as covariates.

For duloxetine response, we observed top hits in regions on chromosome 1, 7 and 19 implicating previously unnoted intergenic variants; however, none of the findings reached genome-wide significance (p<10^{-6}). For placebo response, there was a significant hit on chromosome 3 (p=1.87×10^{-9}) located 150kb from the *STAC1* gene, implicated in neuron-specific signal transduction, and expressed in nociceptive (pain processing) neurons. Carriers of the CC genotype improved on average by 49.6% of MADRS score while non-carriers only improved by 23.9% – a clinically relevant difference. Furthermore, there was a suggestive association (p<10^{-6}) with a marker located in the *TPO* gene involved in thyroid functioning.

Our data provide new insights into genetic pathways implicated in response to antidepressant and in particular to placebo medication, as well as, do not support the notion that similar pathways are involved.

119 | Admixture Mapping in Two Mexican Samples Identifies Significant Associations of Locus Ancestry with Triglyceride Levels in the *ZNF259/APOA5* Region

Andrew Mazurek[1], Christopher Gignoux[2], Alexandra A. Sockell[3], Michael Agostino[1], Andrew Morris[4], Lauren E. Petty[5], Craig L. Hanis[5], Nancy J. Cox[6], Adan Valladares-Salgado[7], Jennifer E. Below[5], Miguel Cruz[7], Esteban J. Parra[1]

[1] Department of Anthropology, University of Toronto at Mississauga, Mississauga, Canada; [2] Department of Genetics, Stanford University, Stanford, California, United States of America; [3] School of Medicine, Stanford University, Stanford, California, United States of America; [4] Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; [5] Department of epidemiology, Human Genetics & Environmental Sciences, University of Texas School of Public Health, Houston, Texas, United States of America; [6] Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America; [7] Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, IMSS, Mexico City, Mexico

We carried out an admixture mapping study of lipid traits in two samples from Mexico City. Native American locus ancestry was significantly associated with triglyceride levels in a broad region of chromosome 11 overlapping the *ZNF259* and *APOA5* genes. Fine-mapping of this region using dense genome-wide data points to the variant rs964184 as the main driver of the association signal: rs964184 is the only marker included in the 99% credible set of SNPs. The frequency of the allele associated with increased triglyceride concentrations (rs964184-G) is between 30–40% higher in Native American populations from Mexico than in European populations. The evidence currently available for this variant indicates that it may be exerting its effect through three potential mechanisms: 1) modification of enhancer activity, 2) regulation of the expression of several genes in *cis* and/or *trans*, or 3) modification of the methylation patterns of the promoter of the *APOA5* gene.

120 | Association Analyses of Myopia in Multiplex African American Families using FBAT and a Rare Variant FBAT with Exome Chip Data

Candace D. Middlebrooks[1], Claire L. Simpson[2], Anthony M. Musolf[1], Laura Portas[3], Federico Murgia[4], Dwight Stambolian[5], Joan E. Bailey-Wilson[1]

[1]National Human Genome Research Institute, Baltimore, Maryland, United States of America; [2]University of Tennessee Health Science Center, Memphis, Tennessee, United States of America; [3]Università degli Studi di Verona, Cagliari Area, Italy; [4]Institute of Population Genetics - National Research council of Italy, Cagliari Area, Italy; [5]Scheie Eye Institute - Penn Presbyterian Medical Center, Philadelphia, Pennsylvania, United States of America

Myopia is an eye condition in which the light entering the eye does not focus on the retina resulting in distant objects appearing out of focus. Within recent years, the incidence and prevalence of myopia have increased, reaching epidemic proportions in several Asian countries. Populations of African descent have lower rates of myopia than Asian populations, but multiplex families from the African American (AA) population may represent an opportunity to identify unique susceptibility and protective alleles.

We have performed an association study examining common and rare variants using Exome Chip genotyping (Illumina HumanExome v1.1 array plus 24,263 custom SNPs) within a family - study framework in 106 African American families from the Philadelphia area who have multiple individuals affected with myopia. Individuals in the families were defined as myopic if their average refractive error was <=−1 Diopter (D) and were considered unaffected if their average refractive error was >0.0 D. After quality control, there were 242,901 SNPs available for analysis. Single-variant analyses using Family Based Association Test (FBAT) resulted in a suggestive signal in African Americans within the psoriasis susceptibility 1 candidate 3 gene (*PSORS1C3*) (rs887468, *P*=1.2E×05) on chromosome 6. This association signal is intriguing but the analysis of each genetic variant individually does not adequately make use of the many rare variants available in this dataset. To increase power, we are performing gene-based tests using a rare variant analysis approach within the FBAT software suite. Using this analysis, we expect to increase our power to identify an association

121 | Exploring the Heritability of Pharmacogene Expression

Sabrina L. Mitchell[1], Eric R. Gamazon[1], Sara P. Hillenmeyer[2], Russ B. Altman[3,4], Nancy J. Cox[1], Lea K. Davis[1]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Program in Biomedical Informatics, Stanford University, Stanford, California, United States of America; [3]Departments of Bioengineering and Genetics, Stanford University, Stanford, California, United States of America; [4]Department of Medicine, Stanford University, Stanford, California, United States of America

Pharmacogenes encompass genes whose products function in drug metabolism. Variation in the expression of pharmacogenes impacts the metabolism of drugs and other substances. Recent studies indicate that pharmacogenes exhibit high inter-individual variability in expression. The observed variation in gene expression can be attributed to both genetic (heritable) and environmental factors. To explore the heritability of pharmacogene expression we leveraged the gene expression and genotype data available in the Genotype-Tissue Expression (GTEx) Project, giving us the ability to examine seven relevant human tissues. The mean expression was calculated for each gene within an individual tissue. Only those genes meeting a minimal threshold of expression (mean RPKM > 0.1) were included in the analysis for a given tissue. In our preliminary analysis we examined the heritability of gene expression in adipose (subcutaneous; *n*=350), brain (hypothalamus; *n*=96), and liver (*n*=119). Heritability of gene expression was calculated for genes in each tissue via GCTA. The Wilcoxon rank sum test was used to compare heritability of gene expression of the PGRN pharmacogenes (PMID: 26856248) and their complement in adipose (*p*=0.11), brain (*p*=0.055), and liver (*p*=0.055). While none of these analyses exceeded the statistical threshold indicating a significant difference in heritability of gene expression between the two gene sets, results from brain and liver demonstrated nominal significance. We are currently exploring the heritability of pharmacogene expression across multiple other tissues available in GTEx. Results from these studies will help illuminate our understanding of the regulation and genetic architecture of pharmacogenes across tissues.

122 | IQML: A Robust Statistical Approach for Isoform Level Quantification from RNA-Seq Data

Pronoy Kanti Mondal[1], Raghunath Chatterjee[1], Indranil Mukhopadhyay[1]

[1]Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Technological advances trigger the generation of massively parallel genome-wide transcriptome data, commonly known as RNA-Sequencing. It enables us to measure the transcription profiles with high precision and accuracy at genome-wide scale. However, the key challenges in quantification of transcripts appear mainly due to problems in comprehending the distribution of reads and ambiguity in mapping to proper isoforms. This leads to problems in modelling and estimation of transcript abundance. Thus estimation of isoform level abundance using millions of reads is still a challenge that needs to be addressed under very realistic conditions of the experiment. Moreover, the reads that hold key information to this estimation problem may be of different types for single-end or pair-end reads.

In order to estimate this abundance, we have developed a statistical method for Isoform level Quantification based on Maximum Likelihood (IQML) under very general and weak conditions of the nature and distribution of reads. We adopt EM algorithm to finally obtain the estimates. Among other features, our proposed IQML method is also robust to the assumption of the probability distribution of reads, thus requiring no prior knowledge about the distribution of reads, which has immense significance in this whole problem.

We have studied our method extensively using simulated as well as real data. Our method shows promising result in comparison to the performance of other available methods. Moreover, this method is scalable with respect to memory allocation and computationally very fast thus making it extremely useful and feasible approach in practical implementation with real data.

123 | Field Synopsis of Genetic Variation and Colorectal Cancer; Unified 2016 Update

Zahra Montazeri[1], Julian Little[1], Christine Nyiraneza[1], Evropi Theodoratou[2], Wei Zheng[3]

[1] School of Epidemiology, Public Health and Preventive Medicine Faculty of Medicine, University of Ottawa, Ottawa, Canada; [2] Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, United Kingdom; [3] Vanderbilt Epidemiology Center Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

According to current predictions of global demographic changes, the World Health Organization (WHO) estimates a 77% increase in Colorectal Cancer (CRC) incidence and 80% increase in CRC-related deaths by 2030 [Bray F, et al: “Global Cancer Transitions According to the Human Development Index (2008–2030): A Population-Based Study.” The Lancet Oncology 2012; 13(8):790–801. doi:10.1016/S1470-2045(12)70211]. Symptoms are less obvious early in the disease development and become commonly noticeable when prognosis is unfavorable, but identifying genetic variants that influence susceptibility to disease could be a powerful approach for prevention. We had completed a field synopsis for CRC in 2010 and results published in JNCI. Now we are in progress of conducting a more comprehensive field synopsis with aim to catalogue all published genetic association studies on CRC up to end of 2015. We considered all published (and some unpublished) genetic association data, included candidate gene and Genome-wide Association Studies (GWAS), for CRC and conducted meta-analyses to summarize risk estimates to determine the direction and magnitude of association between polymorphic genetic variants and CRC.

Based on an exhaustive search more than 200 genes in about 770 Single Nucleotide Polymorphisms (SNPs) were identified that associated with CRC. There was one study for 450 SNPs, two studies for 90 SNPs and 3 or more studies for 230 SNPs. Furthermore, we have access to 4 GWAS data. Meta-analysis will be carried out for SNPs with 3 or more studies, and variants will be considered as “highly credible” and “less credible” according to results. We consider several criteria in our analysis. The identification of genetic variants with influence on CRC risk may reflect an importance of genes involved in colorectal cancer risk. Our data should help results of genetic associations studies to be placed in context and interpreted appropriately and should help direct future research effort.

124 | False Positive Rate Inflation and Admixed Populations in the Quantitative Transmission Disequilibrium Test

Anthony M. Musolf[1], Joan E. Bailey-Wilson[1]

[1] National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America

The Transmission Disequilibrium Test (TDT) is a well-known statistical test for the analysis of trio data that functions as both a test of linkage and a test of association and is robust to the problem of population stratification. TDT has been extended to quantitative phenotypes and this study aims to examine the robustness of the quantitative TDT when phenotypic means differ across populations.

Genome-wide trio data was simulated for two distinct populations using HapMap allele frequencies. Phenotype data was simulated from a normal distribution and assigned to individuals at random. In the equal-mean scenario, data was simulated for both populations using the same mean. In the different-mean scenario, data for each population was drawn same distribution but using different means. Type I error rates were calculated after the populations were analyzed by qTDT both alone and mixed. We have found that the qTDT statistic maintains proper type I error when the phenotypic means are equal, but results in inflation if the means differ across populations. Larger differences between the means of the populations lead to higher levels of type I inflation. This is a critical deficiency in this test, since inflation of false positive rates in the presence of population admixture is the main reason that TDT tests are utilized. We have developed several corrections that restore proper type I error levels. Analyses to determine how these corrections affect power are ongoing.

125 | Prenatal Smoke Exposure Alters Mitochondrial DNA Methylation in Umbilical Cord Blood Dendritic Cells

Michelle L. North[1], Cheng Peng[2], Marco Sanchez-Guerra[2], Hyang-Min Byun[3], Jeff Brook[1,4], Anne K. Ellis[5], Andrea A. Baccarelli[2]

[1]Southern Ontario Centre for Atmospheric Aerosol Research, University of Toronto, Toronto, Canada; [2]Laboratory of Environmental Epigenetics, Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [3]Human Nutrition Research Centre, Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; [4]Air Quality Research Division, Environment Canada, Toronto, Canada; [5]Department of Medicine, Queen's University, Allergy Research Unit, Kingston General Hospital, Kingston, Canada

DNA methylation is related to perinatal factors, such as maternal smoking. Human cells contain nuclear and mitochondrial DNA (mtDNA), but the impact of mitochondrial epigenetics on development has not been adequately explored. Dendritic Cell (DC) mitochondria were targeted because of their sensitivity to oxidative stress. We tested the hypothesis that prenatal smoke exposure would be associated with mitochondrial epigenetic differences in DCs in umbilical Cord Blood (CB).

CB DCs were isolated from a cohort with a known high prevalence of maternal smoking using magnetic sorting (n=91). We analyzed mtDNA regions including the D-loop promoter, transfer RNA phenylalanine (MTTF), and 12S ribosomal RNA (MT-RNR1) by pyrosequencing. Copy Number (CN) was determined using qPCR. CN and mtDNA were analyzed for associations with perinatal factors using models adjusted for maternal age, pre-pregnancy BMI, ethnicity, child's gender, and Socioeconomic Status (SES).

Prenatal smoke exposure was associated with a 1.53% (95% CI: 0.61 - 2.46%, p=0.002, adjusted model) increase in mtDNA methylation in MTTF. The D-loop region in DC mitochondria also demonstrated an increase in methylation associated with maternal smoke exposure during pregnancy, of 3.82% (95% CI: 0.50 - 7.14%, p=0.03, adjusted model). However, we did not observe significant associations between mtDNA methylation and gestational weight gain, maternal allergy or C-section delivery. MtDNA CN was also not associated with perinatal risk factors.

