DOI: 10.1002/gepi.22256

Genetic Epidemiology
OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org
WILEY

ABSTRACTS

Invited Abstract

1 | Data Integration: Data-driven Discovery from Diverse Data Sources

Genevera Allen[1,2]

[1]*Department of Electrical and Computer Engineering and Departments of Statistics and Computer Science, Rice University, Houston, Texas, United States of America;* [2]*Jan and Dan Duncan Neurological Research Institute, Baylor College of Medicine, Houston, Texas, United States of America.*

Data integration, or the strategic analysis of multiple sources of data simultaneously, can often lead to discoveries that may be hidden in individual analyses of a single data source. In this talk, we present several new techniques for data integration of mixed, multi-view data where multiple sets of features, possibly each of a different domain, are measured for the same set of samples. This type of data is common in healthcare, biomedicine, national security, multi-senor recordings, multi-modal imaging, and online advertising, among others. In this talk, we specifically highlight how mixed graphical models and new feature selection techniques for mixed, multi-view data allow us to explore relationships amongst features from different domains. Next, we present new frameworks for integrated principal components analysis and integrated generalized convex clustering that leverage diverse data sources to discover joint patterns amongst the samples. We apply these techniques to integrative genomic studies in cancer and neurodegenerative diseases to make scientific discoveries that would not be possible from analysis of a single genomics data set.

2 | Impact of Reference Panel Choice for Imputation on Genome-wide Association Study Results for Type 2 Diabetes in Arab Population

Hossam Almeer[1]*, **Osama Alsmadi**[2], **Naser Elkum**[3], **Mohamad Saad**[1]

[1]*Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar;* [2]*King Hussein Cancer Center, Amman, Jordan;* [3]*Research Department, Sidra Medicine, Doha, Qatar.*

The prevalence of Type 2 Diabetes Mellitus (T2DM) varies substantially among ethnicities because of dissimilarity in lifestyle and genetic makeup. Genome-Wide Association Studies (GWAS) with T2DM have mostly been conducted in populations of European and Asian origin, and it remains to be confirmed whether they will be similar in the Arab population.

In this study, we conducted a GWAS for T2DM in a cohort of Kuwait-resident Arab population (498 cases and 1141 controls), genotyped on the Illumina HumanCardio-MetaboChip array. We evaluated the accuracy of imputation with Minimac3 and its impact on GWAS results using four reference panels: (1) 1000 Genomes Project data (1KG); (2) a public dataset of 108 healthy Qataris with Whole Genome Sequence (Q108); (3) the two previous datasets merged (1KG+Q108); and (4) a cohort of ~1000 Arab subjects, containing both healthy and T2DM patients, with Whole Exome Sequence.

Our preliminary results showed that the 1KG imputation generally led to the best performance across different Allele Frequency intervals. The Q108 imputation yielded similar performance to 1KG imputation for common variants despite its small size. Merging the two datasets decreased imputation performance.

For all imputations performed, we confirmed two T2DM susceptibility genes, *TCF7L2* (Lead SNP was rs34872471, *P* value = 9.58E-08, OR = 1.6, 95% CI = [1.34,1.9]) and *GLIS3* (Lead SNP was rs10814915, *P* value 2.57E-05, OR = 1.44, 95% CI = [1.22,1.71]). We also identified three genes, using 1KG imputation only, that could play a role in T2DM. These genes were *KCNE4*, *PTPRT*, and *PTRF*.

3 | Addressing the Missing Data Issue in Multi-phenotype Genome-wide Association Studies

Mila D. Anasanti[1]*, **Marika Kaakinen**[1,2], **Marjo-Riitta Jarvelin**[3,4,5,6], **Inga Prokopenko**[1]

[1]*Department of Genomics and Common Disease, Imperial College London, United Kingdom;* [2]*Department of Clinical and Experimental Medicine, University of Surrey, Guildford, United Kingdom;* [3]*Department of Epidemiology and Biostatistics, Imperial College London, United Kingdom;* [4]*Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial*

College London, London, United Kingdom; [5]Center for Life Course Health Research, University of Oulu, Oulu, Finland; [6]Unit of Primary Care, Oulu University Hospital, Oulu, Finland; [7]Biocenter Oulu, University of Oulu, Oulu, Finland.

Joint analysis of multiple phenotypes in genome-wide association studies (MP-GWAS) increases power for locus discovery but suffers from missingness in phenotype values. We investigated properties of missing data imputation methods within MP-GWAS, focusing on single and multiple-imputation (SI/MI) using Bayesian approach and expectation-maximisation bootstrapping (EMB), k-nearest neighbour (kNN), left-censored imputation method (QRILC) and random forest (RF). We simulated genetic data for 5,000/50,000/500,000 individuals using Hapgen2, and highly (r = 0.64) and moderately correlated (r = 0.33) phenotypes (3/9/30/120) for them. We randomly selected common, low-frequency and rare variants to be significantly ($P < 5 \times 10^{-8}$) associated with the simulated phenotypes. We considered several proportions of missing data (1/5/20/50%) under missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). We used the Root Mean Squared Error (RMSE) for the evaluations with complete cases (CC) vs. full data RMSE as the reference level. RF and MI-EMB diverged the most from the reference under MCAR ($P_{RF} = 7.18 \times 10^{-4}$, $P_{MI\text{-}EMB} = 6.27 \times 10^{-4}$). RF also outperformed under MAR ($P = 6.22 \times 10^{-4}$), whereas QRILC outperformed under MNAR ($P = 7.45 \times 10^{-5}$). RF was applied to the Northern Finland Birth Cohorts (NFBC) 1966 and 1986 (4,955 and 2,687 individuals, respectively) for the imputation of anthropometric and glycaemic measurements and 149 serum metabolite levels. MP-GWAS of 31 amino acids showed a novel association at *ADAMTS* after imputation with RF ($P = 2.61 \times 10^{-11}$ vs. $P = 5.68 \times 10^{-7}$ in CC) and improved power at *FCGR3B* ($P = 1.86 \times 10^{-9}$ vs. $P = 1.72 \times 10^{-8}$). We propose improved solutions for phenotype imputation in high-dimensional omics data-analyses and have implemented these into a user-friendly and computationally efficient imputeSCOPA software tool.

4 | Recurrency Approaches Using Random Forests to Identify Genetic Risk Factors While Controlling Family-wise False Positive Rates

Joan E Bailey-Wilson*, Jeremy Sabourin, Anthony M. Musolf, Emily R. Holzinger, James D. Malley

Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America.

Non-parametric machine learning methods are robust to complex models such as are hypothesized for complex genetic diseases and traits. Machine learning methods such as Random Forests (RF) can identify complex effects. We have developed an RF-based variable selection method called relative recurrency variable importance metric (r2VIM) to address hurdles in using RF to identify variables (features) that increase risk of complex phenotypes in genome-wide analyses.

Previously, we have shown that r2VIM improves false positive control and power to identify important features compared to RF, especially when non-linear effects contribute to the phenotype. Here, we use permutation of trait status to control family-wise error rates in variable selection to achieve a selection criterion that better scales with the number of SNPs evaluated and compare performance of several schemes for maximizing power to detect causal variants in simulated data. These data were simulated using GWAsimulator and the 1000 genomes EUR haplotypes to create a set of subjects with LD structure similar to actual EUR population, pruned to match the SNP density of a typical 1 M SNP chip and creating a trait due to: 1) one interaction between 2 SNPs with no main effects on maps of 10,000 to 100,000 SNPs and 2) both interaction effects and independent main effects. The effects of stepwise procedures on power are also being evaluated. Power is affected by multiple factors.

This recurrency approach has the potential to elucidate novel biological pathways which could improve both treatment and prediction of complex human diseases.

5 | Progress and Controversy in Analysis of Complex Phenotypes Based on Genome-wide Association Statistics

David Balding[1,2]*, John Holmes[1], Doug Speed[2,3]

[1]Melbourne Integrative Genomics, Schools of Biosciences and of Mathematics & Statistics, University of Melbourne, Australia; [2]Genetics Institute, University College London, United Kingdom; [3]Aarhus Institute for Advanced Studies, Aarhus University, Denmark.

Recently there has been much interest and great progress in statistical modelling of genome-wide association study test statistics for heritability analyses, genetic correlation estimates and to predict complex phenotypes, correcting for GWAS confounding bias if required. The LDSC model has been widely applied since published in 2015, but it's unrealistic implicit assumption of uniform expected heritability across SNPs is now recognized to have led to poor performance. The model has been further developed

(now called S-LDSC), including SNP-specific expected heritabilities that adjust for effects of minor allele fraction, linkage disequilibrium and genome annotations. We have proposed a different approach, implemented in the SumHer software, finding in some settings dramatically different inferences from those based on LDSC. As the models have improved, so have methods for model comparison, including an improved log likelihood approximation and an approach based on leave-one-chromosome-out prediction of summary statistics. We report latest developments, showing that as the models have improved, inferences are converging. However, we also report that the adjustment for GWAS confounding bias is unreliable in all current approaches even if the assumed heritability model is correct, so that summary-statistic analyses should be limited to settings in which the original GWAS adequately adjusted for confounding.

6 | Estimation of SNP-based Heritability in Multi-ethnic Studies

Saonli Basu

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America.

Heritability estimation provides important information about the relative contribution of genetic and environmental factors to phenotypic variation. Genome-wide complex trait analysis (GCTA) approach is now being routinely used to estimate SNP-based heritability for many complex traits. The accurate estimation of heritability through GCTA heavily depends on the accurate estimation of a high-dimensional genetic relatedness matrix (GRM) and even when the assumptions in GCTA are satisfied correctly, heritability estimation can be biased. Presence of subtle population substructures in the data could also severely impact heritability estimation. In fact, a more pertinent question is to define heritability in presence of population substructure. Here we model the stochastic nature of the GRM to estimate heritability of a trait in a multi-ethnic study, while accounting for the population structures in the data. We allow for different heritability parameters for each subpopulation while modeling the spatial dependency in the genetic relationships among the sampled individuals using reverse Haseman-Elston regression. We study the dynamic nature of heritability for height, BMI and substance-use related traits in NHLBI Trans-Omics for Precision Medicine cohort to investigate the impact of population substructure on SNP-based heritability estimation for low, moderate and highly heritable traits.

7 | A Framework for Transcriptome-Wide Association Studies in Breast Cancer in Diverse Study Populations

Arjun Bhattacharya[1]*, Montserrat García-Closas[2,3], Andrew Olshan[4,5], Charles M. Perou[5,6,7], Melissa Troester[4,7], Michael I. Love[1,6]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America; [3]Division of Genetics and Epidemiology, Institute of Cancer Research, London, United Kingdom; [4]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [7]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America.

The relationship between germline genetic variation and breast cancer survival is largely unknown, especially in understudied minority populations. Genome-wide association studies (GWAS) have interrogated breast cancer survival but are often underpowered due to many clinical covariates and subtype heterogeneity. GWAS often detect loci in non-coding regions, which require follow-up studies to interpret. Recent work in transcriptome-wide association studies (TWAS) show increased power in detecting functionally-relevant, trait-associated loci by leveraging expression quantitative trait loci (eQTLs) in external reference panels in relevant tissues. However, race-specific TWAS reference panels may be needed to draw correct inference in large, racially-heterogeneous cohorts, and such panels for breast cancer are lacking. Here, we provide a framework for TWAS for breast cancer in diverse populations, using data from the Carolina Breast Cancer Study (CBCS), a population-based cohort that oversampled self-identified African American women. We perform an eQTL analysis for 417 breast cancer-related genes to train race-stratified predictive models of tumor expression from germline genotypes. These race-stratified expression models are not always applicable across race, and their predictive performance varies across breast cancer subtypes. At a false discovery-adjusted P value less than 0.05, we identify hazardous associations near *CAPN13* (*2p23.1*) and *VAV3* (*1p13.3*) and a protective association near *IGF2BP2* (*3q27.2*) in TWAS that are underpowered in

GWAS in a CBCS sample of 3,828 women. This approach shows increased power to detect survival-associated genomic loci, demonstrating the relative strength of TWAS over GWAS. TWAS is an efficient approach for understanding the genetics underpinning breast cancer outcomes in diverse populations.

8 | Novel Association of G-quadruplex SNPs in Schizophrenia Candidate Genes with Cognition and Tardive Dyskinesia in a Schizophrenia Cohort

Upasana Bhattacharyya[1]*, Smita N. Deshpande[2], Bittianda Kuttapa Thelma[1]

[1]Department of Genetics, University of Delhi South Campus, New Delhi, India, Department of Psychiatry; [2]Postgraduate Institute of Medical Education and Research-Dr. RML Hospital, New Delhi, India.

Schizophrenia (SZ) is a neuropsychiatric disorder affecting ~0.5–1% of the population with high morbidity rate. SZ patients suffer from other complications such as cognitive impairment (CI) and antipsychotic-mediated movement disorder Tardive dyskinesia (TD) along with disease specific symptoms, leading to low quality of life. TD and CI are quantitative traits that show a huge genetic overlap with SZ. Thus, studying well implicated SZ genes might provide useful etiological insights. Quantitative traits are often linked to differential expression of genes. Gene expression is often affected by the folding kinetics of non-canonical secondary structures such as G-quadruplexes in DNA during transcription. SNPs present within these G-rich sequences (Quad-SNP) have found to modulate the phenotype by compromising the secondary structure in a dose dependent manner, making them a well-suited candidate for studying disorders with additive genetic model of inheritance such as CI and TD. In this study, we have systematically selected and genotyped using fluidigm SNPtype assay, 24 Quad-SNPs present in the promoter and 5′UTR of ~500 well known SZ candidate genes in a north Indian SZ case-control cohort phenotyped for TD (n = 80 cases, 95 controls) and cognition (n = 328 cases, 283 control). Association of two Quad-SNPs from *CACNA1I* and *SEZ6L2* (*P* value = 0.01 and 0.02 respectively) with TD and nine Quad-SNPs from *CUL3*, *ITIH1, PUS7, TCF32, NAB2, SMIM4, CHARNA5* and *ZNF804A* (*P* values = 0.006–0.04) were noted with one or more of the eight cognitive domains. Most of these genes have been implicated in TD/cognition based on association/animal studies but their causal relationship warrants further investigations.

9 | Integrative Omics Approach Identifies Association Between Dementia Risk and Non-coding Variants also Associated with Gene Expression in Brain

Elizabeth E. Blue[1]*, Timothy A. Thornton[2], Charles Kooperberg[3], Alexander P. Reiner[3,4]

[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]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]Department of Epidemiology, University of Washington, Seattle, Washington, United States of America.

Recent studies suggest genetic overlap between dementia and ischemic stroke, suggesting that analyses of dementia in the setting of ischemic stroke may uncover new risk loci influencing shared pathophysiological processes. Our analysis of whole-genome sequence on >11,000 participants in the Women's Health Initiative (WHI) study including 5,000 stroke cases, 1,600 dementia cases, and 1,700 women with brain imaging/cognitive function allow us to further explore shared biology between AD and these other neuro-vascular outcomes. Our multiethnic genome-scan (GWAS) identified significant associations between SNPs near *APOE* (P value $= 1 \times 10^{-26}$), *MYH11* (P value $= 1 \times 10^{-9}$), and *FZD3* (P value $= 4 \times 10^{-8}$) and risk of dementia, adjusting for age, ethnicity, stroke, and venous thromboembolism status using mixed-model logistic regression. The associations near *MYH11* and *FZD3* were robust to *APOE* adjustment and were not associated with stroke risk (ischemic, hemorrhagic, or all) in the WHI. GWAS for late-onset AD have not implicated *MYH11* or *FZD3*, previously, although both have been shown to be differentially expressed in several AD studies. Using an integrative omics approach, we show that these associated SNPs are also associated with the expression of *FZD3, CCDC25*, *DUSP4*, *NPIPA5*, and *NPIPA1* in AD-specific resources, and *FZD3*, *DUSP4*, and *NPIPA5* are differentially expressed in neurons derived from brain tissue with varying levels of AD pathology.

10 | Genomic Imprinting Analyses Reveal Maternal Effects to be a Cause of Genotypic Variability in Type 1 Diabetes and Rheumatoid Arthritis

Inga Blunk[1]*, Hauke Thomsen[2,3], Norbert Reinsch[1], Manfred Mayer[1], Asta Försti[2,4], Kristina Sundqvist[4], Jan Sundqvist[4,5], Kari Hemminki[2,4]

[1]Institute of Genetics and Biometry, Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, Germany; [2]Division of Molecular Genetic Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg,

Germany; [3]GeneWerk GmbH, Heidelberg, Germany; [4]Center for Primary Health Care Research, Lund University, Malmö, Sweden; [5]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, California, USA.

In the past, studies have indicated that the susceptibility to type 1 diabetes (T1D) and rheumatoid arthritis (RA) is influenced by genomically imprinted genes. Imprinted genes contribute to the class of parent-of-origin effects (POEs) as their alleles are epigenetically inactivated depending on their parental origin. However, findings have been contradicting and other POEs including maternal effects–which can be genetic and environmental–may have biased earlier findings.

A mixed model with two random gametic effects was used to investigate the importance of imprinting as well as the influence of maternal genetic and maternal environmental effects on the susceptibility to T1D and RA. Fixed effects including gender, birth year, and social economic index were analyzed. A suitable population database was available through linkage of the Swedish Hospital Discharge Register and the Multigeneration Register. For T1D the dataset contained 27,255 patients with 208,114 ancestors; for RA 15,850 patients with 60,684 ancestors were available.

The susceptibility to T1D did not turn out to be affected by imprinting but by maternal environmental effects explaining 18.8% (±1.81%) of the phenotypic variance. Regarding the susceptibility to RA imprinting effects were not significant but significant variances were found for maternal effects. However, in contrast to T1D, they could not clearly be determined to be either genetic or environmental. Fixed effects turned out to be significant underlining the importance of environmental factors. The results implicate that POEs other than imprinting effects may have biased earlier results and that they must be taken into account in future imprinting studies.

11 | Covariate Adjusted Permutation for Millions of Samples

Ryan J. Bohlender*, Yao Yu, Jiun-Sheng Chen, Chad D. Huff

Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America.

As samples become cheaper and biobank scale datasets become increasingly common, it is important that robust statistical tools are available for the analysis of those data. Covariate adjusted permutation allows for the calculation of correct *P* values for a variety of, potentially more powerful, statistical methods that have been underutilized due to concerns about computational complexity. We provide a memory efficient parallelizable algorithm for the permutation of large case-control sets, with associated covariates. Logistic regression is conducted to estimate individual odds ratios, which are used to sample case or control status from Fisher's non-central hypergeometric (FNCH) distribution. Permutation is conducted using a collapsing procedure where only individuals carrying the minor allele have their state permuted using the FNCH distribution, because only they contribute to the test statistic. Cases are randomly assigned to n + 1 categories, where n is the number of minor-allele carriers. The remaining category represents homozygous major allele carriers. As a result, we avoid permuting millions of samples, and instead permute a number on the order of tens to thousands for a rare-variant association study, while still providing improved control of Type-1 error. For rare variant association, we can leverage sparse matrices that drastically reduce the memory usage for genotype storage. Memory usage for one million samples is below 10 GB for 28 threads in the full program, providing an upper bound on the amount of memory used for the permutation process.