Maternal smoking was associated with differences in CBDC mitochondrial DNA methylation. This may be related to the known effects of smoking on oxidative stress balance, and may affect innate immunity.

126 | Polymorphisms in the NADPH Oxidase Complex are Associated with Hepatitis C-induced Fibrosis and Inflammation

Sandra J. Page[1], Maria Rivera[1], David E. Kleiner[2], Xiongce Zhao[3], Sungyoung Auh[4], Elaine F. Remmers[5], Theo Heller[1]

[1]Translational Hepatology Unit, Liver Diseases Branch, NIDDK, NIH, Bethesda, Maryland, United States of America; [2]Laboratory of Pathology, NCI, NIH, Bethesda, Maryland, United States of America; [3]Center for Veterinary Medicine, FDA, Rockville, Maryland, United States of America; [4]Office of Clinical Director, Liver Diseases Branch, NIDDK, NIH, Bethesda, Maryland, United States of America; [5]Inflammatory Disease Section, NHGRI, NIH, Bethesda, Maryland, United States of America

Infection with the Hepatitis C virus (HCV) causes hepatitis, potentially leading to cirrhosis or hepatocellular carcinoma (HCC). Over 150 million people are chronically infected with HCV, about a half million people die annually from HCV-related complications. HCV is an enveloped virus with a positive-sense RNA genome encoding 10 structural and non-structural proteins. HCV proteins induce oxidative stress by stimulating NADPH oxidases (Nox). Previous studies have shown that Nox expression coincides with hepatic fibrosis. We hypothesized that polymorphisms in Nox family members may affect the progression of HCV. We genotyped 359, HCV-infected patients (288 Caucasians; 71 Africans) for HapMap-defined tagSNPs in 14 genes encoding Nox enzymes or their regulatory proteins. After dividing the cohort by race and performing quality control steps, we compared SNP composition with each of the biopsy-derived measures of inflammation and fibrosis. Fisher Exact Tests with FDR correction for multiple comparisons were used. In the African cohort, rs12753665 (*NCF2,* neutrophil cytosolic factor 2, a subunit of Nox2) was associated with the highest Ishak fibrosis scores achieved over the course of treatment (FDR p=0.02), while rs760519 (*NCF4*, neutrophil cytosolic factor 4, another subunit of Nox2) was associated with the highest Ishak scores at the onset of treatment (FDR p=0.02). In the Caucasian cohort, rs2292464 (*DUOX1,* dual oxidase 1, a H_2O_2-producing Nox) was significantly associated with periportal inflammation (FDR p=0.04). Although exploratory, this study suggests that genetic links between HCV-associated liver disease and Nox family members may exist.

127 | Multivariate Analysis of Anthropometric Traits using Summary Statistics of Genome-Wide Association Studies from GIANT Consortium

Haeil Park[1,2], Xiaoyin Li[1], Yeunjoo E. Song[1], Karen Y. He[1], Xiaofeng Zhu[1]

[1]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America; [2]Department of Laboratory Medicine, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

Despite the hundreds of variants that have been detected by GWAS, those variants only explain small fractions of phenotypic variations. Cross-Phenotype Association Analysis (CPASSOC) can further improve statistical power by identifying variants that contribute to the variation of multiple traits, which is often relevant to pleiotropy. In this study, we performed CPASSOC analysis on the summary statistics from the Genetic Investigation of ANthropometric Traits (GIANT) consortium using a novel method recently developed by our group. Sex-specific meta-analysis data for height, Body Mass Index (BMI), and Waist-to-Hip Ratio Adjusted for BMI (WHRadjBMI) from discovery phase of the GIANT consortium study were combined using CPASSOC for each trait

as well as 3 traits together. The conventional meta-analysis results from the discovery phase data of GIANT consortium studies were used to compare with that from CPASSOC analysis. Our result showed CPASSOC is able to identify variants missed by conventional meta-analysis. When combining male and female data, a total number of 10 variants were significantly associated with either height, BMI, or WHRadjBMI using CPASSOC but missed by the conventional meta-analysis. When combining those three related traits together by CPASSOC, we identified 7 significant loci missed by the conventional meta-analysis in GIANT discovery stage. Additionally, several gene sets that were not enriched from the result of conventional meta-analysis for individual trait were identified using CPASSOC by combining the three traits. Our result suggests that CPASSOC can identify loci that may be missed by single trait meta-analysis and provide novel biological insight by analyzing multiple traits together.

128 | Genome-Wide Association Analysis of Time-to-Metastasis of Colorectal Cancer Based on Mixture Cure Model

Michelle E. Penney[1], Yildiz E. Yilmaz[1,2,3], Jane S. Green[1,3,4], Patrick S. Parfrey[3], Sevtap Savas[1,4]

[1]Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada; [2]Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Canada; [3]Discipline of Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada; [4]Discipline of Oncology, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

Differentiating between colorectal cancer patients who will be long-term survivors without metastasis and who will experience metastasis within a short time is a critical aim in healthcare. This study aimed to identify Single Nucleotide Polymorphisms (SNPs) associated with the risk of experiencing metastasis and time-to-metastasis in patients who experience metastasis after diagnosis. We focus our efforts on genome-wide SNP genotype data (810,622 SNPs) obtained from a sub-cohort of colorectal cancer patients participating in the Newfoundland Colorectal Cancer Registry. The patient cohort consists of long-term metastasis-free survivors and patients who experienced metastasis after diagnosis. These two groups can be analyzed separately but simultaneously using the Mixture Cure (MC) model. We applied the univariable MC model under different genetic models to investigate the effect of each SNP independently on the probability of being a long-term metastasis-free survivor and on the time-to-metastasis in patients who experienced metastasis after diagnosis. We detected a number of SNPs significantly associated with early metastasis ($p<6e-8$). We also discovered SNPs whose specific genotypes were not detected in patients who developed metastases. This is the first study to apply the MC model to such an extensive project and the first study to investigate genetic associations with time-to-metastasis in colorectal cancer patients using such a large genetic data set. Our results suggest that the MC model can provide novel insight into outcome-genetic variation relationships. The identified associations, once replicated in multivariable models and in other patient cohorts, could assist in the development of personalized treatment strategies for colorectal cancer patients.

129 | Differential Shrinkage as a Way of Integrating Prior Knowledge in a Bayesian Model to Improve the Analysis of Genetic Association Studies

Miguel Pereira[1], John R. Thompson[2], Christian X. Weichenberger[3], Duncan C. Thomas[4], Cosetta Minelli[1]

[1]National Heart and Lung Institute - Imperial College London, London, United Kingdom; [2]University of Leicester, Leicester, United Kingdom; [3]EURAC, Bolzano, Italy; [4]Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America

We propose a method of integrating external biological information about SNPs in a Bayesian hierarchical shrinkage model that jointly estimates SNP effects with the aim of increasing the power to detect variants in genetic association studies. Our method induces shrinkage on the SNP effects that is inversely proportional to prior information: SNPs with more information are subject to little shrinkage and more likely to be detected, while SNPs without prior information are strongly shrunk towards zero (no effect).

The performance of the method was tested in a simulation study with 1000 datasets, each with 500 subjects and ∼1200 SNPs, divided in 10 Linkage Disequilibrium (LD) blocks. One LD block was simulated to be truly associated with the outcome. The method was further tested on an empirical example using BMI as the outcome and data from the European Community Respiratory Health Survey: 1,829 subjects and 2,614 SNPs from 30 blocks, 6 of which known to be truly associated with BMI. Prior knowledge was retrieved using the bioinformatic tool Dintor and incorporated in the model.

The Bayesian model with inclusion of prior information outperformed the classical analysis. In the simulation study, the mean ranking of the true LD block was 2.8 for the Bayesian model vs. 3.6 for the classical analysis. Similarly, the mean ranking of the six true blocks in the empirical example was 8.3 vs. 11.7 in the classical analysis. These results suggest that our method represents a more powerful approach to detect new variants in genetic association studies.

130 | A Unified Association test for the Meta-Analysis of Multiple Traits using GWAS Summary Statistics

Debashree Ray[1], Michael Boehnke[1]

[1] *Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America*

Genome-Wide Association Studies (GWAS) for complex diseases have primarily focused on the univariate analysis of each trait characterizing the disease. For example, GWAS on risk factors for coronary artery disease analyze genetic associations of plasma lipids such as total cholesterol, low-density lipoprotein, high-density lipoprotein, and triglycerides separately. However, traits underlying a disease are often correlated and a joint analysis may yield improved statistical power for association over multiple univariate analyses. We propose a unified association test of a single genetic variant with multiple traits that utilizes univariate GWAS summary statistics. This novel test does not require individual-level data, and uses only publicly available summary statistics from existing GWAS to test genetic associations of categorical and/or continuous traits. One can also use this test to analyze a single trait over multiple studies with overlapping samples. The software for the proposed test reports an approximate asymptotic P value for association and is computationally efficient for implementation at a GWAS level. Our simulation experiments show that our method can maintain proper type-I error at low error levels. It has appreciable statistical power across a wide array of association scenarios (which is unknown apriori for real data), while existing methods have widely varying power curves. When applied to plasma lipids summary data of Teslovich et al (Nature, 2010), our test detected significant genetic variants beyond the ones identified by existing tests. In summary, the proposed method can potentially provide novel insights into the genetic underpinnings of a disease.

131 | Homozygosity and Health-Related Phenotypes in an Asthma Cohort

Marie-Hélène Roy-Gagnon[1], Kelly M. Burkett[2], Jean-François Lefebvre[1], Hélène Vézina[3], Catherine Laprise[4], Emmanuel Milot[5]

[1] *School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada;* [2] *Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada;* [3] *Department of Human and Social Sciences, Université du Québec à Chicoutimi, Chicoutimi, Canada;* [4] *Department of Fundamental Sciences, Université du Québec à Chicoutimi, Chicoutimi, Canada;* [5] *Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, Canada*

Animal studies have shown that non-additive genetic variation (underlying homozygosity effects) can be as important as additive genetic variation in shaping complex traits, but the latter is most often studied. Thus, the health impact of increased homozygosity levels on complex traits is not well understood. Large founder populations such as Quebec allow the measurement of a wide range of homozygosity levels, necessary to efficiently study homozygosity and its impact on health. In this study, we used genome-wide genotypic data and deep-rooted genealogies (7608 founders in 18 generations) from the ongoing Saguenay–Lac-Saint-Jean (SLSJ) asthma study, including 1394 participants, to quantify homozygosity and relate it to health phenotypes. We determined Runs Of Homozygosity (ROHs) from genotypic data using PLINK, estimated inbreeding coefficients empirically from these ROHs and compared them to genealogical inbreeding. We then used linear mixed models to investigate the association between homozygosity levels and asthma-related phenotypes. Mean length of ROHs >1500kb was 3416kb (standard deviation SD=3842), ranging from 1500 to 79424kb. Mean empirical inbreeding was 1.06% (SD=0.86), ranging from 0 to 7.1%, and correlated well with genealogical inbreeding (intraclass correlation=0.6, increasing to 0.7 when considering ROHs>2500kb). The proportion of ROHs varied across the genome, with a peak at the MHC region as previously reported. We found a negative association between empirical inbreeding estimates and blood concentration of neutrophils in the adult SLSJ participants, although not reaching statistical significance (p=0.065). We also investigated homozygosity locally along the genome. Our research provides insights on non-additive genetic effects.

132 | A Comparison of Genetic Risk Prediction and Subtyping for Generalized Vitiligo

Stephanie A. Santorico[1,2,3], Subrata Paul[1], Daniel Yorgov[1], Ying Jin[2,4], Tracey Ferrara[2], Richard A. Spritz[2,4]

[1] *Mathematical and Statistical Sciences, University of Colorado, Denver, Colorado, United States of America;* [2] *Human Medical Genetics and Genomics Program, University of Colorado School of Medicine, Aurora, Colorado, United States of America;* [3] *Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado, Aurora, Colorado, United States of America;* [4] *Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado, United States of America*

Generalized Vitiligo (GV) is an autoimmune disease resulting in melanocyte destruction. Previously, we discovered and replicated 46 autosomal GV loci via meta-analysis of three genome-wide association studies, with independent replication, consisting of a total of 4,680 vitiligo cases and 39,586 controls of European (EUR) ancestry with 8,801,562 genotyped variants. Loci mostly encode immune regulators and apoptotic regulators, some of which are also associated with other autoimmune diseases, as well as several melanocyte regulators. Here, we present a “polygenic risk score” for vitiligo, optimizing the set of variants for inclusion, and compared the predictive ability to a “major loci risk score" based on our 46 replicated autosomal GV loci. The polygenic risk score is built from estimated logistic regression coefficients, after adjustment for ancestry. Variants were included in risk scores for levels of significance, α, in $\{5\times10^{-8}$, $0.001,0.002\ldots0.009,0.01,0.02,\ldots,0.09,0.1\}$ after LD clumping with an index variant P value of 0.1. Performance was assessed by 10-fold cross-validation. Over this range of significance levels, the maximal Area Under the Curve (AUC) of 0.769 occurred at an inclusion α of 0.008; however, this was lower than the AUC for the major loci risk score of 0.776. In addition, recursive partitioning was applied to the 46 replicated autosomal GV loci. Six loci were found to best separate cases and controls into sub-groups. These loci included *HLA-A*, *PTPN22*, *CD44*, *GZMB*, *TYR*, and *MC1R*. Additional clustering methods are being considered and will be compared to the current results. Long term these will be correlated with vitiligo sub-phenotype data.