12 | Association of Polygenic Risk Scores for Body Mass Index and Systolic Blood Pressure in a Pediatric Cohort Requiring Surgery for Congenital Heart Defects

Joseph H. Breeyear[1]*, Jacob M. Keaton[1,2,3], Eric S. Torstenson[1,2], Andrew H. Smith[4,5], Sara L. Van Driest[4], Jeffery G. Weiner, Prince J. Kannankeril[5], Todd L. Edwards[1,2,3]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee; [2]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee; [3]Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, Tennessee; [4]Departments of Pediatrics and Medicine, Vanderbilt University Medical Center, Nashville, Tennessee; [5]Thomas P. Graham Jr. Division of Cardiology, Department of Pediatrics, Monroe Carell Jr. Children's Hospital at Vanderbilt, Nashville, Tennessee.

Congenital malformations are the leading cause of infant mortality in the US. Congenital heart disease (CHD) is the most prominent congenital malformation, affecting 40,000 US births per year, most children requiring surgical intervention(s). To date, no studies have evaluated associations of polygenic risk scores (PRS) for anthropometric traits with postoperative outcomes nor the utility of PRS to predict outcomes such as hypertension and obesity

in pediatric subjects. We used imputed genotypes from pediatric participants requiring surgery for CHD (mean age ± SD = 3.38 ± 5.39 years, n = 1,978 subjects). Base data for body mass index (BMI) PRS was the GIANT consortium GWAS 2015 BMI data (n_{max} = 322,154 subjects); the systolic blood pressure (SBP) PRS used published GWAS SBP data (n_{max} = 760,226 subjects). Calculating scores in PLINK, target data were pruned for linkage disequilibrium at an r^2 threshold of 0.1 at a maximum distance of 250 kilobases. Associations of PRS for BMI and SBP with BMI and length of hospital stay following surgery (LOS) were modeled using linear regression in R and adjusted for age, sex, and 10 PCs; SBP model adjusted for BMI. BMI PRS was associated with BMI (PRS *P* value threshold = 0.001, SNPs = 1,603, $\beta \pm SE$ = 0.58 ± 0.15 kg/m^2, *P* value 8.5x10^{-5}) and LOS (PRS *P* value threshold = 0.05, SNPs = 16,244, 1.67 ± 0.83 days, *P* value = 0.045). The SBP PRS was associated with LOS (PRS *P* value threshold = 0.0001, SNPs = 2,985, −0.45 ± 0.15 days, *P* value = 3.02×10^{-3}). Our results demonstrate the ability of PRSs developed in adults to predict pediatric traits and outcomes, with the BMI PRS with BMI association providing proof of principle.

13 | IMHOTEP–a Composite Score Integrating Popular Tools for Predicting the Functional Consequences of Non-synonymous Sequence Variants

Amke Caliebe[1]*, Carolin Knecht[1], Matthew Mort[2], Olaf Junge[1], David N. Cooper[2], Michael Krawczak[1]

[1]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom.

The *in silico* prediction of the functional consequences of mutations is an important goal of human pathogenetics. However, bioinformatic tools that classify mutations according to their functionality employ different algorithms so that predictions may vary markedly between tools. We therefore integrated nine popular prediction tools (PolyPhen-2, SNPs&GO, MutPred, SIFT, MutationTaster2, Mutation Assessor and FATHMM as well as conservation based Grantham Score and PhyloP) into a single predictor. The optimal combination of these tools was selected by means of a wide range of statistical modelling techniques, drawing upon 10 029 disease-causing single nucleotide variants (SNVs) from Human Gene Mutation Database and 10 002 putatively "benign" non-synonymous SNVs from UCSC. Predictive performance was found to be markedly improved by model-based integration, whilst maximum predictive capability was obtained with either random forest, decision tree or logistic regression analysis. Comparison of our approach with other integrative approaches such as Condel, CoVEC, CAROL, CADD, MetaSVM and MetaLR using an independent validation dataset, revealed the superiority of our proposed integrative approach. An online implementation of this approach, IMHOTEP ("Integrating Molecular Heuristics and Other Tools for Effect Prediction"), is provided at http://www.uni-kiel.de/medinfo/cgi-bin/predictor/. Currently, we are working on improving the practical usability of IMHOTEP.

14 | Next Generation MB-MDR: Taking the Challenge to Enhance Replication and Interpretation in Epistasis Studies for Complex Traits

Aldo Camargo[1]*, Junior Ocira[1], Diane Duroux[1], Francois Van Lishout[1], Kristel Van Steen[1,2]

[1]Biostatistic Biomedicine Bioinformatics – GIGA-R Medical Genomics, University of Liège, Liège, Belgium; [2]WELBIO researcher, University of Liège, Liège, Belgium.

Model-Based Multifactor Dimensionality Reduction (MB-MDR) is a genome-wide association approach to screen for interactions between categorical variables such as SNPs, although addressing some concerns related to first implementations of Multifactor Dimensionality Reduction (MDR). Since its conception, the approach has been encapsulated in an entire framework for epistasis detection with SNPs, although accommodating different trait types and study designs. Towards addressing issues of replication and interpretation of epistasis findings, we have developed several extensions to the data format, which allow defining new units of analysis such as those based on the integration of multi-omics data, or, exploiting meiotic recombination patterns in the data. Furthermore, the software release MBMDR.v5.0 introduces visualization tools and statistical tests to check marginal distributions of MB-MDR's test statistic and to check the validity of maxT implementations based on approximations via gamma distributions. It also facilitates expression quantitative trait loci (eQTL) epistasis analyses and has enhanced features for parallelized analyses. Via real-life data applications, we show how these new features not only enhance epistasis research but also opens up new avenues for set-based gene-mapping analyses with common and rare variants.

15 | An Efficient Identity by Descent Mapping Test for Biobank-scale Cohorts

Han Chen[1,2]*, Ardalan Naseri[2], Degui Zhi[1,2]

[1]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America.

In the past decade, genome-wide association studies (GWAS) have identified thousands of genetic loci associated with complex human diseases and quantitative traits. Most GWAS have focused on testing the associations with genotypes (e.g., the number of minor alleles for single nucleotide polymorphisms or copy number variations) from genotyping arrays or DNA sequencing, including common and rare genetic variations but ignoring the phased haplotype information. However, little is known about the roles of mid-range and long-range haplotypes on the genetic architecture of complex traits. Here we leverage the Identity by Descent (IBD) segments inferred from a random projection-based IBD detection algorithm to represent shared haplotypes between individuals, in the mapping of genetic associations with complex traits, and propose a computationally efficient statistical test for IBD mapping in biobank-scale cohorts. Simulation results show that our new method appropriately controls the type I error under the null hypothesis of no genetic association in large biobank-scale samples, and outperforms traditional GWAS approaches, especially when the causal variants are untyped, rare, and/or in linkage equilibrium with adjacent genotyped markers. We also apply our method to IBD mapping of multiple quantitative traits using real data from the UK Biobank.

16 | Transcriptomic and Exonic Profiles of Hispanic Individuals, Comparing Obese with Abnormally Low Triglycerides to Those of Normal Weight with Hypertriglyceridemia

Hung-Hsin Chen[1]*, Lauren E. Petty[1], Joseph McCormick[2], Susan P. Fisher-Hoch[2], Eric R. Gamazon[1,3], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, United States of America; [2]School of Public Health, University of Texas Health Science Center at Houston, Brownsville, TX, United States of America; [3]Clare Hall, University of Cambridge, Cambridge, United Kingdom.

Triglyceride levels usually increase with increase in body mass index, and both are associated with risk of cardiovascular disease and type 2 diabetes. Yet, the genetic pathways controlling triglyceride levels are not well defined. In this study, we examine genome-wide expression by applying exome-sequencing and RNA-sequencing. We compare a cohort of Hispanic individuals with high BMI and low triglycerides (<75 mg/dl) to those with normal weight and very high triglyceride levels (> = 300 mg/dl). Such extremely discordant phenotypes would be expected to be primarily under genetic control. We selected 14 individuals with normal weight but hypertriglyceridemia and 24 obese individuals with extreme low triglycerides from the Cameron County Hispanic Cohort. We applied DESeq. 2 to identify differentially expressed genes, and GATK was used for exome variants calling. We identified six genes with significant differential expression (FDR adjusted P value <0.05) between these two groups that also harbored rare variants (minor allele frequency <0.05 in all 1000 Genome populations) that may impact expression or function: *TRIM10*, *C8orf59*, *DBI*, *RPS14*, *MPHOSPH6*, and *RGS1*. Next, we annotated all variants observed in our sample using SeattleSeq Annotation 138 to determine functionality, highlighting one rare, missense variant in TRIM10, which creates a substitution from tryptophan to cysteine (PolyPhen score = 1). Methodologically, our study presents a novel transcriptome-based approach to prioritizing rare variants in exome-wide analysis with broad implications on detection and interpretation in exome sequencing studies. Our results implicate several genes and biochemical pathways that may be involved in genetic control of triglyceride levels, independent of obesity.