133 | Risk of Cardiovascular Event in Relation to Age-at-Menopause Associated Genetic Variants in the Framingham Heart Study

Chloé Sarnowski[1], Kathryn L. Lunetta[1], Joanne M. Murabito[2]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]Framingham Heart Study, Framingham, Massachusetts, Section of General Internal Medicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, United States of America

Early-menopause is associated with increased risk of cardiovascular diseases. Genome-wide association studies identified 56 independent SNPs associated with Age-at-Natural Menopause (ANM). We sought to determine the association between time-to-first-cardiovascular-event and a Genetic Risk Score (GRS) comprising the ANM decreasing alleles in the Framingham Heart Study (FHS). GRS was computed with ANM SNPs from FHS Haplotype Reference Consortium imputed data and weighted by the effect size. Cardiovascular event was defined as myocardial infarction, ischemic stroke or death from coronary heart disease. Association between GRS and time-to-first-cardiovascular-event was evaluated using Cox proportional-hazards model, accounting for familial relatedness and adjusting for sex, age, principal components and cardiovascular risk factors. A total of 7,448 FHS individuals with 325 post-baseline cardiovascular events were analyzed, including 3,331 men (242 events) and 4,117 women (283 events). We confirmed the strong association between GRS and ANM ($b=-1.2\pm0.1$, $P=1.4\times10^{-28}$). When analyzing all individuals, GRS was not associated with cardiovascular event risk (HR=1.1 [1.0-1.2], P=0.22). However, when adding GRS*sex interaction, both GRS and interaction were significant (HR_{GRS}=0.6 [0.4-0.9], P=0.005 & $HR_{GRS*sex}$=1.4 [1.2-1.7], P=0.0005). Stratification on sex showed that GRS increased cardiovascular event risk in women (HR=1.3 [1.1-1.5], P=0.001) while a non-significant opposite effect was detected in men (HR=0.9 [0.8-1.1], P=0.19). A comparable significant effect was detected in an analysis restricted to 1,756 post-menopausal women (160 events) and adjusted for ANM. Thus, this study evidenced that ANM decreasing alleles increased cardiovascular event risk in women and interacted with sex to influence cardiovascular event risk. Replication is needed to confirm this finding.

134 | Taking into Account Gene-by-Early Environmental Tobacco Smoke Exposure Interactions to Detect Genetic Variants Influencing Time-to-Asthma Onset

Chloé Sarnowski[1,2], Marie-Hélène Dizier[1,2], Raquel Granell[3], Debbie Jarvis[4,5], Markus J. Ege[6,7], Catherine Laprise[8], Pierre-Emmanuel Sugier[1,2,9], Patricia Margaritte-Jeannin[1,2], William O. C. Cookson[10], Miriam Moffatt[10], Mark Lathrop[11], Isabelle Pin[12,13,14], Erika von Mutius[6,7], Valérie Siroux[12,13,15], A. John Henderson[3], Manolis Kogevinas[16,17,18,19], Florence Demenais[1,2], Emmanuelle Bouzigon[1,2]

[1]Inserm, UMR-946, F-75010, Paris, France; [2]Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, F-75007, Paris, France; [3]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; [4]Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College, London, United Kingdom; [5]MRC-PHE Centre for Environment & Health, London, United Kingdom; [6]Dr von Hauner Children's Hospital, Ludwig Maximilian University, Munich, Germany; [7]Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research, Munich, Germany; [8]Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, Canada; [9]Université Pierre et Marie Curie, Paris, France; [10]Section of Genomic Medicine, National Heart Lung Institute, Imperial College London, London, United Kingdom; [11]McGill University and Génome Québec Innovation Centre, Montreal, Canada; [12]Université Grenoble Alpes, IAB, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, F-38000 Grenoble, France; [13]Inserm, IAB, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, F-38000 Grenoble, France; [14]CHU de Grenoble Alpes, Pediatrics, Grenoble, France; [15]CHU de Grenoble, IAB, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, F-38000 Grenoble, France; [16]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [17]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; [18]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain; [19]Universitat Pompeu Fabra, Barcelona, Spain

The number of genetic factors identified for asthma remains limited. The study of gene-by-environment interactions may facilitate the discovery of new genes. Environmental Tobacco Smoke (ETS) exposure *in utero* and/or during infancy is a known risk factor for childhood-onset and late-onset asthma. Our goal was to identify genetic variants interacting with early ETS exposure that influence Time-to-Asthma Onset (TAO). We conducted a large-scale meta-analysis of five Genome-Wide Interaction Studies (GEWIS) of TAO (totaling 3,643 exposed (ETS^+) and 5,275 non-exposed (ETS^-) subjects of European ancestry) using survival analysis methodologies. Since the power of GEWIS depends on the statistical test used according to the underlying genetic model which is unknown, two tests were performed: 1) a joint test of SNP and G×ETS interaction effects and 2) a test of G×ETS interaction alone. While the joint test identified two asthma regions (9p24 & 17q12-q21) interacting with ETS on TAO at $P\leq10^{-13}$, the interaction test revealed three potential new loci ($P<10^{-6}$): 13q21, 16p13 and 19q13. Further analysis of

9p24 and 17q12-q21 loci stratified on asthma age-of-onset (≤6yrs *versus* >6yrs) confirmed the 17q12-q21×ETS interaction in childhood-onset asthma and evidenced a complex effect of 9p24 SNP on asthma risk with: 1) a strong effect in ETS$^+$early-onset group (HR [95%CI]=1.4 [1.3-1.6]), 2) intermediate effects in both ETS$^-$early-onset and ETS$^+$later-onset groups (HR [CI]=1.2 [1.1-1.4]), and 3) no effect in ETS$^-$later-onset group. This study evidences a role of both early ETS exposure and 9p24 genetic variants in asthma risk and age-of-onset. Funding: FRSR, GABRIEL, ANR-GWIS-AM

135 | Physical Activity and DNA Methylation: An Epigenome-Wide Approach

Sergi Sayols-Baixeras[1,2], Isaac Subirana[3,1], Sebastian Torres-Cuevas[1], Carla Lluis-Ganella[1], Alberto Zamora[4,5], Alina Velescu[6,1], Jaume Marrugat[1], Stella Aslibekyan[7], Roberto Elosua[1]

[1] Cardiovascular Epidemiology and Genetics Research Group, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain; [2] Universitat Pompeu Fabra (UPF), Barcelona, Spain; [3] CIBER Epidemiology and Public Health, Barcelona, Spain; [4] TransLab Research Group, Medical Sciences Department, University of Girona, Girona, Spain; [5] Lipid and Atherosclerosis Unit, Department of Internal Medicine, Hospital de Blanes, Girona, Spain; [6] Angiology and Vascular Surgery Department, Hospital del Mar, Barcelona, Spain; [7] Department of Epidemiology, University of Alabama at Birmingham, Birmingham, Alabama, United States of America

The mechanisms of Physical Activity (PA)-induced health benefits are not yet well understood. Our aim was to identify CpGs showing differential methylation related to PA practice. We designed a cross-sectional two-stage epigenome-wide association study: i) the discovery sample included 645 individuals from the REGICOR (REgistre GIroní del COR) study, a population-based cohort in Catalonia (Spain); ii) for validation we obtained methylation and phenotype data from 2542 participants of the Framingham Offspring Study attending exam 8 through the Database of Genotypes and Phenotypes (dbGAP; http://dbgap.ncbi.nlm.nih.gov; project number #9047). DNA methylation was assessed using the Illumina HumanMethylation450 BeadChip. Light, moderate and vigorous intensity PA practice was assessed by validated questionnaires. Robust multivariable linear regression models adjusted for sex, age, smoking status, batch, estimated cell count and surrogate variables were used. We selected for validation those CpGs showing an association with PA practice with a P-value<1×10^{-5}in the discovery sample (n=30 CpGs) and a meta-analysis using the discovery and validation samples for the 30 CpGs was done. We validated 1 CpG site showing differential methylation inversely associated with moderate-vigorous PA practice (cg09349128; P-value 1.22×10^{-7}). This CpG is located in chromosome 22 in the intronic region of the gene *ALG12* and close to a non-coding RNA and the *MLC1* gene. This CpG site has been previously and directly associated with body mass index. Methylation levels in this locus could partially mediate some of the health benefits associated with PA practice.

136 | Statistical Methods for Pleiotropy: Sequential Test to Determine which Traits are Associated

Daniel J. Schaid[1]

[1] Mayo Clinic, Rochester, Minnesota, United States of America

The statistical association of a single trait with genetic data has revolutionized human genetics, with many genome-wide association studies providing guidance on genetic factors influencing human health. Pleiotropy – the association of more than one trait with a genetic marker – is believed to be common, yet current multivariate methods do not formally test pleiotropy. Current multivariate methods, such as multivariate regression of multiple traits on a genetic marker, or reverse regression of a genetic marker on multiple traits, test the null hypothesis that no traits are associated with a genetic marker; a statistically significant finding could result from only one trait driving the association. We developed a new formal test of pleiotropy, so that rejection of the null hypothesis implies at least two traits are associated with the marker. We further refined our approach to sequentially test the number of associated traits, in order to identify which traits are statistically associated, while accounting for the correlation among the traits. The new methods, with simulations illustrating it properties, will be presented, as well as application to a study of the genetics of response to small pox vaccination.

137 | A Genome-Wide Study of Gene- Fine Particulate Air Pollution Interaction Effects on Carotid Intima-Media Thickness - the Heinz Nixdorf Recall Study

André Scherag[1], Marie Henrike Geisel[1,2], Frauke Hennig[3], Barbara Hoffmann[3], Susanne Moebus[2], Raimund Erbel[2], Karl-Heinz Jöckel[2]

[1] Clinical Epidemiology, Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany; [2] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital, University of Duisburg-Essen, Essen, Germany; [3] Institute of Occupational, Social and Environmental Medicine, Center for Health and Society, University of Düsseldorf, Düsseldorf, Germany

Carotid intima media thickness (cIMT) is a marker of subclinical atherosclerosis. Risk factors for the progression of cIMT have been investigated extensively and have more recently also included genetics and particulate air pollution. Recent studies suggest that susceptibility to air pollution may vary based on different genetic profiles, however the literature on gene-air pollution interactions effects on cIMT is scarce.

Here, we conducted a genome-wide interaction study with fine particulate air pollution ($PM_{2.5}$).

Participants of the population-based Heinz Nixdorf Recall study were genotyped by genome-wide Illumina SNP arrays and subsequently imputed using the 1000 Genomes reference panel (phase1 version 3). For cIMT, sex-specific, inverse normal transformed residuals were used as quantitative traits after accounting for potential age and population stratification effects. We assumed an additive genetic model and performed a 2df-test for the genetic main effect and the interaction effect with $PM_{2.5}$. Sex- and array-specific analyses were then combined by a fixed-effect meta-analysis. We will show details of our quality control and analysis strategy and report preliminary result of 2471 participants. Consistency checks with regard to the direction of effect across data sets and validations of best findings (all loci with a suggestive association signal of $p \leq 1 \times 10^{-5}$of the 2df-test) in independent cohorts with similar exposure assessments are ongoing.

This work was supported by the German Research Council (DFG: SCHE1648/1-3, HO3314/2-3,).

138 | Genome Wide Copy Number Profiling in North Indian Gallbladder Cancer Patients

Aarti Sharma[1], Ashok Kumar[1], Neeraj Kumari[2], Narendra Krishanani[2], Balraj Mittal[3]

[1]Department of Surgical Gastroenterology, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India; [2]Department of Pathology Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India; [3]Department of Genetics, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India

Gallbladder Cancer (GBC) is a malignancy of biliary tract having an unusual geographic distribution. Recent studies have shown that molecularly targeted therapies are critically dependent on genetic profile of tumor. So understanding molecular features of GBC is critical towards improving treatment paradigm for this disease. Present study is focused on determination of Copy Number Variations (CNVs) and loss of hetrozygosity in genome wide major cancer related genes in GBC patients. 13 histopathologically confirmed GBC cases were included in this study. GBC cells were micro-dissected from formalin fixed paraffin embedded tissue blocks having >80% of tumor cells. DNA extraction was done by standard protocols. Copy number/mutation profiling was done by "OncoScan^{TM} FFPE Assay Kit" (Affymetrix, CA). Data analysis was done by SNPFASST2 algorithm using the Nexus Express for Oncoscan software version 3.0 and 7.5 (Biodiscovery, Inc., CA USA). A substantial number of CNVs detected were recurrent. Recurrent gains were at chromosome 12q (53%), 20p (69%) and recurrent losses were at 4q (46%). The common genes affected in CN loss process were NPP6, IRF2, CASP3, PRIMPOL, MLF1IP, CENPU, ACSL1, SLED1 in chromosome 4, whereas genes involved in copy number gain process were HMGA2, RPSAP52 on 12q14.3, RASSF3 gene on 12q14.2 and MACROD2 on 20p12.1. Presence of LOH was also found to be a significant character in GBC. More studies with greater sample size are required for validation of findings.

139 | Significant Role of PLCE1 and LXRs Receptor Sequence Variants in Gallbladder Cancer Predisposition: A Multi-analytical Strategy

Kiran Sharma[1], Aarti Sharma[1], Sanjeev Misra[2], Ashok Kumar[3], Balraj Mittal[1]

[1]Department of Genetics, Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGIMS), Lucknow, India (Present address: Post doc fellow, Biochemistry and medical Genetics, University of Manitoba, Winnipeg, Canada); [2]Department of Surgical Oncology, King George medical University, KGMU, Lucknow, India; [3]Department of Surgical Gastroenterology, Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGIMS), Lucknow, India

Gallbladder cancer (GBC) is a violent neoplasm associated with late diagnosis, unsatisfactory treatment and poor prognosis. The disease shows complex interplay between multiple genetic variants associated inflammation and tumorigenesis. We analyzed 15 polymorphisms in nine genes involved in various pathways to find out combinations of genetic variants contributing to GBC risk. The genes included in the study were (matrix etallopeptidase-2, MMP-7, MMP-9), tissue inhibitor of metalloproteinases (TIMP-2), cytochrome P450 (CYP)1A1, CYP1B1, phospholipase C epsilon 1 (PLCE1), liver X receptor (LXR)-alpha and LXR-beta. Genotypes were determined by PCR-RFLP and TaqMan probes. Statistical analysis was done by SPSS version 16. Multilocus analysis was performed by Classification and RegressionTree (CART) analysis and multifactor dimensionality reduction (MDR) to gene–gene interactions in modifying GBC risk. In silico analysis was done using various bioinformatics tools (F-SNP,FAST-SNP). Single locus analysis showed association of MMP-2(−735 C>T, −1306 C>T), MP-7 − 181 A>G, MMP-9 (P574R,R668Q), TIMP-2 − 418 G>C, CYP1A1-MspI, CYP1A1-Ile462Val,PLCE1 (rs2274223 A>G, rs7922612 T>C) and LXR-beta T>C(rs3546355 G>A, rs2695121 T>C) polymorphisms with GBC risk (p<0.05) whereas CYP1B1 and LXR-α variants were not associated with GBC risk. Multidimensional reduction analysis revealed LXR-β(rs3546355 G>A, rs2695121 T>C), MMP-2 (−1306 C>T), MMP-9 (R668Q), and PLCE1 rs2274223 A>G to be key players in GBC causation (p<0.001, CVC=7/10). The results were further supported by independent CART analysis (p<0.001). In-silico analysis of associated variants suggested change in splicing or transcriptional regulation. Interactome and STRING analysis showed network of associated genes. The study found PLCE1 and LXR-β network interactions as important contributory factors for genetic predisposition in gallbladder cancer.

140 | Fine Mapping of Lung Function Association in the MHC Region by Haplotype Imputation Reveals an Amino Acid change Underlying SNP Associations

Nick Shrine[1], Louise V. Wain[1], Andrew P. Morris[2], David P. Strachan[3]; UK BiLEVE consortium, Ian P. Hall[4], Martin D. Tobin[1]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; [3]Population Health Research Institute, St George's, University of London, London, United Kingdom; [4]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom

We performed a genome-wide association study (GWAS) of lung function quantitative traits forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC) and their ratio (FEV_1/FVC) in 48,493 samples of European Ancestry (UK BiLEVE study) which, in addition to confirming 4 previously identified lung function signals in the major histocompatibility complex (MHC) on chromosome 6, identified 2 new secondary independent signals.

In order to fine map the MHC GWAS signals, we used a published reference panel to impute HLA classical alleles and amino acid changes and tested their association with lung function traits. The new second most significant signal across the MHC region for both FEV_1 (P=5.7×10^{-13}) and FEV_1/FVC (P=1.2×10^{-20}) was an association with *HLA-DQB1* amino acid 57 Alanine present/absent (Alanine frequency=36.6%). After conditioning on the amino acid variant, 1 of 2 genome-wide significant (P<5×10^{-8}) GWAS signals for FEV_1 and 5 of 6 for FEV_1/FVC were strongly attenuated (minimum P=2.1×10^{-5}). Stepwise conditional analyses showed that the majority of the MHC lung function association signal could be explained by just the lead independent GWAS SNP and the amino acid change for both traits.

Using haplotype imputation allowed us to build upon lung function GWAS discovery to pinpoint a potential causal variant in the MHC (*HLA-DQB1* amino acid change 57, previously linked to type 1 diabetes risk) that explains a substantial proportion of the variance previously attributed to GWAS SNPs in this region.

141 | Admixture Mapping using Linear Mixed Models

Daniel Shriner[1], Charles N. Rotimi[1]

[1]Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland, United States of America

Linear mixed models have become popular in association testing because they account for population structure, familial relatedness, and cryptic relatedness and also can account for the polygenic fraction of heritable variance, thereby simultaneously controlling the false positive error rate and potentially improving power. Here, by substituting the genetic relatedness matrix estimated from genotypes with the ancestral similarity matrix estimated from local ancestry, we extend linear mixed models for admixture mapping. First, we show that the expected values of local ancestry have the same expected values of genetic relatedness in terms of kinship coefficients, but with substantially larger variance. The former indicates that power can be gained due to increasing sample sizes by including all study participants, whether nominally unrelated or in pedigrees. The latter constrains how much phenotypic variance due to genetics is specifically attributable to local ancestry. Second, we show that the "leave one chromosome out" approach is invalid. To address this issue, we introduce an approach that uses multivariate adaptive regression splines to account for correlation of local ancestry between loci. Third, if the amount of variance explained by local ancestry is small, as is typical of anthropometric and cardio-metabolic traits in admixed African Americans, then little power can be gained by accounting for the variance explained by all loci other than the one being tested.

142 | Statistical and Analytical Challenges in Microbiome Analysis of Saliva Samples using Nanopore Sequencing Compared with 16S Sequencing

Claire L. Simpson[1]

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Sciences Center, Memphis, Tennessee, United States of America

The microbiome has been established as an important factor in the development and prognosis of many diseases that interacts with genomic and environmental risk factors. A range of microbiome projects now exist with the purpose of identifying and characterizing the array of microbial organisms in both healthy and disease populations. Assaying the spectrum of microbes that are present at a body site is typically performed by genotyping fixed content arrays or more recently sequencing the bacterial specific ribosomal RNA 16S rRNA. However, both fixed content arrays and 16S sequencing may miss many low abundance or previously uncharacterized species. Alternative approaches, which use a more agnostic approach to sequencing all the available microbial genetic material may provide a broader picture of species presence in particular tissues and may be useful in identifying associations with specific disease states.

The Oxford Nanopore Technologies MinION sequencing system passes nucleic acids through a nano-scale pore and passes an ionic current through that pore, measuring the changes in

current as the molecule passes through the hole. The property of nucleotides in nucleic acids to disrupt this current in characteristic fashion allows the differentiation of the nucleotides as the molecule passes through the nanopore.

Here I present pilot data from the use of the Oxford Nanopore Technologies MinION sequencing system, and compare the results with data produced by standard 16S sequencing. I also compare relative processing overheads and processing times, including the data analytic challenges presented by each method.

143 | Genome-wide Gene by Stress Interaction Analysis Reveals Sex Difference for Kidney Function among Hispanic Individuals

Abanish Singh[1,2,3], Elizabeth Hauser[2,4,5], Nora Franceschini[6], Mildred A. Pointer[7,8]

[1] Behavioral Medicine Research Center, Duke University Medical Center, Durham, North Carolina, United States of America; [2] Duke Molecular Physiology Institute, Durham, North Carolina, United States of America; [3] Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America; [4] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, United States of America; [5] Durham Epidemiologic Research and Information Center, Durham Veterans Affairs Medical Center, Durham, North Carolina, United States of America; [6] Department of Epidemiology, The University of North Carolina at Chapel Hill, North Carolina, United States of America; [7] Cardio-metabolic Research, Julius L. Chambers Biomedical/Biotechnology Research Institute, North Carolina Central University, Durham, North Carolina, United States of America; [8] Department of Biological & Biomedical Sciences, North Carolina Central University, Durham, North Carolina, United States of America

Chronic kidney disease (CKD) is a progressive disease with high morbidity and mortality. Risk factors for CKD include age, gender, race, and those for cardiovascular disease (CVD). Psychosocial stress is one of the CVD risk factors that may have differential impact on CKD depending on gender and race. We used a GWAS framework to investigate the impact of the interaction of SNPs and chronic psychosocial stress (G×E) stratified by race on estimated glomerular filtration rate (eGFR). We used a linear regression model with an additive SNP term, a term for chronic stress, and a SNP-by-stress interaction term with population ancestry correction as well as age and sex adjustment in all ethnic groups in the MESA GWAS dataset - Whites, Chinese Americans, Blacks, and Hispanics. We identified a genome-wide significant G×E interaction in the Hispanic group. The two genome-wide significant SNPs (SNP_A-4245032, P=5.83E-09; SNP_A-8495005, P=2.18E−08) were in perfect LD (R^2= 1.0) and were located downstream of the *C20orf197* gene. We evaluated the direction of association and identified that the mean of eGFR increased with the increasing stress level for the minor allele group (TC/TT) but not for those homozygous of major allele (CC) of the most significant SNP (SNP_A-4245032). We evaluated gender differences and saw that mean eGFR increased with increasing stress levels for the minor allele group primarily in males. This locus was not significant on interethnic meta-analysis suggesting the potential for differential effects of stress and genes. This finding may help to elucidate gender differences in CKD prevalence rates.

144 | Follow-up of G×E Interactions: *EBF1* G×E Association, Synthetic Chronic Psychosocial Stress, and Dropout from a Structured Exercise Program

Abanish Singh[1,2,3], Elizabeth R. Hauser[2,4,5], Johanna L. Johnson[2], Michael A. Babyak[1,3], Beverly H. Brummett[1,3], Rong Jiang[1,3], Cris A. Slentz[2], Kim M. Huffman[2,6], Ilene C. Siegler[1,3], Redford B. Williams[1,3], William E. Kraus[2,6,7]

[1] Behavioral Medicine Research Center, Duke University Medical Center, Durham, North Carolina, United States of America; [2] Duke Molecular Physiology Institute, Durham, North Carolina, United States of America; [3] Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America; [4] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, United States of America; [5] Durham Epidemiologic Research and Information Center, Durham Veterans Affairs Medical Center, Durham, North Carolina, United States of America; [6] Department of Medicine, Duke University Medical Center, Durham, North Carolina, United States of America; [7] Duke Center for Living, Duke University Medical Center, Durham, North Carolina, United States of America

Regular physical exercise has long been recognized as a path to cardiovascular (CV) health and sustaining metabolic balance. Whereas regular physical exercise improves lipid and carbohydrate metabolism, chronic psychosocial stress has a negative effect on cardiometabolic CV risk factors. In this study, using data from two independent exercise clinical trials – Studies of a Targeted Risk Reduction Intervention through Defined Exercise (STRRIDE-RT, STRRIDE-PD), designed to study effects of exercise on cardiovascular health – we explored the relationship between genetic variants, chronic psychosocial stress, and exercise. We constructed a synthetic score of chronic psychosocial stress using the algorithm as described in Singh et al. 2015 (Gen. Epi.) and observed a statistically significant correlation between stress and depression scores (Rho=0.51, $P\leq0.0001$). We replicated a previous finding [Singh et al. 2015, EJHG] of gene-by-stress association of baseline hip circumference with *EBF1* SNP rs4704963 (P=0.005). This interaction was not significant at the post-exercise timepoint; however, post-exercise mean stress (0.15) and depression (0.56) were lower compared to pre-exercise (0.24, 0.65). We performed an ethnicity-stratified logistic regression on a dichotomous variable (0/1) for “dropout” from the structured exercise program, with chronic psychosocial stress score at the pre-exercise timepoint (age and gender adjusted) to evaluate the association between stress and dropout. This association was statistically significant (P-value=0.014) for STRRIDE White participants that was driven primarily by female dropouts (Female P-value=0.024, Male P-value=0.352) with high levels of stress.

This analysis demonstrates methods for validating G×E associations and that psychosocial stress may predict gender and race-specific dropout from an exercise program.

145 | Missing Data in Canonical Correlation Analysis of Multiple Phenotypes and Multiple SNPs

Emily Slade[1], Peter Kraft[1]

[1] *Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America*

Tests of association between multiple phenotypes and a set of SNPs can improve the power to detect a genotype-phenotype association. Canonical Correlation Analysis (CCA) provides a measure of association between two multivariate sets of variables measured on the same individuals, such as SNPs and multiple phenotypes. In settings where many variables are measured on each individual, the proportion of subjects with missing data on at least one variable can be high. When performing CCA, missing data has traditionally been handled through complete case analysis, simple mean imputation, or k-nearest neighbors approaches. We show through simulation that these methods can lead to bias of the first canonical correlation as well as loss of power and improper size in a global test of association between the two multivariate sets of variables. When there is no association between the two sets of variables, estimates of the first canonical correlation can be nonzero in finite datasets, even in the absence of missing data. The decrease of sample size in a complete case analysis exacerbates this problem. In some situations, complete case analysis overestimates the first canonical correlation over 2.5× relative to the fully observed data, even when the data were missing completely at random. Mean imputation and k-nearest neighbors imputation methods perform less poorly than complete case analysis, but researchers should consider employing even more sophisticated methods such as multiple imputation to handle missing data in canonical correlation analysis.