17 | PCSK9 Variants and Type 2 Diabetes Risk in People of African Ancestry: a Meta-analysis Study N = 30,000

Tinashe Chikowore[1,3]*, Anubha Mahajan[2], Michèle Ramsay[3], Andrew P. Morris [2,4,5]

[1]MRC/Wits Developmental Pathways for Health Research Unit, Department of Pediatrics, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [2]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK; [3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [4]School of Biological Sciences, University of Manchester, Manchester, UK. [5]Department of Biostatistics, University of Liverpool, Liverpool, UK.

Lifelong low-density lipoprotein cholesterol (LDL-C) reductions, coronary heart disease (CHD) protection and absence of circulating PCSK9 in loss-of-function (LOF) *PCSK9* variant carriers of African Ancestry, motivated the development of PCSK9 inhibitor drugs. However, studies that assess the risk for diabetes of PCSK9 inhibitor drugs, using LOF *PCSK9* variants as proxies, have been performed only in individuals of European ancestry. These studies have focused on R46L *PCSK9* variant, which is very rare in African ancestry populations. The association of type 2 diabetes (T2D) with large effect LDL-C lowering, LOF *PCSK9* variants, common in African ancestry individuals (C679X, A443T, L253S and Y142X), has yet to be evaluated.

Our study included 29,501 African Ancestry individuals (6,601 T2D cases and 22,900 controls) from five studies. T2D association summary statistics for C679X, A443T, L253S and Y142X *PCSK9* variants, adjusted for age, sex and body mass index (where available) were aggregated using fixed-effects meta-analysis in Metasoft. A Bonferroni-corrected threshold $P = 0.0125$ was considered significant.

The C679X variant was significantly associated with reduced T2D risk (OR = 0.73; $P = 1.1 \times 10^{-5}$). A443T, L253F and Y142X associations with T2D risk were not significant. Our findings of the protective effect of the C679X variant is contrary to reports of increased risk for R46L carriers in European ancestry populations. However, they support the previously reported association of C679X with low fasting glucose in black South Africans. Our study highlights the importance of genetic diversity in studies using LOF variants as proxies for potential side effects of drugs.

18 | Investigating the Use of Machine Learning Methods to Build Risk Prediction Models for Complex Disease

James P. Cook[1]*, John Y. Goulermas[2], Andrew P. Morris[1]

[1]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; [2]Department of Computer Science, University of Liverpool, Liverpool, United Kingdom.

Large-scale population biobanks offer exciting opportunities to develop risk prediction models for complex diseases because of the availability of genetic data and extensive lifestyle and clinical information. Unlike traditional polygenic risk scores, machine learning methods can be utilized to build risk prediction models that include both genetic and non-genetic features, and interactions between them.

We have performed a simulation study to assess the utility of several machine learning methods (gradient boosting machines, deep learning neural networks, and random forests) to generate prediction models for type 2 diabetes (T2D), applied using the H2O package, using data from the UK Biobank. Twenty thousand participants were randomly selected according to their T2D status (10,000 cases and 10,000 controls). Five relevant clinical factors (age, sex, body mass index, diastolic blood pressure and systolic blood pressure) were selected for entry into the model alongside a set of 1–100 SNPs, simulated with varying minor allele frequency and relative risk of disease. Irrelevant clinical factors were also selected to assess whether the methods identify them as unimportant for disease prediction.

All methods successfully identified the most strongly associated genetic and non-genetic factors as the most important features for prediction, and assigned the least importance to the irrelevant factors. Results also indicated that the inclusion of strongly associated genetic variants increases the predictive accuracy of the model compared to using clinical factors alone, while the inclusion of more modestly associated variants does not appear to improve predictive power.

19 | A Principal Component Approach to Polygenic Risk Scores to Avoid Over- and Underfitting

Brandon J. Coombes[1]*, Matej Markota[2], Sue L. McElroy[3], Mark A. Frye[2], Joanna M. Biernacka[1,2]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States of America; [2]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, United States of America; [3]Lindnor Center of HOPE/University of Cincinnati, Cincinnati, OH, United States of America.

Polygenic risk scores (PRSs) have become an increasingly popular tool in genetics to construct risk prediction models, to establish evidence of genetic effect when no genome-wide significant variants exist, and to establish common polygenic signals between two different traits. Proposed methods to construct PRSs use either pruning and thresholding, Bayesian shrinkage, or penalization approaches. Regardless of method, construction of PRSs require tuning parameters to optimize the prediction, which can result in overfitting. This can be guarded against by using validation data, cross-validation, or split-validation. For smaller studies, however, splitting the data can result in an underpowered analysis. Alternatively, one could *a* priori choose a single parameter setting to

construct the PRS, but this can lead to underfitting and, likewise, an underpowered analysis. Here, we propose computing PRSs under a grid of tuning parameter values, performing principal component (PC) analysis on the resulting PRSs, and keeping the first PC for association testing. The first PC achieves the highest signal-to-noise ratio and can thus have the effect of concentrating much of the signal of the different possible settings into this PC. We compare the performance of the PRS-PC approach with optimization and *a* priori selection of tuning parameters to test for association of a variety of PRSs with subphenotypes from two bipolar disorder datasets. This comparison is performed over a variety of PRSs constructed from summary statistics from the largest studies of psychiatric disorders and related traits. We find that the PRS-PC approach outperforms the other strategies in most scenarios.

20 | A Novel Locus Identified in Chromosome 14 of Mouse Modulates Lens Weight

Jennifer B. Cordero[1]*, Robert W. Williams[1], Lu Lu[1], Claire L. Simpson[1,2]

[1]*Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, United States of America;* [2]*Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, TN, United States of America.*

Abnormalities of the size and shape of the lens will affect ocular refractive power and impair vision. In previous work, we mapped eye size of 700 mice to define quantitative trait loci (QTL) that modulate eye size and retinal area, but failed to detect any loci specifically controlling lens mass. Here, we exploited the power of new advanced mapping methods and improved genotypes to remap our original lens data and to define a novel lens-specific locus. Lens weight was measured in 122 cases from 26 BXD strains and parents C57BL/6 J and DBA/2 J. Measurements were corrected for variance associated with sex and age. The original study used Haley-Knott mapping methods and about 300 markers. Our reanalysis exploited the Genome-wide Efficient Mixed Model Association software with leave-one-chromosome-out scheme as well as 7,000 markers in GeneNetwork2. We uncovered a locus that affects lens weight in the mouse but has no detectable effect on overall eye weight. It maps to chromosome 14 between 58.2–63.5 Mb (LOD 4.8). We identified two candidate genes *Fgf9* and *Ctsb*. We also detected a secondary locus at chromosome 5 (LOD 3.7) which aligns with a locus that controls overall eye weight in mouse and myopia in humans. The chromosome 5 locus therefore has a global influence on both eye and lens whereas the newly discovered locus on chromosome 14 is lens-specific. The discovery of a locus modulating lens weight may contribute to the understanding of genetics of lens development, coordination of ocular growth and disorders caused by structure abnormalities.

21 | QTL Remapping of Murine Eye Weight Reveals Novel Candidate Genes for Ocular Growth

Roberto Y. Cordero[1]*, Robert W. Williams[1], Lu Lu[1], Claire L. Simpson[1,2]

[1]*Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, TN, United States of America;* [2]*Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, TN, United States of America.*

The global prevalence of myopia is growing and may affect almost half of the world's population by 2050. A pioneering study of mouse eye weight used interval mapping to locate quantitative trait loci (QTLs) that control normal variation in the architecture of the eye, lens, and retina in laboratory mice and found novel QTLs *Eye1* and *Eye2*. In this study, we increased the sample size, resulting in a four-fold increase in strains and 16-fold increase in cases. Cases totaling 11,761 from 112 BXDs and progenitors (C57BL/6 J and DBA/2 J), had an average age of 200 days (Mean eye weight = 23.9 ± 0.2 mg). Measurements were corrected by multiple linear regression analysis to statistically control for covariance between eye weight and variables such as body weight, sex, and age. QTL mapping was conducted using 2017 BXD genotypes of GeneNetwork employing a linear mixed model with leave one chromosome out approach. Candidate genes were compared to known myopia genes in the Consortium for Refractive Error and Myopia (CREAM) study. We performed QTL remapping of adjusted eye weight using the new 2017 BXD Genotypes and found a significant locus on Chromosome 19 from 53–58 Mb (LOD 4.9). We found strong candidate genes *Shoc2, Tcf7l2* and *Dclre1a* which are well expressed in the eye and associated with a very significant cis-eQTL. A GWAS study by CREAM implicates *TCF7L2* in myopia. Our findings illustrate the power of using enhanced bioinformatics tools and new mouse genotypes in mapping to improve localization of QTLs and identify promising genes.

22 | Developing a Genetic Risk Index for Peanut Allergy

Sidney Liu[1], Ying Yi[1], Aida Eslami[1], Yuka Asai[2], Andrew Sandford[1], Ann E Clarke[3] and Denise Daley[1,4]*

[1]Centre for Heart and Innovation, St. Paul's Hospital, Vancouver, Canada; [2]Division of Dermatology, Department of Medicine, Queen's University, Kingston, Canada; [3]Division of Rheumatology, Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. [4]Division of Respiratory Medicine, Department of Medicine, University of British Columbia, Vancouver, Canada.