146 | Quantifying risk of bias in systematic reviews of genetic association studies

Zahra N. Sohani[1,2,3], Shohinee Sarma[4], Russell J. de Souza[1,2], David Meyre (1,2,), Sonia S. Anand[1,2,5]

[1] *Population Genomics Program, Department of Clinical Epidemiology & Biostatistics, Hamilton, Canada;* [2] *Chanchalani Research Centre, McMaster University, Hamilton, Canada;* [3] *Faculty of Medicine, University of Toronto, Toronto, Canada;* [4] *Michael DeGroote School of Medicine, McMaster University, Hamilton, Canada;* [5] *Department of Medicine, McMaster University, Hamilton, Canada*

Systematic reviews are a common method of providing synthesized effect of a genetic variant on a trait of interest. But, summary estimates are subject to bias due to the varying methodological quality of individual studies. We developed, validated, and are currently empirically evaluating a tool that assesses the risk of bias in systematic reviews of genetic association studies. Published guidelines and recommendations in consultation with experts were used to create items included in the Q-Genie tool. Evaluation was performed in two parts. First, four reviewers rated 30 studies randomly selected from a published meta-analysis. These ratings were used to assess construct validity, reliability, and item discrimination. We report G-coefficients as measures of inter-rater reliability, internal consistency, and overall reliability of the tool, as well as item-total correlations and Cronbach's alpha to assess the discriminative ability of each item. Second, we apply the tool to 50 meta-analyses to evaluate its ability to (i) increase precision after exclusion of low quality studies, (ii) decrease heterogeneity after exclusion of low quality studies and (iii) agreement with experts on quality rating by Q-Genie. Preliminary analyses demonstrate excellent psychometric properties and generate a score for each study with corresponding ratings of 'low', 'moderate', or 'high' quality. When applied to individual meta-analyses to exclude studies of low quality, we found a decrease in heterogeneity and an increase in precision of summary estimates. Integration of Q-Genie into meta-analyses can inform selection of studies for inclusion, conduct sensitivity analyses, and perform meta-regressions.

147 | The Power and Type I Error of Tiled Regression Analysis Depend on the Selection Criteria at all Stages

Alexa J.M. Sorant[1], Jeremy A. Sabourin[1], Heejong Sung[1], Alexander F. Wilson[1]

[1] *Genometrics Section, Computational and Statistical Genomics Branch, NHGRI, NIH, Baltimore, Maryland, United States of America*

Tiled regression, a staged procedure that performs variable selection on subsets of genetic variants in hotspot-based tile, chromosome, and genome-wide regions, is controlled by selection parameters at each stage. The effects of different sets of critical values for stepwise regression for the three stages were studied using simulations based on the genotypes of unrelated individuals in the Trinity Student Study. A quantitative trait was modeled on the additive effects of five independent SNPs, each with locus-specific heritability 0.01. Two hundred replicates were analyzed with tiled regression, as implemented in TRAP v2.0, for two critical value sets: (1) 10^{-5} for all stages and (2) 10^{-3}, 10^{-4} and $10^{-5}$for the tile, chromosome and genome-wide stages, respectively. The observed type I error rate (considering SNPs on chromosomes without causal SNPs) for set 1 was 3.25×10^{-6}, while that for set 2 was 5.72×10^{-6}, closer to the final critical value (10^{-5}). Power estimates with sets 1 and 2 were 0.386 and 0.395, respectively. The "power" to detect tiles containing causal SNPs was higher: 0.589 and 0.598 for sets 1 and 2, respectively. Analogously, rates of selecting tiles on chromosomes without causal SNPs relative to the total number on those chromosomes were also higher: $4.00\times10^{-5}$and $7.02\times10^{-5}$for sets 1 and 2, respectively. In these simulation experiments, using a liberal critical value at the first stage of tiled regression and more conservative values at later stages increased type I error rate and power, compared to using the final critical value at all stages.

148 | Genome-Wide Analysis of Copy Number Variation and Common Facial Variation in a Large Cohort of Bantu Africans

Megan Sorenson[1], David Astling[2,3], Hung-Chun Yu[2], Feyza Yilmaz[2,4], Joanne Cole[2,5], Stephanie A. Santorico[1,5,6], Richard A. Spritz[2,5], Tamim H. Shaikh[2,5], Audrey E. Hendricks[1,5,6]

[1]Mathematical and Statistical Sciences, University of Colorado - Denver, Denver, Colorado, United States of America; [2]School of Medicine, University of Colorado - Anschutz Medical Campus, Aurora, Colorado, United States of America; [3]Department of Biochemistry and Molecular Genetics, University of Colorado–Anschutz Medical Campus, Aurora, Colorado, United States of America; [4]Department of Integrative Biology, University of Colorado–Denver, Denver, Colorado, United States of America; [5]Human Medical Genetics and Genomics Program, University of Colorado–Anschutz Medical Campus, Aurora, Colorado, United States of America; [6]Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, United States of America

The face is one of the most distinguishing characteristics of the human body, and similarity between relatives points towards a strong genetic component in normal facial development. However, little is known about the genetic factors underlying normal facial appearance, particularly Copy Number Variations (CNVs). CNVs are a substantial source of normal genetic variation and CNVs have been directly implicated in several genetic disorders that exhibit dysmorphic facial features. We hypothesize that some variability in normal facial appearance results from genetic variation mediated by CNVs. We test this using genotype and phenotype data from ∼3700 Bantu African children. These individuals were genotyped using high-density SNP-based microarrays and photographed using 3D cameras, which enable quantification of facial distances and shapes that explain the majority of facial shape variation. CNVs were called by PennCNV, DNACopy, and VanillaICE, and a high-quality set called by multiple algorithms were used for analysis. We have detected CNVs in several genes previously implicated in craniofacial development and dysmorphic disorders, including *SHH*, *PAX3*, *FGFR1*, *FGFR2*, and *IRF6*. We present results from single variant (CNV frequency $> 1\%$) and gene-region based (CNV frequency $\geq 5\%$) analyses using EMMAX to adjust for population structure and hidden relatedness, and including known covariates (e.g., age, sex). In addition to providing a comprehensive analysis of the relationship between CNVs and facial morphology in this population, our study provides a unique source of information about CNVs in a large African sample.

149 | PhenoScanner: A Database of Human Genotype-Phenotype Associations

James R. Staley[1], James Blackshaw[1], Mihir A. Kamat[1], Steve Ellis[1], Robin Young[1,2], Adam S. Butterworth[1,3]

[1]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Robertson Centre for Biostatistics, University of Glasgow, Glasgow, United Kingdom; [3]NIHR Blood and Transplant Research Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

PhenoScanner is a curated database of publicly available results from large-scale genetic association studies. This tool aims to facilitate “phenome scans”, the cross-referencing of genetic variants with many phenotypes, to help aid understanding of disease pathways and biology. The database currently contains over 350 million association results and over 10 million unique genetic variants, mostly single nucleotide polymorphisms. It is accompanied by a web-based tool that queries the database for associations with user-specified variants, providing results according to the same effect and non-effect alleles for each input variant. The tool provides the option of searching for phenotype associations with proxies of the input variants, calculated using the European samples from 1000 Genomes phase 3 and HapMap 2.

In the near future, we plan to release a second version of PhenoScanner with additional databases containing results on gene expression (>140 million results) and metabolomics (>2 billion results), as well as increasing the number of variant associations with diseases and traits to more than 600 million. This version will also contain variant and gene annotations along with phenotype ontology mappings. Furthermore, the functionality of PhenoScanner will be increased to allow greater number of variant queries and effect allele specification as well as include gene-based and regional-based query options. In tandem, we intend to release an R package that will allow users to query the PhenoScanner databases from within R.

PhenoScanner is available at: www.phenoscanner.medschl.cam.ac.uk

150 | A statistically Efficient Gene-Mapping Method that Reduces Sequencing Costs and Better Prioritizes Candidate Genes

William C.L. Stewart[1,2,3], Christopher W. Bartlett[1,3]

[1]The Research Institute at Nationwide Children’s Hospital, Columbus Ohio, United States of America; [2]Department of Statistics, The Ohio State University, Columbus Ohio, United States of America; [3]Department of Pediatrics, The Ohio State University, Columbus Ohio, United States of America

Although causal genes for some disorders were first mapped through the analysis of large multiplex pedigrees (e.g., HD, BRCA1, BRCA2, PS1, and GJB2), this approach usually results in candidate gene regions that are still too large. Typically, dozens (if not hundreds) of genes are implicated, and the sequencing costs are often prohibitive. To address these concerns, we developed an efficient interval estimator that reduces the size of candidate gene regions, and a powerful test of association (POPFAM+) that better prioritizes the genes within them. Our interval estimator is the 95% confidence interval for trait location, and relative to competing estimators, it has increased precision because it uses all of the available dense SNP data. Moreover, it is specifically designed to handle large multiplex pedigrees. After constructing the confidence interval, we prioritize the genes within it on the basis of each gene's evidence for association.

We applied our new interval estimator to the dense SNP data of four large families segregating a specific language impairment gene on chromosome 13. Furthermore, we applied POPFAM+ to these same data, and to the dense SNP data of 417 matched controls. We found that our interval estimator reduced the candidate gene region by more than 20%, and that POPFAM+ increased our ability to detect associated variants underneath the linkage peak. Overall, these two complementary methods should greatly reduce the costs of targeted resequencing efforts, and dramatically increase the rate at which disease genes are found.

151 | Using Polygenic Risk Scores to Predict Response to Methotrexate in Rheumatoid Arthritis

Jenna L. Strathdee[1], John C. Taylor[1], Tim Bongartz[2], James I. Robinson[1], Ann W. Morgan[1], Jennifer H. Barrett[1]; MATURA Consortium

[1] School of Medicine, University of Leeds, Leeds, United Kingdom; [2] Division of Rheumatology, Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, United States of America

Rheumatoid arthritis (RA) is an autoimmune disease that can lead to permanent joint damage if not treated effectively at an early stage. The usual first-line treatment is methotrexate (MTX), but some patients fail to respond to this drug. Factors influencing response are currently unknown but could include a genetic component, although single variant analysis has so far not returned any genome-wide significant results. An alternative approach is to analyze combinations of variants by constructing polygenic risk scores (PRS) from potentially relevant traits. A first PRS was derived from a recent RA susceptibility meta-analysis, summing over genome-wide significant variants from ~100 loci, weighting the number of risk alleles by the coefficient from the RA susceptibility study. The PRS was used to predict outcome in ~1000 RA patients, assembled from different sources by the MATURA consortium, all treated with MTX monotherapy and followed up after 6 months. Four outcomes were analyzed: 6-month change in disease activity score (DAS28CRP3) and in its three components (C-reactive protein (CRP), swollen joint count, and tender joint count), using linear regression adjusted for age, sex and baseline measure and meta-analysed across datasets. No significant associations were found between this PRS and MTX response. A similar approach is now underway to analyze PRSs derived from other phenotypes, including other autoimmune diseases, serum biomarkers related to the immune system and inflammation and gene expression levels.

152 | Replication of a Single Nucleotide Polymorphism Variant in *CETP* Gene Associated with Large-HDL Particle in the *ClinSeq*® Study

Heejong Sung[1], Maureen Sampson[2], Katie Lewis[3], David Ng[3], Stephen G. Gonsalves[3], NISC Comparative Sequencing Program[4], James C. Mullikin[4,5], Alan Remaley[6], Leslie G. Biesecker[3], Alexander F. Wilson[1]

[1] Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, NIH, Baltimore Maryland, United States of America; [2] Clinical Center, NIH, Bethesda, Maryland, United States of America; [3] Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland, United States of America; [4] National Institutes of Health Intramural Sequencing Center (NISC), National Human Genome Research Branch, NIH, Bethesda, Maryland, United States of America; [5] Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland, United States of America; [6] Lipoprotein Metabolism Section, National Heart, Lung, and Blood Institute, NIH, Bethesda, Maryland, United States of America

ClinSeq® is a large-scale medical sequencing study to investigate associations of sequence variants with traits related to Coronary Artery Disease (CAD). The study currently includes more than 1000 non-smoking participants between the ages of 45 and 65 with coronary artery calcification scores. About 200 CAD-related traits were measured at the NIH Clinical Research Center. Whole-exome sequencing was performed with the Agilent SureSelect 38Mb and 50Mb capture kits for 387 and 325 individuals, respectively, at the NIH Intramural Sequencing Center. Cryptic relatedness and misspecified population stratification were checked by multidimensional scaling analysis, with 635 unrelated European Americans (EAs) remaining. SNVs common to both capture regions, with EAs in each region having some variation and at least 50% call rate, were merged, yielding 439,807 SNVs. Of these, the SNVs with MAF<0.01 were collapsed into a single derived variant for each genomic region defined by hotspot blocks. Lipoprotein particle profiles including High Density Lipoprotein particle (HDLp) were measured by nuclear magnetic resonance spectroscopy at LipoScience/LabCorp Global Research Services. Tests of association between each SNV and each classified HDLp were performed on untransformed, log-transformed and rank-inverse transformed HDLp with simple linear regression, adjusting for age, sex, BMI and use of medication. The SNVs rs1532625 and rs7205804 in the intron of ***CETP*** gene were associated (p<1e−05) with untransformed and transformed large-particle HDLp (combining HDLp for estimated particle diameters of 9.7, 10.5 and 12 nm). This finding replicates the association by Reilly et al. (2013) between rs1532625 and HDL level.

153 | A Genome-Wide Association Study of Multiple Longitudinal Traits with Related Subjects

Yubin Sung[1], Zeny Feng[1], Sanjeena Subedi[1]

[1] Department of Mathematics and Statistics, University of Guelph, Guelph, Canada

Pleiotropy is a phenomenon in which a single gene inflicts multiple correlated phenotypic effects, often characterized as traits, involving multiple biological systems. We propose a two-stage method to identify pleiotropic effects on multiple longitudinal traits from a family-based data set. The first stage analyzes each longitudinal trait via a three-level generalized mixed-effects model. Random effects predicted at the subject-level and at the family-level measure the subject-specific genetic effects and between-subjects intraclass correlations within families, respectively. The second stage performs a simultaneous association test between a single nucleotide polymorphism and all subject-specific random effects corresponding to the multiple longitudinal traits analyzed in the first stage. The simultaneous genetic association test is conducted based a generalized quasi-likelihood scoring method (GQLSM) in which the correlation structure among related subjects is adjusted. We conduct two simulation studies to assess the performance of our proposed method and demonstrate its applicability by analyzing the Genetic Analysis Workshop 16 Problem 2 cohort data drawn from the Framingham Heart Study.

155 | Comparison of Variant Calling Software for Pooled Sequencing Studies

Silke Szymczak[1], Stefanie Müller[2], Franziska Hopfner[1,2], Gregor Kuhlenbäumer[2], Michael Krawczak[1], Astrid Dempfle[1]

[1] Institute of Medical Informatics and Statistics, University of Kiel, Kiel, Germany; [2] Department of Neurology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany

Pooling multiple individuals in next generation sequencing studies is a cost efficient approach for rare variant detection. To enable identification of variant carriers, special designs can be used so that each individual is sequenced in several pools. This strategy requires that pool based genotypes are estimated accurately.

We compared several variant calling algorithms that have been developed for pooled sequencing studies. A targeted sequencing study was simulated for different patterns of Single Nucleotide Variant (SNV) carriers, coverages per pool and frequencies of the rare variants.

Across all scenarios, CRISP had the largest empirical power to detect true SNVs and to estimate genotypes, followed by VarScan, LoFreq and vipR. For most scenarios, power of CRISP was nearly 100% for both detection and estimation if the SNV was covered by at least 50 reads per sample in each pool. FreeBayes performed differently for varying scenarios and had the largest false positive rates. However, false positive rates were usually low and only increased to about 10% for settings with a low coverage of 12.5× per sample in each pool.

In conclusion, we recommend CRISP for calling rare variants in pooled sequencing studies. In low coverage regions, calling might be additionally performed with VarScan to reduce false positive findings.

156 | Mediation Analysis of Bidirectional Associations with Application to Obesity and Diabetes

Rajesh Talluri[1], Sanjay Shete[1,2]

[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America; [2] Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

Obesity and diabetes are both major public health issues and are risk factors for each other. Recent genome wide studies have identified several genetic variants associated with either obesity, diabetes, or both. Because of the known interdependence between obesity and diabetes we hypothesize that some of these SNPs that are associated with both obesity and diabetes may be mediated through obesity to affect diabetes or through diabetes to affect obesity. Identifying these mediated relationships would further enhance our knowledge about diabetes and obesity. We propose a framework for performing bidirectional mediation analyses. Through, simulations, we showed the statistical properties of our methods. In many scenarios, the residuals of mediator and outcome are correlated because of unmeasured predictors not included in the mediation model. We showed that the proposed model gives unbiased estimates in this scenario too. We applied the proposed model to investigate the bidirectional relationship between the diabetes and obesity using the genome wide association study data from the MESA cohort. We identified 6 SNPs that were associated with both diabetes and obesity. Two SNPs (rs3752355 and rs6087982) had indirect effects on obesity mediated through diabetes (0.28; 95% CI [0.01, 0.67] and 0.36; 95% CI [0.08, 0.85], respectively). The remaining four SNPs (rs7969190, rs4869710, rs10201400 and rs12421620) directly affect diabetes and obesity without any mediation effects.

157 | Genome-Wide Meta-Analysis of Response to Methotrexate in Rheumatoid Arthritis Patients on behalf of the MATURA Consortium

John C. Taylor[1], Suzanne Verstappen[2], Jianmei Wang[3], Tim Bongartz[4], Ian Scott[5], Paul Emery[1], Costantino Pitzalis[6], Anne Barton[2], Ann W. Morgan[1], Jennifer H. Barrett[1]

[1] School of Medicine, University of Leeds, Leeds, United Kingdom; [2] Centre for Musculoskeletal Research, The University of Manchester, Manchester, United Kingdom; [3] Roche Products, Welwyn Garden City, United Kingdom; [4] Vanderbilt University, Nashville, Tennessee, United States of America; [5] King's College, London, United Kingdom; [6] Queen Mary University, London, United Kingdom

The recommended first treatment for rheumatoid arthritis (RA) is methotrexate (MTX), but only about half of patients respond adequately. Patients not responding are offered alternative drugs, but irreversible damage may already have occurred. Better prediction of treatment response would increase the chance of achieving early disease control. Previous published genetic studies of response to MTX in RA have been small and are mainly candidate-gene studies. We carried out a genome-wide study of response to MTX in approximately 1000 RA patients of European ancestry.

Patients were from observational cohorts of early RA or the placebo arm of clinical trials. Genotypes were imputed to 1000 Genomes. Response was measured as change over 3–6 months in DAS28CRP3, a measure of disease activity based on C-reactive protein levels, swollen joint count and tender joint count. Four genome-wide linear regression analyses were carried out - of change in DAS28CRP3 and each of its components adjusted for baseline. The analysis was conducted on five separate cohorts and results meta-analyzed.

Although no (SNP) reached genome-wide significance, one region on chromosome 11 reached $<10^{-7}$for DAS28, and other regions reached 10^{-6}; these are being followed up in independent cohorts. No support was seen for associations previously reported in RA. There was weak support for association with *ZMIZ1*, previously reported as showing some evidence of association in a genome-wide study of response to MTX in juvenile idiopathic arthritis (P=$6\times10^{-6}$from meta-analysis across diseases). Larger studies will be required to detect genetic factors that robustly predict treatment response.

158 | Personalized Prevention? Causal Inference Methods for Evaluating Genetic Targeting Strategies for Screening

Duncan C. Thomas[1]

[1] University of Southern California, Los Angeles, California, United States of America

We previously showed that risk stratification for colorectal cancer based on 27 replicated GWAS loci yielded a 10-year range for optimal ages at start of endoscopy, but we did not investigate the timing of subsequent screens. Because screening behavior depends on previous screening history and family members' behaviors, these can act as both confounders and intermediate variables on a causal pathway, leading to bias in conventional analyses of the causal effect of screening. Inverse propensity score weighting provides a way to analyze observational data as if it were a sequence of trials where screening is applied at random rather than self-selected. I simulated family data under plausible models for the underlying disease process and for screening behavior to assess whether a targeted screening approach based on individual's risk factors would lead to a greater reduction in cancer incidence than a uniform screening policy. These showed that conventional analyses produced a substantial positive bias, which was eliminated by using propensity score weighting. A large case-control study of colonoscopy and colorectal cancer from Germany showed a large protective effect, which inverse propensity score weighting made even stronger. Targeted screening approaches based on either fixed risk factors or family history are predicted to yield greater reductions in cancer incidence, with fewer screens needed to prevent one cancer than population-wide approaches, but the differences may not be large enough to justify the additional effort required. GWAS analyses currently underway may allow clearer delineation of optimal screening schedules in the future.

159 | A Prospective Likelihood Copula-Based Approach for the Analysis of Secondary Phenotypes in Ascertained Samples

Fodé Tounkara[1], Geneviève Lefebvre[1], Celia Greenwood[2], Karim Oualkacha[1]

[1] Université du Québec à Montréal (UQAM), Montréal, Canada; [2] Lady Davis Institute for Medical Research, Jewish General Hospital and McGill University, Departments of Oncology, Epidemiology, Biostatistics & Occupational Health, and Human Genetics, Montréal, Canada

Data collected for a GWAS of a primary phenotype are often used for additional genome-wide association analyses of secondary phenotypes. However, when the primary and secondary traits are dependent, analyses of secondary phenotypes may induce spurious associations in non-randomly ascertained samples, and naïve analyses are inappropriate to adjust for them. Previously, a retrospective likelihood copula-based method has been proposed in this context; however this method has only been applied to studies with case-control sampling. We propose a prospective likelihood copula-based approach to appropriately detect genetic variant/secondary phenotype association in the presence of selected samples. We use the prospective likelihood to incorporate the sampling mechanisms, and use copulas to allow for several dependence structures between the primary and the secondary phenotypes. The simulations show that our methods take correctly into account the sampling ascertainment, control type 1 error and are robust to the misspecification of the prevalence. We are currently analyzing data from the Twins United Kingdom study. We have ascertained individuals from this cohort based on their weight, and we are examining genetic associations with the related phenotype of bone mineral density, measured at both lumbar spine and femoral neck. The results of the real data analyses are still to come. Numerical results show that the proposed methods produce similar results as those obtained under the retrospective likelihood copula-based method. We plan to extend and compare both retrospective and prospective copula-based approaches to more general ascertainment mechanisms such as sampling from extreme phenotypes.

160 | Impact of Genetic Variants on Latent Class Modeling of Parenting Behavior in Neurodevelopmental Studies

Eva Unternaehrer[1], Keelin Greenlaw[2], Katherine T. Cost[3], Andrée-Anne Bouvette-Turcot[1], Hélène Gaudreau[1], Kieran J. O'Donnell[1], Marla B. Sokolowski[4], James L. Kennedy[5], Meir Steiner[6], Leslie Atkinson[7], John E. Lydon[8], Alison S. Fleming[3], Michael J. Meaney[1], Antonio Ciampi[2], Celia Greenwood[2]; On Behalf of the MAVAN Research Team

[1]Douglas Mental Health University Institute, McGill University, Montreal, Canada; [2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [3]University of Toronto Mississauga, Toronto, Canada; [4]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Canada; [5]Center for Addiction and Mental Health, Toronto, Canada; [6]St. Joseph's Medical Center, Hamilton, Canada; [7]Department of Psychology, Ryerson University, Toronto, Canada; [8]Department of Psychology, McGill University, Montreal, Canada

Parental behavior shapes offspring psychosocial and neuronal development. Thus, understanding and predicting parenting is essential for identifying children at risk for mental disorders. A parent needs to adapt childcare to the developmental stage of the child. The evolution of this dynamic process is usually assessed by longitudinal measures of appropriate scales, which can be seen as multivariate trajectories. These trajectories are influenced by various psychosocial and biological factors, including genetic variants implicated in mental health and disease. Latent Class Modeling is a useful tool to model such trajectories, and to identify distinct patterns of parenting behavior (clusters). The aim of this study was to examine the impact of genetic variants in candidate genes on parenting trajectories. We assessed maternal self-reports of parenting and mental well-being annually, across the first 6 years postpartum in a sample of 375 mothers participating in the Maternal Adversity, Vulnerability and Neurodevelopment Study. We used Latent Gold® Software for latent class modeling with the aim of clustering mothers based on their changes in parenting variables. Preliminary analyses identified six distinct clusters of mothers, demonstrating that this method can extract useful information from psychosocial data. We will present results on the influence of different covariates on these trajectory clusters, including SNPs in selected candidate genes, maternal psychosocial stress exposure, and socioeconomic status. Besides novel insight into predictors of maternal parenting trajectories, the suggested approach might provide stronger maternal phenotypes in the context of neurodevelopmental studies, as compared to cross-sectional or single parenting measures.

161 | A Gene-by-Environment Analysis of 184,428 Subjects Reveals a Role for Adaptive and Innate Immunity in Coronary Artery Disease in Subjects with Type 2 Diabetes

Natalie R. van Zuydam[1,2], Benjamin F. Voight[3], Claes Ladenvall[4], Juan Fernandez[1], Rona J. Strawbridge[5]; On Behalf of CARDIoGRAM and SUMMIT

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, London, United Kingdom; [2]Medical research institute, University of Dundee, Dundee, United Kingdom; [3]University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [4]University of Uppsala, Uppsala, Sweden; [5]The Karolinksa Institute, Stockholm, Sweden

Subjects with type 2 diabetes (T2D) have accelerated atherosclerosis and a larger burden of coronary artery disease (CAD) compared to subjects without diabetes. We aimed to identify variants and pathways that modified the risk of CAD in the context of T2D.

We combined summary statistics from 1000G/HapMap2 imputed genotypes and CardioMetabochip genotypes from European ancestry subjects that were either CAD cases (N=27,706) or controls (N=38935) with no history of CAD. Of these 24,257 had T2D (10,012 cases) and 42,384 had no T2D (ND; 17,694 cases). Statistics were combined in a fixed effects inverse-variance weighted meta-analysis and tested for interaction with T2D status. Replication was sought for independent loci (P-value<1×10^{-4}) from the diabetes stratified (nloci=511) and interaction analysis (nloci=175), in further subjects: 11,537 with T2D (3,706 cases) and 106,250 without T2D (12,988 cases).