Background: Over 200 single nucleotide polymorphisms (SNPs) have been found to be associated with food allergy (FA) in genome-wide association studies (GWAS). A Genetic risk score (GRS), can be derived from GWAS to summarize genetic risk, risk stratification and/or prediction. Our objective was to use information from the Canadian Peanut Allergy Registry (CanPAR) GWAS study to develop a GRS using the weighted sum of the number of risk alleles (with values 0/1/2) by the natural log of their respective odds ratio (OR). The positive predictive value of the GRS was evaluated in the Canadian Asthma Primary Prevention Study (CAPPS). Models were fit with and without principal components. Using a p-value threshold of 1.49E-06 and LD pruning we identified 25 independent SNPs for use in the GRS, of these 13 SNPs were identified in the CAPPS study. We then evaluated the area under the curve (AUC) which is used to determine the effectiveness of the classification and the positive predictive value (PPV).

Results: A summary of GRS risk models is shown in Table 1.

Conclusions: Only minimal differences were observed between models with 25 SNPs vs.13, and preliminary findings indicate replication of the model in the CAPPS study.

23 | Statistical Interaction and Mendelian Randomization: Are They The Same?

Mariza de Andrade, PhD

Division of Biomedical Statistics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America.

There are several similarities between statistical interaction as gene-environmental (GXE) interaction and mendelian randomization (MR). The approaches used for both look similar when genetic variants are used to test hypotheses towards the role of the environmental exposures. However, this is not true if the association between the phenotype under study (e.g. low cholesterol levels) and increased cancer rates is resulted through reverse causation. To better understand the causality of this particular association one has to compare the genotypes of the apolipoprotein E (APOE) gene between cases and controls since the APOE is responsible for the reduction or clearance of cholesterol from plasma that depends on the different APOE genotypes. If there are other confounding from other sources such as smoking status or low-fat diet, the Mendelian Randomization should be used instead of GXE. I will present different scenarios involving real applications using GXE and MR.

24 | Towards an Accurate Cancer Diagnosis Modelization: Comparison of Random Forest Strategies

Ahmed Debit[1,2]*, Christophe Poulet[1], Claire Josse[3], Chloé-Agathe Azencott[4], Guy Jerusalem[3], Kristel Van Steen[2], Vincent Bours[1,5]

[1]University of Liege, GIGA-Research, Laboratory of Human Genetics, Liege, Belgium; [2]University of Liege, GIGA-Research, Medical Genomics, BIO3, Liege, Belgium; [3]University Hospital (CHU), Department of Medical Oncology, Liege, Belgium; [4]Centre for Computational Biology (CBIO) of Mines ParisTech, Institut Curie and INSERM, Paris, France; [5]University Hospital (CHU), Center of Human Genetics, Liege, Belgium.

TABLE 1 Summary of GRS risk models

		#of SNPs selected			Statistical measures	
p-value threshold	Study	Genotyped	Imputed	Total	AUC (95% CI)	PPV
1.49E-06*	CanPAR	6	19	25	0.878 (0.862–0.894)	0.8313570
1.49E-06	CanPAR	5	8	13	0.638 (0.613–0.664)	0.6062246
1.49E-06	CAPPS	10	3	13	0.579 (0.404–0.754)	0.6666667

Machine learning approaches are heavily used to produce models that will one day support clinical decisions. To be reliably used as a medical decision, such diagnosis and prognosis tools have to harbor a high-level of precision. Random Forests have been already used in cancer diagnosis, prognosis, and screening. Numerous Random Forests methods have been derived from the original random forest algorithm. Nevertheless, the precision of their generated models remains unknown when facing biological data. The precision of such models can be therefore too variable to produce models with the same accuracy of classification, making them useless in daily clinics. Here, we perform an empirical comparison of Random Forest based strategies, looking for their precision in model accuracy and overall computational time. An assessment of 15 methods is carried out for the classification of paired normal - tumor patients, from 3 TCGA RNA-Seq datasets: BRCA (Breast Invasive Carcinoma), LUSC (Lung Squamous Cell Carcinoma), and THCA (Thyroid Carcinoma). Results demonstrate noteworthy differences in the precisions of the model accuracy and the overall time processing, between the strategies for one dataset, as well as between datasets for one strategy. Therefore, we highly recommend testing several random forest strategies prior to modeling. This will certainly improve the precision in model accuracy while revealing the method of choice for the candidate data.

25 | Identification of Selective Sweeps Through Deep Learning in Whole Genome Sequenced Malaria Parasites

Wouter Deelder[1,2]*, Ernest Diez Benavente[1], Jody Phelan[1], Susana Campino[1], Luigi Palla[1§], Taane G. Clark[1§]

[§]*Equally contributing authors*

[1]*London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom;* [2]*Dalberg Advisors, 7 Rue de Chantepoulet, CH-1201 Geneva, Switzerland.*

Background: Whole-genome sequencing (WGS) is increasingly applied to Plasmodium malaria isolates to identify genetic determinants of malaria pathogenesis. The detection of genomic regions under selection pressure has revealed the genetic determinants underpinning drug resistance, as well as candidate vaccine targets. Typically, "positive" selective sweeps are detected by using population genetic methods that find genetic regions with extended-haplotype homozygosity. However, these approaches can be computationally expensive and the results may differ depending on which SNPs are included, the definition of populations, and statistical significance thresholds.

Methods: We apply a machine "deep" learning approach using a Convolutional Neural Network to identify extended haplotype signals in WGS data. We evaluate if this deep-learning method, which does not require prior extraction or selection of population-genetic features, can classify selective sweeps of *P. falciparum* (n = 1,950) and *P. vivax* (n = 830) WGS isolates.

Results: The application of this method, after training on loci known to be associated with drug resistance, revealed a new set of loci for both *P. falciparum* and *P. vivax* that are potentially under positive selection pressure.

Conclusion: Overall, we have shown that a deep learning approach holds potential for the detection of positive selection signatures in malaria parasites. This method may play a supporting role in resistance surveillance for malaria and likely has wider applications outside the field of malaria.

26 | Epigenome-wide Association Study of Immunoglobulin E Levels Using High-resolution Dna Methylation Profiling

Laura Beaumier[1], Sébastien Chanoine[1], Warren Cheung[2], Anaïs Malpertuis[1], Emmanuelle Bouzigon[3], Miriam F. Moffatt[4], Isabelle Pin[1], Florence Demenais[3]*, Elin Grundberg[2], Valérie Siroux[1]

[1]*Team of Environmental Epidemiology, Institute for Advanced Biosciences, Institut National de la Santé et de la Recherche Médicale U1209, Université Grenoble Alpes, Grenoble, France;* [2]*Center for Pediatric Genomic Medicine, Children's Mercy Kansas City, Kansas City, Missouri, United States of America;* [3]*Team of Genetic Epidemiology and Functional Genomics of Multifactorial Diseases, Institut National de la Santé et de la Recherche Médicale UMR1124, Université de Paris, Paris, France;* [4]*Section of Genomic Medicine, National Heart and Lung Institute, London, United Kingdom.*

Asthma displays important phenotypic heterogeneity. Allergy features, mediated by Immunoglobulin E (IgE) levels, define a major endophenotype of asthma. The identification of differentially methylated CpGs (DMC) or regions (DMR) associated with IgE levels in asthmatics may bring further insight into asthma heterogeneity.

High-resolution methylome data in regulatory elements of immune cells obtained by MethylC-Capture Sequencing (2.8 million CpGs) were analysed in association with IgE levels in 599 blood samples from asthmatic subjects of the Epidemiological study on the Genetics and Environment of Asthma (EGEA). A binomial mixed model was fitted for each CpG, considering the proportion of methylated reads weighted for sequence read

coverage as the dependent variable and log(IgE) as predictor while adjusting for age, sex, smoking and proportions of leucocytes. Associations reaching a *P*-value $\leq 10^{-4}$ (N = 317 DMCs) were followed-up in 194 subjects from the Medical Research Council for Eczema (MRCE) study. The results from the two datasets were meta-analyzed.

Meta-analysis of EGEA and MRCE results identified a DMC at chr21:45628025 associated with IgE at genome-wide significance (*P* value = 8.5×10^{-9}). The DMC maps to an immune-cell specific distal regulatory element 20 kb downstream of *ICOSLG*. In addition, we found suggestive associations (*P* value $< 5 \times 10^{-6}$) at 13 other loci. Further investigation of DMRs, based on combining spatially correlated DMC *P*-values, showed that out of 15 DMRs detected in EGEA, the one replicating in MRCE included *ICOSLG*.

This study identified a novel IgE-associated DMC, nearby *ICOSLG*, a gene of major biological relevance for allergy but not yet reported by genome-wide studies of allergic diseases.

27 | An Online Platform for Densely Imputed GWAS Summary Statistics

Ayse Demirkan[1,2]*, Liudmila Zudina[3], Zhanna Balkhiyarova[1,4], Marika Kaakinen[1,4], Inga Prokopenko[1,4]

[1]Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, United Kingdom; [2]Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands; [3]Department of Biomedical Engineering, Imperial College London, London, United Kingdom; [4]Department of Genomics of Common Disease, Imperial College London, London, United Kingdom.

Genotype imputation is a necessary step in the Genome-Wide Association Study (GWAS) meta-analyses combining summary statistics of primary GWAS. A survey of publicly available data in the GWAS Catalog and the GRASP database showed that only less than 30% of summary statistics are imputed to the 1000 Genomes (1KG)/Haplotype Reference Consortium (HRC) variant density, with the majority being imputed to the HapMap reference panel. After GWAS meta-analyses are published, genotype imputation is cumbersome and often impossible due to restricted access to individual level data. We implemented a recently published approach, ssimp to publicly available HapMap-based GWAS summary statistics and we propose to increase their usage by introducing a new web-based platform containing a resource of GWAS summary statistics for multiple phenotypes imputed to 1KG. This removes the necessity for summary statistics re-formatting, parallelizing the processes and preparing the scripts by each researcher in need to use summary statistics imputation. Starting with traits associated with glucose homeostasis, including HOMA-IR, HOMA-B, fasting glucose and insulin, imputed to all-ancestries 1KG, our online resource is constantly growing. The average number of SNPs in HapMap-based summary statistics is 2.5 million before imputation and reaches an average of 8.5 million after imputation to 1KG after stringent quality control. We are in the process of releasing initial set of GWAS summary statistics imputed to the European-ancestry HRC panel.

Our platform will advance the research by simplifying the analytical procedures required in both GWAS meta-analyses and other analyses based on summary statistics.

28 | A Powerful and Versatile Colocalization Test

Yangqing Deng*, Wei Pan

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, United States of America.

Testing colocalization has become increasingly popular in establishing causal relationships: if a GWAS trait and a gene's expression share the same causal SNP, it may suggest a regulatory role of the causal SNP through gene expression to the trait. Accordingly, it is of interest to develop and apply various colocalization testing approaches. The existing approaches all have some severe limitations. For instance, in some methods the null hypothesis to be tested is that there is colocalization, which may not be ideal when often the null hypothesis cannot be rejected due to limited statistical power (with too small sample sizes). Some other methods impose a strong restriction on the number of causal SNPs in a locus, which may lead to loss of power in the presence of wide-spread allelic heterogeneity. Importantly, most methods cannot be applied to either GWAS/eQTL summary statistics or cases with more than two possibly correlated traits. We develop a simple and general approach based on conditional analysis of a locus on multiple traits, overcoming many shortcomings of the existing methods. We demonstrate that compared with other methods, our new method can be applied to a wider range of cases and perform better in certain scenarios, using both simulated and real data, including a large-scale Alzheimer's disease GWAS summary dataset and a gene expression dataset, and a large-scale blood lipid GWAS summary association dataset. An R package will be publicly available.

29 | Epigenome-wide Association Study of Change in Blood Metabolite Levels From Young- to Middle Adulthood in the Northern Finland Birth Cohort 1966

Harmen Draisma[1,2]*, Igor Pupko[3], Liudmila Zudina[4], Valeria Goffert[5], Mila D. Anasanti[2], Mika Ala-Korpela[6,7], Marjo-Riitta Jarvelin[7,8], Natalia Pervjakova[9], Marika Kaakinen[1,2,10], Inga Prokopenko[1,2]

[1]Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, United Kingdom; [2]Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, United Kingdom; [3]Department of Medicine, Imperial College London, London, United Kingdom; [4]Department of Life Sciences, Imperial College London, London, United Kingdom; [5]University of Tartu, Tartu, Estonia; [6]Baker Heart and Diabetes Institute, Melbourne, Australia; [7]Center for Life Course Health Research, University of Oulu, Oulu, Finland; [8]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [9]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [10]Centre for Pharmacology and Therapeutics, Department of Medicine, Imperial College London, London, United Kingdom.

Several associations between metabolite levels and *DNA methylation* (DNAm) have been reported. However, the relationship between *longitudinal* changes in metabolite levels and differential DNAm is underexplored. We assessed associations between epigenome-wide blood DNAm and change in blood metabolomics-based variables. For 595 non-diabetic individuals from the Northern Finland Birth Cohort 1966 for whom nuclear magnetic resonance-based metabolomics data were available at both *ages 31* (T1) and *46* (T2) as well as concurrent blood DNAm data at T2, we calculated for each of the 228 metabolomics-based variables the average change in level per year between T1 and T2. We used our methylSCOPA software, which enables both longitudinal and multi-phenotype *epigenome-wide association studies* (EWAS), for single-phenotype EWAS of change residuals – corrected for sex – for each metabolomic variable versus the degree of DNAm for 832,569 markers on the Illumina (San Diego, CA) MethylationEPIC BeadChip. We quality-controlled, residualized, and normalized the DNAm data, and mapped genomic locations to CGCh37/hg19. We detected 67 epigenome-wide significant ($P < 1 \times 10^{-7}$) associations between a DNAm site and change in the level of a metabolomic variable, involving 28 unique DNAm sites and 53 unique metabolomic variables. Nine DNAm sites associated significantly with change in the levels of multiple metabolomic variables, and change in the levels of five metabolomic variables associated with multiple DNAm sites. When looked together, effects of these 28 DNAm sites formed four robust clusters of effects on metabolite levels. Using a novel powerful *methylSCOPA* approach, we demonstrated that longitudinal changes in blood metabolite levels are associated with DNAm.

30 | Host Genome-wide Association Study of Infant Susceptibility to Shigella-associated Diarrhea

Dylan Duchen[1]*, Rashidul Haque[2], Genevieve Wojcik[3], Laura Chen[1], Poonum Korpe[1], Beth Kirkpatrick[4], William A. Petri[5], Priya Duggal[1]

[1]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States of America; [2]International Center for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh; [3]Stanford University, Stanford, CA, United States of America; [4]University of Vermont, Burlington, VT, United States of America; [5]University of Virginia, Charlottesville, VA, United States of America.

Shigella is a leading cause of moderate-to-severe diarrhea in African and South Asian children and the causative agent of shigellosis and dysentery. Associated with 80 - 165 million cases of diarrhea and up to 600,000 deaths annually, exposure to shigella is ubiquitous in many regions while colonization or infection is heterogenous. To characterize the host-genetic susceptibility to shigella-associated diarrhea, we performed two independent genome-wide association studies (GWAS) including 589 Bangladeshi infants, 429 from the PROVIDE birth cohort and 160 infants from a Cryptosporidium-focused study birth cohort in Dhaka, Bangladesh. We classified children as ever having shigella associated diarrhea or not in the first 13 months of life. A qPCR Ct distribution of the ipaH gene, carried by all four shigella species and enteroinvasive E. coli, identified a total of 143 infants with a shigella-associated diarrheal event and 446 infants with no evidence of shigella-associated diarrhea within their first 13 months of life. Host GWAS's were performed using the Illumina Infinium 5 Multiethnic Global Array and analyzed under an additive genetic model. A joint analysis (imputed variants N = 6,547,362) identified loci of interest on chromosomes 11 (rs582240, within the *KRT18P59* pseudogene, average MAF = 29.4%, $P = 8.37 \times 10^{-8}$) and 8 (rs12550437, within the lincRNA *RP11–115J16.1*, average MAF = 38.1%, $P = 1.69 \times 10^{-7}$). This study suggests host genetic factors may influence the response to shigella colonization and pathogen-associated diarrhea. Additional replication and further research on the function of these genes and their association with

shigellosis and other pathogen-associated diarrheal diseases is warranted.

31 | Improving Efficiency in Epistasis Detection with a Gene-based Analysis Using Functional Filters

Diane Duroux[1,§]*, Hector Climente-González[2,3,4,§], Aldo Camargo[1], Lars Wienbrandt[5], David Ellinghaus[5], Chloe-Agathe Azencott[4,2,3], Kristel Van Steen[1,6]

[§]Equally contributing authors

[1]BIO3 - GIGA-R Medical Genomics, University of Liège, Liège, Belgium; [2]Institut Curie, PSL Research University, F-75005 Paris, France; [3]INSERM, U900, F-75005 Paris, France; [4]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France; [5]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany; [6]WELBIO researcher, University of Liège, Liège, Belgium.

Reproducibility and interpretability are acknowledged concerns in Genome-Wide Association Interaction Studies, due to biological and methodological variability. Increasing the exploitation of functional information in these studies may alleviate some of these concerns. In this context, we took real-life data on Inflammatory Bowel Disease (IBD) from the International IBD Genetics Consortium as case study. After quality control (including HWE p0.001; MAFand LD pruning at r^2 0.75), 66,280 samples and 38,225 SNPs remained. We then supplemented epistasis testing, including via Model-Based Multifactor Dimensionality Reduction, with pre- and post-analysis functional knowledge.

We employed a modified version of FUMA to functionally map and annotate the available SNPs to genes, based on several physical, expression and chromatin interaction information protocols. Then Biofilter_2.4 was used as a search engine to find candidate interacting gene pairs, supported by evidence in at least 3 databases called by Biofilter (without using trait information). Back-traced SNP-pairs to Biofilter prioritized gene-pairs were submitted to subsequent epistasis analysis, adjusting for main effects and confounders, and correcting for multiple testing. Epistasis results thus obtained, and epistasis results obtained by exhaustively screening the 38,225 SNPs and back-traced pool of SNPs were presented as statistical epistasis networks with nodes representing genes and edge weights computed by aggregating over SNPxSNP interactions (truncated p-value product) when contributing to retained Biofilter gene-pairs. Gene communities were analyzed for their enrichment in Reactome pathways. Our preliminary results indicate increased robustness and plausibility of epistasis findings in IBD when using pre- and post-functional knowledge over exhaustive screening, across analytic strategies.

32 | A Powerful Gene-set Analysis Method Identifies Novel Associations and Improves Interpretation in Uk-Biobank

Diptavo Dutta[1,2]*, Seunggeun Lee[1,2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States of America; [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, United States of America.