Overall, known CAD loci were associated with CAD irrespective of T2D status. Two loci showed replicable evidence of interaction with T2D status: rs79705396, near *IL15RA* and *IL2RA* (OR T2D=0.68, OR ND=1.20, P.int=5.2×10^{-6}), and rs7952366, near *PVRL1* (OR T2D=1.10, OR ND=0.97, P.int=1.6×10^{-5}). Pathway analysis showed interaction loci were enriched for the Fc$_{\gamma}$R mediated phagocytosis (KEGG) pathway (FDR<0.05); protein-protein interaction analysis identified enrichment in the semaphorin-plexin signalling pathway (GO) specifically in subjects with T2D (p=2×10^{-17}). The interaction loci and pathway analysis approaches suggest a differential role for the immune/inflammatory system in CAD in subjects with T2D, a concept well supported by other approaches in the literature.

162 | Investigating Interactions Among Genetic and Environmental Risk Factors in Longitudinal Family Studies

Cheng Wang[1], Jean-François Lefebvre[1], Lise Dubois[1], Kelly M. Burkett[2], Marie-Hélène Roy-Gagnon[1]

[1] School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada; [2] Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada

Although many genetic variants have been associated with complex phenotypes, the marginal effect from these loci alone has failed to fully explain the trait heritability in populations. It is hypothesized that interaction effects between genes and environmental factors may account for some of the "missing heritability". Longitudinal family studies can potentially provide a more powerful approach to study Gene-Environment (GE) interactions, but it also presents additional challenges due to the temporal and familial correlations in the data structure. In this study, we evaluate the performance of various approaches to detect GE interactions in longitudinal twin data.

We performed a simulation study based on data from the ongoing Quebec Newborn Twin Study (QNTS), a twin cohort from Montreal, Canada. We used a modified linear mixed model to simulate arbitrary interaction relationships between hypothetical genetic and environmental exposures and Body Mass Index (BMI) at 6 time points. We compared power and type I error of the interaction test from the classical twin model, 3-level hierarchical linear mixed model and nonparametric Partition Based score I (PBI) test. The twin model and PBI methods were applied to the mean over time summary statistic. The classical twin model performed generally well, except for clearly non-linear interactions. PBI had low power to detect linear interactions. All methods had low power for some interaction scenarios such as those with low marginal effects. Our results will help guide our analysis of the QNTS data and other family-based longitudinal studies of genetic interactions.

163 | Enhancing Power of Rare Variant Association Test by Zoom-Focus Algorithm (ZFA) to Locate Optimal Testing Region

Maggie Haitian Wang[1], Haoyi Weng[1], Rui Sun[1], Benny Chung-Ying Zee[1]

[1] Division of Biostatistics, JC School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR; CUHK Shenzhen Research Institute, Shenzhen, China

Exome or targeted sequencing data exerts analytical challenge to test Single Nucleotide Polymorphisms (SNPs) with extremely small Minor Allele Frequency (MAF). Various rare variant tests were proposed to increase power by aggregating SNPs within a fixed genomic region, such as a gene or pathway. However, a gene could contain from several to thousands of markers, and not all of them may be related to the phenotype. Combining functional and non-functional SNPs in arbitrary genomic region could impair the testing power. We propose a Zoom-Focus Algorithm (ZFA) to locate the optimal testing region for a given fixed collapsing region, which can be applied in conjunction with a variety of rare variant association tests. In the first Zooming step, a given genomic region is partitioned by order of two, and the best partition is located within all partition levels. In the next Focusing step, boundaries of the zoomed region are refined. Simulation studies showed that ZFA substantially enhanced the statistical power of rare variant tests by over 10 folds, including the WSS, SKAT and W-test. The method is applied on real exome sequencing data of hypertensive disorder, and identified biologically relevant genetic markers to metabolic disorder that are undiscoverable by testing using full gene. The proposed method is an efficient and powerful tool to increase the effectiveness of rare variant association tests for exome sequencing datasets of complex disorder.

164 | The Influence of Genetic Susceptibility and Calcium Plus Vitamin D Supplementation on Fracture Risk

Youjin Wang[1], Jean Wactawski-Wende[1], Lara E. Sucheston-Campbell[2], Leah Preus[1], Kathleen M. Hovey[1], Jing Nie[1], Rebecca D. Jackson[2], Samuel K. Handelman[2], Rami Nassir[3], Carolyn J. Crandall[4], Heather M. Ochs-Balcom[1]

[1] University at Buffalo, The State University of New York, Buffalo, New York, United States of America; [2] The Ohio State University, Columbus, Ohio, United States of America; [3] University of California Davis, Davis, California, United States of America; [4] University of California Los Angeles, Los Angeles, California, United States of America

A recent meta-analysis of Genome-Wide Association Studies (GWASs) identified multiple Bone Mineral Density (BMD) and fracture-associated loci. We evaluated whether GWAS-identified loci modify the association between Calcium plus vitamin D (CaD) intake and fracture risk. Data from 5,823 white Women's Health Initiative CaD trial participants were included. Participants received 1,000 mg elemental calcium with 400 IU vitamin D_3 daily or placebo. The fracture GRS (Fx-GRS) was composed of 16 fracture and BMD-associated variants, and BMD-GRS was created with 50 BMD-associated variants from a meta-GWAS. We used both Cox regression and a case-only approach to test for multiplicative interaction. Additive interaction was assessed with the Relative Excess Risk due to Interaction (RERI). We categorized participants into three groups based on GRS quartile (Q1, Q2-3, and Q4). We observed no interaction between the Fx-GRS and CaD on fracture risk; however, we observed a significant multiplicative interaction between the BMD-GRS and CaD ($p_{\text{interaction}}$=0.01). There was a significant negative additive interaction between placebo assignment and higher BMD-GRS; RERI (95% CI), Q2-3: −0.61 (−1.18, −0.05), Q4: −0.76 (−1.43, −0.09). The protective effect of CaD on fracture risk was observed in women in the lowest BMD-GRS

quartile (HR=0.60, 95% CI=0.44, 0.81) but not in women with higher BMD-GRS. We observed significant effects of CaD on fracture risk only in women with lowest genetic predisposition to low BMD. Future large-scale studies with functional characterization of GWAS findings are warranted to assess the utility of GRS in analysis of risks/benefits of CaD for bone health.

165 | Discovery of Novel Loci Associated with Heart Rate from Exome Chip Analysis

Helen R. Warren[1,2], Marten E. van den Berg[3], Pim van der Harst[4], Niek Verweij[4], Mark Eijgelsheim[5], Bruno H. Stricker[6,7,8,9], Patricia B. Munroe[1,2]; On Behalf of the CHARGE Consortium EKG Working Group

[1] Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [2] NIHR Barts Cardiovascular Biomedical Research Unit, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [3] Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands; [4] Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [5] University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [6] Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; [7] Netherlands Genomics Initiative-sponsored Netherlands Consortium for Healthy Ageing, Rotterdam, The Netherlands; [8] Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; [9] Inspectorate of Health Care, The Hague, The Netherlands

Electrocardiographic (EKG) measures are useful for diagnosing cardiac conditions. Well-established EKG-based cardiovascular risk factors include a short RR interval, which is inversely proportional to a high heart rate (HR).

The largest (N∼180,000) published genome-wide association study (GWAS) for HR increased the number of loci to 21, but much of the genetic variation remains unknown.

Our discovery data from 30 studies included ∼105,000 European descent participants genotyped at ∼250,000 exome-chip variants. The RR phenotype was taken directly from EKGs or converted from HR. Association analyses were performed using R-SeqMeta. Sensitivity analyses showed no bias from participants taking beta-blockers, and no requirement for phenotype transformation.

Replication was sought from an independent GWAS for HR from UK Biobank (∼135,000 European ancestry participants), giving a total of N∼240,000 in the sample size weighted combined meta-analysis in METAL.

Single variant analyses identified nine novel loci, reaching genome-wide significance from combined meta-analysis and Bonferroni-adjusted significance from replication, with two missense lead variants predicted to be damaging. Three loci (RNF207, SCN10A, GBF1) have published associations with other EKG traits (QT, PR, QRS). Novel loci results were looked-up in non-European ancestries (N∼11,000: African Americans, Hispanics, Chinese Americans) but non-significant, likely due to lack of power.

New variants, not in LD with known SNPs, were validated within nine known HR loci, including some low-frequency and missense variants. Two known HR genes (CCDC141, KIAA1755) yielded significant gene-based associations, driven by the single independent additional low-frequency variants. Our new findings advance our knowledge of the genetic architecture of HR and other EKG traits.

166 | Locus Discovery in Genome-wide Association Studies using Bivariate Analysis

Nicole M. Warrington[1,2], David M. Evans[1,3,4]

[1] The University of Queensland Diamantina Institute, The University of Queensland, Translational Research Institute, Brisbane, Queensland, Australia; [2] School of Women's and Infants' Health, The University of Western Australia, Perth, Western Australia, Australia; [3] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [4] School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

Statistical theory and simulations have shown that bivariate genetic analysis has increased statistical power over univariate analysis to detect genetic variants that contribute pleiotropically to the two phenotypes, particularly when the effect of the variants on the two traits is in the opposite direction to the prevailing phenotypic correlation. Despite this, bivariate strategies have been under-utilized in the gene mapping literature. Observational epidemiological studies have shown that an individual's birth weight is inversely related to their blood pressure in later life. We used a subset of individuals of European descent in the UK Biobank with self-reported birth weight, blood pressure measurements and information on doctor diagnosed high blood pressure. We conducted a variance components analysis in BOLT-LMM to estimate the genetic correlation between birth weight and high blood pressure (r_g=-0.13 [SE=0.04], r_e=−0.04 [SE=0.01], N=70,645), systolic blood pressure (r_g=−0.19 [SE=0.03], r_e=−0.04 [SE=0.01], N=67,064) and diastolic blood pressure (r_g=−0.09 [SE=0.03], r_e=−0.03 [SE=0.01], N=67,075). Given these estimates indicate that a subset of genetic variants contribute pleiotropically to the phenotypes, we conducted bivariate genome-wide association analysis between birth weight and the three phenotypes. There were more than 100 genome-wide significant loci across the three phenotypes; however, only one locus showed association in the bivariate analysis and not the univariate analysis (top SNP: $P_{Univariate}=5.4\times10^{-8}$, $P_{Bivariate}=3.1\times10^{-8}$). This lack of locus discovery using bivariate analysis could be because the residual correlation between the phenotypes is too small or due to maternal genotype effects on the intrauterine environment influencing birth weight.

167 | To Adjust or not Adjust: Genomic Screens Using a Compound of Models Including Adjusted Phenotypes can Distinguish Biological Models for Obesity

Thomas W. Winkler[1], Zoltán Kutalik[2,3], Iris M. Heid[1]

[1]Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; [2]Institute of Social and Preventive Medicine, CHUV-UNIL, 1010 Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Genome-wide association studies searching for disease loci increasingly utilize multiple phenotypes and statistical models adjusting for covariates. However, such genomic screens depict not only genetic factors for the primary phenotype, but also those of the covariate to an extent that depends on the correlation. A deeper understanding of how to interpret results of adjusted model screens is currently lacking.

We exemplify results and interpretation of adjusted model screens on the example of Body Mass Index (BMI), Waist-Hip-Ratio (WHR) and WHR adjusted for BMI (WHRadjBMI) with data from the GIANT consortium (N>300,000).

We identify 159 independent genome-wide significant loci across the three screens and classify loci into four different biological models based on their relative association results: factors that affect^{1}BMI and WHR in the direction as expected by the overall correlation (82 loci); ^{2}BMI-only (25 loci); ^{3}WHR-only (28 loci), and^{4}BMI and WHR into opposite directions highlighting fat distribution dysfunction (24 loci). Pathway results support a notion of distinct biology for these four models that extends previous differentiations based on BMI- and WHRadjBMI-signals. We yield similar results when computing the WHR association results from the BMI and WHRadjBMI results and the correlation between BMI and WHR.

We illustrate the ability of the WHRadjBMI screen to not necessarily depict genetic factors affecting WHR-only (without effect on BMI), but primarily genetic factors associated with unusual fat distribution. Our results support future application of adjusted models particularly in a compound with the pure phenotypes in order to differentiate biological models.

168 | The Role of Early-Life Growth Development, FTO Gene and Exclusive Breastfeeding on Child BMI Trajectories

Yan Yan Wu[1], Steve Lye[2], Laurent Briollais[2,3]

[1]Department of Public Health Sciences, University of Hawai'i at Manoa, Honolulu, Hawai'i, United States of America; [2]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada; [3]Dalla Lana School of Public Health, Biostatistics Division, University of Toronto, Toronto, Canada

Recent studies have implicated the FTO gene in child and adult obesity. About 70% of the Caucasian population carries at least one FTO rs9939609 risk variant. A longer duration of Exclusive Breastfeeding (EXBF) has been shown to reduce BMI and the risk of being overweight in the general population and among FTO gene carriers. However, it remains unclear whether the preventive effect of EXBF could be explained by its impact on early life growth development, e.g. ages at Adiposity Peak (AP), Adiposity Rebound (AR), BMI velocities in first years of life, which are major determinants of overweight and obesity later in life. We studied 5,590 children from the British Avon Longitudinal Study of Parents and Children (ALSPAC) cohort and modeled their longitudinal BMI profiles from birth to 17 years of age as well as their age at AP, AR and BMI velocities in relation to FTO gene variant and EXBF. A longer duration of EXBF (i.e., at least 5 months) has significant impact on BMI growth trajectories among children carrying FTO adverse variant by modulating the timing of AP, AR and the BMI velocities. EXBF could help "rewind" the clock of critical child development periods when those are altered by adverse genetic effects. EXBF influences early life growth development and thus plays a critical role in preventing the risks of overweight and obesity, especially when those are exacerbated in the first of years of life due to genetic susceptibility or other predisposing conditions.