Biologically functional gene-set association (GSA) analysis can be complementary to single-variant or single-gene test and can provide insights into the genetic architecture of complex diseases. Existing GSA methods have low statistical power when only a small fraction of the genes are associated with the phenotype. Additionally, since most of the existing methods cannot identify active genes interpreting results is challenging. We introduce Gene-Set association Using Sparse Signals (GAUSS), a method for GSA with summary statistics which additionally selects the subset of genes with maximum association signals. The simulation-based p-value for GAUSS can be efficiently calculated by using pre-computed correlation structure of test statistics from a reference data. Numerical experiments show that GAUSS can increase power over several existing methods while controlling type-I error. We analyzed summary statistics from the UK-Biobank data for 1,201 phenotypes with 10,679 gene-sets to demonstrate that GAUSS can identify novel associations across a large number of phenotypes and gene-sets. For example, GAUSS detected two fatty-acid related gene-sets associated with E.Coli-infection (p-values $<10^{-06}$) which illuminate the antibacterial role of fatty-acids. Additionally, GAUSS allows us to investigate active genes for different phenotypes using phenome-wide analysis of a given gene-set, which has been unexplored until now. For example, the association of ATP-binding-cassette transporter gene-set in KEGG database with digestive diseases like Cholelithiasis is driven by *ABCG5* while that with Celiac disease is driven by *TAP2*. The novel associations detected by GAUSS along with the information on active genes and its computational scalability make it an attractive choice to perform GSA.

33 | Optimal Two-phase Designs in Practice: Considerations and an Illustration

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

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

Two-phase design and analysis are cost-reduction techniques that can be used to collect expensive molecular data (G) in a subset of a large GWAS cohort for post-GWAS investigation thereby improving cost-efficiency under budgetary limitations. At phase 1, a GWAS on a quantitative trait, Y, determines a genomic region of interest with a corresponding top-GWAS SNP as a surrogate covariate, Z. In the second phase, G is measured in a subsample selected according to values of Y and Z, thus making G missing by design for the remaining subjects. Lastly, inference on G via semiparametric maximum likelihood benefits from using available data from phases 1 and 2.

Motivated by recent developments in two-phase designs for statistical fine-mapping, we propose strategies to determine a phase 2 subsample at the design stage when investigators can postulate a range of hypothesized design factors (i.e. genetic effects, minor allele frequencies, linkage disequilibrium patterns). These strategies select the subsample by combining information across the range of postulated factors under the proposed optimal designs in two ways: 1) by using a min-median approach and 2) by resampling based on the frequency in which each subject is selected across designs/factors. We illustrate these strategies using the Northern Finland Birth Cohort 1966 comprised of 5,402 subjects from the two northernmost provinces of Finland. We argue that two-phase studies can drastically reduce the costs of gathering molecular data without substantial loss of power to detect genetic associations.

34 | Using External Controls to Account for Mating Asymmetry in Maternal Genetic Association

Joycelyne E. Ewusie[1]*, Kelly Burkett[2], Marie-Hélène Roy-Gagnon[1]

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

In studying the effects of maternal and child genes on the risk of a disease with early onset, current analytic approaches depend on the assumption of mating symmetry when using case-parent triad data. Mating symmetry refers to the assumption that for any possible parental genotype pair, the frequency of a given mother-father genotype assignment in the population is equal to the reverse genotype assignment. The violation of this assumption (mating asymmetry) may lead to spurious maternal associations in studies of case-parent triads. Our study modified the hybrid design (HD) approach by using control parents from an external dataset. Simulations were performed in the context of a real dataset of orofacial cleft case-parent triads. We used different levels of mating asymmetry (MA) and sample sizes (n) of external control-parent dyads to assess the effect on the type 1 error and power to identify maternal effects. Results from our simulation study showed that the type 1 error rate was around 5% when the MA in the external control-parent dyads was no more than ± 0.15 points from the MA in the case-parent triads. The hybrid design with external controls however performed worse than case-parent triad analysis when the MA was more than ± 0.3 points and $n(\text{control-dyads}) \geq n(\text{case-triads})$. Power of the hybrid design using external controls was above 80% for most scenarios considered. In conclusion, the HD with external controls can account for MA and provide valid tests for maternally contributed genotype effects when symmetry assumption fails.

35 | Role of Functional Non Coding Variants in the Germline DNA in the Ovarian Cancer Predisposition

Suzana A.M.Ezquina[1]*, Ed Dicks[1], Rosario I. Corona[2], Kate Lawrenson[2], Simon Gayther[2], Michelle R. Jones[2], Matthew L. Freedman[3], Ronny Drapkin[4], Paul Pharoah[1]

[1]Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]The Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA, United States of America; [3]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, United States of America; [4]University of Pennsylvania, Philadelphia, PA, United States of America.

Known rare ovarian cancer susceptibility alleles are in the coding sequence whereas most common risk alleles are in the non-coding genome. Rare non-coding variation may also be important.

Statistical power limits the detection of individual rare susceptibility variants. Power is increased by using burden testing, which has identified several ovarian cancer genes. The aim of this study is to find rare, non-coding ovarian cancer susceptibility alleles using burden testing on regulatory elements.

We focussed on PAX8 binding sites – PAX8 is a transcription factor expressed in normal ovarian surface and fallopian tube epithelia. High expression of PAX8 is a marker of high grade serous ovarian cancer.

We identified high-fidelity PAX8 binding sites by the overlap of PAX8 ChIPSeq peaks in two normal fallopian tube cell lines and three ovarian cancer cell lines. These were further refined by selecting those peaks that are in active chromatin regions defined by at least 3 of 5

H3K27acetylation ChIPSeq experiments of high grade serous ovarian cancer cell lines.

We called variants in these regions in germline genomes of 247 ovarian cancer patients and 1102 non-cancer controls sequenced for the UK 100k Genomes Project. Variants with a frequency >0.01 were excluded. There were 56 cases with at least one rare PAX8 binding site variant and 173 controls 22% v 15%, P = 0.011, OR = 1.57, CI 1.12–2.21).

Our results suggest that burden testing across non-contiguous regions of the non-coding genome is a promising approach to the identification for uncommon and rare disease susceptibility alleles.

36 | Multi-phenotype Transcriptome-wide Association Study (TWAS) Tests Using Summary Statistics

Helian Feng[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.

There is a great interest in multi-trait genetic association tests, which have been well developed for single SNPs. Here, we extend these methods to multi-SNP TWAS tests to improve power of detecting genes associated with phenotypes regulated through similar pathways. We show that the TWAS test statistic for multiple phenotypes has the same form as the single SNP statistic, replacing the Z-score vector from single SNP tests for multiple traits with Z-scores from TWAS. Thus, established methods for combining single-SNP test statistics across multiple traits can be easily extended to the TWAS case, including SUM, Wald, and ASSET tests. We evaluated several such methods in simulation under different alternatives (different covariance and effect sizes among the phenotypes). Out tests have proper Type I error (when SNPs are not associated with any of the traits). Our results suggest that conducting TWAS with multiple phenotypes jointly improves the power of TWAS. The simulation showed improvement in power compared to individual non-combined tests and a simple combined test with Bonferroni correction. However, we observed no uniform supreme multi-trait method, since the power of each method varies across different alternatives. The Wald test was near-optimal in most of scenarios. We then jointly analyzed the TWAS results from the Global Lipids Genetics Consortium of 4 traits (LDL, HDL, TG, and total cholesterol) with cross-tissue weights built with sparse canonical correlation analysis on GTEx gene expression data. The joint analysis identified additional trait-associated genes and provided new information into the gene regulation architecture for these traits.

37 | Case-only Design to Investigate Interactions Between Genetic Factors and Tobacco Smoke in Patients with Aggressive Periodontitis

Sandra Freitag-Wolf[1]*, Arne S. Schäfer[2]

[1]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; [2]Charité - University of Medicine, Berlin, Germany.

The common disease periodontitis has a complex etiology: genetic susceptibility variants and smoking play important roles. We investigated interactions between genes and smoking (G × S) that affect the susceptibility for aggressive periodontitis by a powerful case-only approach. In short, under two assumptions, namely (i) the disease is sufficiently rare (i.e. prevalence <5%) and (ii) G and S are uncorrelated in the general population, any association between S and G in the cases, points towards a G × S interaction at the population level. We used imputed genotypes of genome-wide data of the OmniExpress BeadChip that included 851 aggressive periodontitis (AgP) patients (650 from Austria and Germany, 171 from The Netherlands) and compared never vs. ever smokers. Thereby, we identified 16 loci for which the G × S interaction analysis suggested association with $P < 5 \times 10^{-5}$, nine of which were within the same topologically associated domains as SNPs that were previously reported to be associated with smoking related traits and phenotypes. Moreover, we analyzed these 16 loci in the case-control design and compared both approaches. In conclusion, we demonstrated that the genetic predisposition to severe early-onset forms of periodontitis is in parts triggered by smoking. Specifically, we suggest genetic variants in the genes *SSH1* and *ST8SIA1* to increase the disease susceptibility for AgP by interaction with cigarette smoke. Our results will advance our understanding of the molecular mechanisms, especially the interplay of alveolar bone homeostasis, smoking and inflammation against the background of genetic susceptibility.

38 | Applications of Multidimensional Time Model for Probability Distribution Function and Time Scales to Investigations in the Immune System Behavior

Michael Fundator[1,2]*

[1]Division of Material Physics, National Academy of Sciences, Engineering, and Medicine, Washington DC, United States of America; [2]Division of Behavioral and Social Sciences and Education, National Academy of Sciences, Engineering, and Medicine, Washington DC, United States of America.