169 | The Association Between Telomere Length and BMI in Middle-Aged and Older Adults

Yan Yan Wu[1]

[1]Department of Public Health Sciences, University of Hawai'i at Manoa, Honolulu, Hawai'i, United States of America

Body Mass Index (BMI) classifications by World Health Organization (WHO) were developed based on associations between BMI and chronic disease and mortality risk in healthy populations. And these classifications are used in adults regardless of age and race/ethnicity. Aging is generally accompanied by weight loss, however how BMI trajectories change in older adults are not clear. In addition, recent research has found that telomeres are key markers of cellular and biological aging - each time a cell divides, the telomeres get shorter until it becomes "senescent" or it dies. This shortening process is associated with aging and higher risk of obesity, cancer, and mortality. Study of data from the Health and Retirement Study have found that: (1) there is a significant gender-and-racial/ethnic-interaction of BMI trajectories in Whites, Blacks and Hispanics. Hispanics and Blacks have higher BMI and faster declining rates compared to the Whites; (2) longer telomere length delays the timing of weight loss, i.e. the biological aging process. The study provides valuable insight of aging mechanics and sets the foundations for future gene-environmental studies.

170 | A Novel Region-Based Bayesian Approach for Rare Variant Association Test with Application to a Lung Cancer Study from Toronto

Jingxiong Xu[1,2], Wei Xu[1,3], Rayjean Hung[1,2], Geoffrey Liu[1,3], Helene Massam[4], Michael Escobar[1], Laurent Briollais[1,2]

[1] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [2] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [3] Princess Margaret Cancer Center, Toronto, Canada; [4] Department of Mathematics and Statistics, York University, Toronto, Canada

The discovery of rare variants associated with disease outcomes, such as cancers, is a very challenging issue in the field of human genetics. It could help deciphering the genetic basis of many complex human diseases and provide insights into chemoprevention and drug resistance mechanisms. Many approaches for rare variants lack efficiency since they are based on simple summary statistics on a test statistic that is only specified under the null hypothesis (score tests). As an alternative, we introduced a novel region-based statistic based on the Bayes Factor (BF). Both the null and alternative hypotheses are considered when evaluating the marginal likelihood of the data and some gain in efficiency is obtained by specifying a prior distribution for the rare variant count in a specific region, with different hyper-parameters for cases and controls under the alternative hypothesis. A permutation test is used to assess the distribution of the BF under the null hypothesis of no association. Our simulations studies showed that the new BF statistic outperforms popular methods such as the Burden test and SKAT (SNP-set Kernel Association Test) under most situations. In our real application to a sequencing study of lung cancer from Toronto, including 258 cases and 257 matched controls, the association with CHEK2 and TERT was more significant with the BF approach than with SKAT and Burden test. A marginal association with CLPTM1L was only found with the BF approach (P-value=0.04). Sensitivity analyses and further developments of the BF approach to multi-region NGS analysis will also be presented.

171 | SNP-Treatment Interactions of Cardiovascular Medications and Risk of Acute Coronary Syndrome Recurrence

Peng Yin[1], Andrea Jorgensen[1], Andrew Morris[1], Richard Turner[2], Richard Fitzgerald[2], Rod Stables[3], Anita Hanson[2], Munir Pirmohamed[2]

[1] Department of Biostatistics, University of Liverpool, Liverpool United Kingdom; [2] Department of Molecular & Clinical Pharmacology, University of Liverpool, Liverpool, United Kingdom; [3] Liverpool Heart and Chest Hospital, Liverpool, United Kingdom

To identify loci associated with response to cardiovascular drugs in cases of Acute Coronary Syndrome (ACS), we have undertaken a GWAS in 1470 patients recruited to a UK pharmacogenetic study. Patients were treated with a range of drugs including clopidogrel, statins and beta-blockers. Nevertheless, ~14% of patients had another cardiovascular event in five years of follow-up after hospital discharge, including Myocardial Infarction (MI), stroke or cardiovascular death.

We began by identifying clinical risk factors for response to cardiovascular drugs (time to recurrence of ACS) in a Cox proportional hazards regression model: age (p=1.3×10^{-14}), BMI (p=0.015), prior MI (p=2.1×10^{-8}), ACE inhibitor use pre-admission (p=0.0049), and aldosterone antagonist use (p=0.0058). Patients were genotyped using the lllumina OmniExpress array, and after quality control, the genotype scaffold was imputed up to the 1000 Genomes Phase I reference panel (all ancestries, March 2012 release).

We tested for SNP-treatment interactions for each cardiovascular drug with outcome under an additive dosage model after adjusting for clinical risk factors, main effects of SNP and treatment, and principal components to account for population structure. We identified an intergenic variant (rs4936159) with genome-wide significant (p<5×10^{-8}) evidence of interaction with clopidogrel: minor allele frequency 0.05, p=1.2×10^{-8}, hazard ratio (95% CI) 19.3 (7.0-53.3). The variant maps near *NCAM1*, which is involved in left ventricular remodelling. Nominal evidence of clopidogrel interaction was also observed for variants mapping in/near: *EVC2, CNR1, SV2B, LIG3/RFFL*. Our study highlights several biological candidate genes that are associated with response to cardiovascular drugs in ACS recurrence time.

172 | Multi-Variant Linear Regression Tests with Reduced Degrees of Freedom for Association Analysis of Common Variants

Yun Joo Yoo[1,2], Lei Sun[3,4], Andrew D. Paterson[3,5], Shelley B. Bull[3,6]

[1] Department of Mathematics Education, Seoul National University, Seoul, South Korea; [2] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea; [3] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [4] Department of Statistical Science, University of Toronto, Toronto, Canada; [5] Program in Genetics and Genome Biology, Hospital for Sick Children Research Institute, Toronto, Canada; [6] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

As an alternative to multiple single-SNP test statistics, combined analysis of multiple SNPs within a gene, or a specified region, is a natural and interpretable analytic strategy. Various test statistics have been advocated, including minimum P value, generalized Wald, variance-component, and principal-component methods. We investigate Multiple Linear Combination (MLC) test statistics for analysis of common variants under realistic trait models with Linkage Disequilibrium (LD) based on HapMap Asian haplotypes. MLC is a directional test that exploits LD structure in a gene to construct clusters of closely correlated variants, recoded independently of trait such that the majority of pairwise correlations are positive. It combines variant effects within the same cluster linearly, and aggregates cluster-specific effects in a quadratic sum of squares and cross-products, producing an asymptotic chi-squared statistic with reduced degrees of freedom (*df*) equal to the number of clusters.

By simulation studies of 1000 genes from across the genome, we demonstrate that MLC is unbiased under the null and a well-powered and robust choice among alternative methods across a broad range of gene architectures. The mean power of MLC is never much lower than that of other methods, and can be higher, particularly when there are multiple causal variants. Moreover, the variation in gene-specific MLC type1 error and power across 1000 genes is less than that of other methods, suggesting it is a complementary approach for discovery in genome-wide analysis. The cluster construction of the MLC test statistics helps reveal within-gene LD structure, allowing interpretation of clustered variants as haplotypic effects.

173 | Germline Copy Number Variations (CNVs) that Affect Genes and Relapse-Free Survival in Colorectal Cancer

Yajun Yu[1], Salem Werdyani[1], Georgia Skardasi[1], Jingxiong Xu[2], Konstantin Shestopaloff[3], Wei Xu[2,3], Elizabeth Dicks[4], Jane Green[1,5], Patrick Parfrey[4], Yildiz Yilmaz[1,4,6], Sevtap Savas[1,5]

[1]Discipline of Genetics, Faculty of Medicine, Memorial University, St. John's, Canada; [2]Department of Biostatistics, Princess Margaret Hospital, University of Toronto, Toronto, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [4]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University, St. John's, Canada; [5]Discipline of Oncology, Faculty of Medicine, Memorial University, St. John's, Canada; [6]Department of Mathematics and Statistics, Faculty of Science, Memorial University, St. John's, Canada

Identification of new prognostic markers can help with better control, treatment, and prognosis of colorectal cancer. Copy Number Variations (CNVs) are candidate prognostic markers, which are large DNA segments that exist in variable numbers among individuals. In this study, we aimed to examine the roles of germline CNVs in relation to risk of relapse in colorectal cancer.

We examined 495 patients recruited to Newfoundland Colorectal Cancer Registry. Genome-wide CNV profiles were predicted by CNV calling algorithms. Multivariable Cox-regression method adjusting for stage, microsatellite instability status, and tumor location was applied to test the associations between 75 genic CNVs and Relapse-Free Survival (RFS). Score test under the multivariable Cox model with time-varying coefficients was used to identify the CNVs with time-varying effects on RFS.

We identified two CNVs ($p<0.05$) in the Cox multivariable models. These CNVs were located in *TGFBR3* and *STEAP2* genes, where the patients with ≥1 copy of the variants had increased risk of relapse during the follow up. Interestingly, we also identified two additional CNVs associated with the risk of relapse within 3-years post-diagnosis ($p<0.05$), but not after that.

This study identified novel CNVs and genes that can biologically affect the risk of relapse in colorectal cancer patients. We also, for the first time, identified genetic variants that can potentially be early-relapse markers in colorectal cancer. Overall, if replicated, our results will have critical implications for clinic management as well as survival chances of colorectal cancer patients.

174 | A General Framework for Adaptive Set-Based Testing of Gene-Environment Interactions

Shirong Zhang[1], Juan Pablo Lewinger[1]

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America

Gene-Environment Interactions (G×E) play a critical role in complex diseases like cancer and diabetes, yet identifying specific G×E's requires exceedingly large sample sizes. One approach to increase the power to detect G×E's is to use set-based methods, which test the interaction between an exposure and a set of SNPs rather than individual variants. We propose a general Adaptive Combination Test (ACT) framework for aggregating multiple individual interaction signals to detect a global interaction of a set of SNPs with an exposure. The ACT framework, which flexibly extends previously proposed set based tests for gene main effects, adaptively combines individual SNP×E test statistics taking into the account the direction and magnitude of the estimated interaction effects, and can also incorporate two-step G×E screening statistics as weights. The framework can combine individual G×E statistics obtained by fitting any Generalized Linear Model (GLM) (e.g. logistic linear, Poisson) and can be easily extended to also handle Cox-regression. Because permutation is not a valid method for assessing the significance of interactions, we propose a fast simulation alternative based on the joint asymptotic distribution of the marginal test statistics of G×E GLMs fitted separately for each SNP. In a comprehensive simulation study we show that the ACTs protect the type I error rate even with large SNP sets containing hundreds or even thousands of SNPs, and have better power than existing G×E set-based approaches including burden type tests, variance component tests like GESAT, and mixed approaches like eSBERIA.

175 | Leveraging Cross-Disease Genetic Correlations and Large-Scale DNA-Linked Electronic Medical Records to Improve Risk Prediction of Disease

Xue Zhong[1, 2], Qiang Wei[2,3], Rui Chen[2,3], Nancy J. Cox[1,2], Bingshan Li[2,3]

[1] Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America; [2] Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America; [3] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America

There is a growing appreciation of the utility of Genetic Risk Scores (GRS) for detecting shared genetic etiology among diseases, despite the limited power of a single GRS in predicting the risk of a single disease. Here, we propose to leverage the potential cross-disease nature of such scores to circumvent limitations and improve the power of risk prediction by borrowing information from genetically-related diseases. Specifically, we calculated genetic risk scores of over 800 GWAS-reported diseases/traits in ∼10,000 patients of European ancestry from BioVU, a Vanderbilt BioBank linked to detailed Electronic Medical Records (EMR), and carried out a systematic phenome-wide scan of ∼1,800 phenotypes for these patients. Our analyses revealed pervasive genetic sharing between GWAS-ascertained diseases/traits and EMR phenotypes, with ∼200 GRS each being associated (nominal $p<0.001$) with three or more EMR phenotypes. As a proof of concept, we built a risk prediction model for Alzheimer's disease using all genetic risk scores and a logistic regression model with elastic net regularization. The resulting model demonstrated an improved prediction power with an AUC of 0.64, compared to an AUC of 0.58 from the basic model using only a single GRS based on current Alzheimer's-disease-biomarkers. Our approach has the potential to identify better biomarkers and may eventually lead to (personalized) risk prediction for a comprehensive set of diseases.

176 | A Novel Method to Detect Associations between Multiple Phenotypes and Genetic Markers

Huanhuan Zhu[1], Shuanglin Zhang[1], Qiuying Sha[1]

[1] Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

Many complex diseases like diabetes, hypertension, and metabolic syndrome are measured by multiple correlated phenotypes. However, most genome-wide association studies (GWAS) focus on one phenotype of interest or study multiple phenotypes separately for identifying genetic markers associated with complex diseases. Analyzing one phenotype or the related phenotypes separately may lose power due to ignoring the information obtained by combining phenotypes, such as the correlation between phenotypes. In order to increase statistical power to detect genetic markers associated with complex diseases, we developed a novel method to test the optimally weighted combination of phenotypes (TOW-P). This TOW-P method combines multiple phenotypes to one variable by using a linear combination of phenotypes. We performed extensive simulation studies to compare the TOW-P method with some of the existing methods. Our simulation results showed that TOW-P has correct type I error rates and is either the most powerful test or comparable to the most powerful ones among the methods we compared.