Characterization of immune systems behavior during the epidemic and vaccination by chaotical and bifurcation type transformations requires application of Multidimensional Time Model for Probability Distribution Function and Time Scales to investigation. This abstract presented by the winning World Championship in Multidimensional Time Model author presents different aspects of Multidimensional Time Model for Probability Distribution Function and Time Scales analysis to investigations in the immune system behavior with some historical introduction and its application to immune system investigation.

39 | Ordered Multinomial Regression for Genetic Association Analysis of Ordinal Phenotypes

Christopher A. German[1]*, Janet S. Sinsheimer[1,2,3], Hua Zhou[1], Jin J. Zhou[4]

[1]Department of Biostatistics, University of California Los Angeles Fielding School of Public Health, Los Angeles, CA 90095, United States of America; [2]Department of Human Genetics, David Geffen School of Medicine at University of California Los Angeles, Los Angeles, CA, United States of America; [3]Department of Biomathematics, David Geffen School of Medicine at University of California Los Angeles, Los Angeles, CA, United States of America; [4]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, United States of America.

Genome-wide association studies (GWAS) query the entire genome to identify genetic variants associated with complex phenotypes. When the trait is binary, as in case-control studies, the primary analysis method is logistic regression. For multi-category traits, multinomial regression is a natural extension of logistic regression. Modern studies often generate complex phenotypes with ordered, discrete values. Prime examples include (1) subtypes defined from multiple sources of clinical information and (2) derived phenotypes generated by specific phenotyping algorithms for electronic health records. A common strategy divides these phenotypes into two categories and carries out traditional case-control GWAS. The choice of cutoff is arbitrary and different cutoff values may generate inconsistent findings. Multinomial regression, however, ignores trait value hierarchy, potentially losing power. As a solution, we use ordered multinomial regression to analyze ordinal trait GWAS data. Exploiting the ordering in phenotype values, this approach lies between quantitative trait and binary case-control studies, increases power, and generates interpretable results. We derive efficient algorithms for computing test statistics, making GWAS computationally practical for Biobank scale data. We provide a software package, OrdinalGWAS.jl, in the dynamic programming language Julia. Application of our method to the COPDGene study confirms previously found signals based on binary case-control status, with the *P* value of the top signal an order of magnitude smaller than that from the logistic model. Additionally, we demonstrate the ability of our package to analyze very large data sets using UK Biobank data.

40 | Association Mapping of Multivariate Phenotypes in the Presence of Missing Data

Saurabh Ghosh*, Kalins Banerjee

Human Genetics Unit, Indian Statistical Institute, 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. Majumdar, Witte and Ghosh (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 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 or count phenotype, the trait value is estimated using a conditional normal or 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-a-vis the ideal situation with no missing data as well as analyses based only on individuals with complete data. We illustrate an application of our method using data on Coronary Artery Disease.

41 | Whole-exome Sequencing and Protein Interaction Networks to Prioritize Candidate Genes for Susceptibility to Melanoma

Sally Yepes[1], Margaret A. Tucker[1], Hela Koka[1], Cancer Genomics Research Laboratory[1,2], Bin Zhu[1,2], Belynda Hicks[1,2], Neal D. Freedman[1], Stephen J. Chanock[1], Xiaohong R. Yang[1], Alisa M. Goldstein[1]*

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, United States of America; [2]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America.

Known cutaneous melanoma (CM) genes account for melanoma risk in 20–40% of melanoma-prone families, suggesting the existence of additional high-risk genes. Whole exome sequencing (WES) of high-risk families often produces many potential candidates thus requiring prioritization strategies to identify the top genes. We explored the framework of protein-protein interaction networks to investigate prioritization of potential candidate genes associated with CM predisposition. We conducted WES on 98 patients from 27 CM-prone families. Three hundred sixty-four variants that were rare (<0.1% in public and in-house control datasets), loss-of-function or missense predicted to be deleterious and showed disease co-segregation were included in the network analysis. We used known CMM genes (gene list in Goldstein et al. Hum Mol Genet, 2017) as seed proteins. We then applied three different network interactomes (to minimize the impact of network selection on gene prioritization) based on the network propagation concept to rank genes and explore gene modules. Degree-aware algorithm (DADA) was used to rank the set of candidate genes; HotNet and GeneMANIA tools were used to identify modules or subnetworks within the filtered dataset. The rankings of the candidate genes were fairly consistent using different sources of interactome data. Top genes included both known CM genes and potentially new genes that are unknown in CM susceptibility, and were involved in modules that reflect telomere biology, cell cycle, and DNA repair processes. These results highlight the importance of known CM genes and their protein interaction networks in CM susceptibility and demonstrate the value of network approaches in gene prioritization.

42 | Circulating Sex Hormone Levels and DNA Methylation in Blood – an Analysis of Repeated Samples from Men

Justin Harbs[1]*, Sabina Rinaldi[2], Robin Myte[1], Xijia Liu[3], Bethany Van Guelpen[1,4], Sophia Harlid[1]

[1]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden; [2]International Agency for Research on Cancer, Section of Nutrition and Metabolism, Lyon, France; [3]Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden; [4]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden.

The incidence of colorectal cancer (CRC) is higher in men than in women across all world regions, suggesting inherent differences (e.g. hormonal variation) as a likely explanation. The estrogen receptors (ER-α and ER-β) are significantly expressed by circulating leukocytes, especially B-cells. In addition, estrogen mediated mechanisms have been suggested to modulate the immune response, including inflammatory changes that could be important for long term CRC risk. We hypothesize that transcriptional regulation of steroid receptor expression, and in turn the sex-hormone axis, likely includes epigenetic changes such as DNA methylation. Using data on 77 male participants from the Västerbotten Intervention Program (VIP), this study therefore aimed to investigate whether levels of circulating sex hormones, including: testosterone, androstenedione, progesterone, estradiol, estrone, dehydroepiandrosterone and sex hormone binding globulin (SHBG), are associated with altered DNA methylation in circulating immune cells. Sex hormone levels were measured in plasma samples and DNA methylation levels were measured in white blood cell (buffy coat) DNA using the 850K Illumina Infinium MethylationEPIC BeadChip. Each study subject had donated two blood samples collected ten years apart, making it possible for us to monitor time dependent changes. Possible associations were estimated using mixed effect models, adjusted for cell composition and confounders (BMI, smoke status, batch variations etc.). As expected, levels of androstenedione, testosterone, dehydroepiandrosterone and progesterone decreased significantly over the 10-year interval. Preliminary analyses also identified a number of CpG sites significantly associated with circulating sex hormone levels, and efforts to replicate these findings are currently underway.

43 | An African Ancestry Uterine Fibroids Polygenic Risk Score (PRS) Identifies Associations with Other Gynecologic Conditions in the Clinical Phenome

Jacklyn N. Hellwege[1,2]*, Jacqueline A. Piekos[2], Yanfei Zhang[3], Eric S. Torstenson[2,4], the Electronic Medical Record and Genetics (eMERGE) Network, Sarah A. Pendergrass[3], Dan M. Roden[4], Josh C. Denny[2,5], Todd L. Edwards[2,6], Digna R. Velez Edwards[2,5,6]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [3]Genomic Medicine Institute, Geisinger, Danville, Pennsylvania, United States of America; [4]Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center Nashville, TN, United States of America; [5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States of America; [6]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, United States of America.

Uterine fibroids (UF) are the most common female pelvic tumor with prevalence up to 77% by menopause in African Americans. African American women have increased risk of UF than other continental populations, as well as larger and more numerous UFs. The known genetic architecture of UF currently includes ~30 loci, most discovered in European populations. We conducted a phenome-wide association study (PheWAS) of UF polygenic risk scores (PRS) to understand the shared genetic contributions across clinical phenotypes. We constructed an African-ancestry UF PRS of 317 SNPs using a P threshold of $<1 \times 10^{-4}$. Effect sizes were derived from imaging-confirmed UFs genome-wide association study (GWAS) of 1,272 cases and 1,379 controls from the Electronic Medical Records and Genomics (eMERGE) network. We then performed PheWAS across 6,061 independent women from eMERGE using the PRS as the predictor for clinical diagnoses adjusted for age, body mass index, and principal components. We identified 26 significant ($P < 2.5 \times 10^{-5}$) results among 1,381 diagnoses. The top association was UFs ($P = 6.42\text{x}10^{-66}$), showing that the PRS is well-calibrated even when not restricted to imaging-confirmed cases and controls. Other top results included genitourinary conditions related to UFs, including six diagnoses related to menstrual dysregulation. There were also novel associations with other gynecological conditions including endometriosis ($P = 2.51 \times 10^{-13}$), ovarian cysts ($P = 1.26 \times 10^{-7}$), and abnormal Papanicolaou smear ($P = 1.55 \times 10^{-6}$). Overall, results suggested that this imaging-confirmed UFs PRS is a valid correlate of the genetic risk underlying UF development. We also observed associations with other gynecologic conditions, including neoplastic phenomena, suggesting shared genetic risks.

45 | Leveraging Genetic Ancestry for New Insights into Complex Traits in Admixed Populations

Andrea R.V.R. Horimoto[1]*, Nora Franceschini[2], Timothy A. Thornton[1,3]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [2]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States of America; [3]Department of Statistics, University of Washington, Seattle, Washington, United States of America.

Genetic studies in multi-ethnic cohorts offer great potential to elucidate the genetic factors influencing complex traits. A variety of statistical methods have recently been developed to overcome special challenges for whole genome association analysis of complex traits in large-scale multi-ethnic cohorts. With existing methodology, however, genetic ancestry differences among sampled individuals are often treated as a confounder to be adjusted for in an analysis to protect against spurious association. Leveraging genetic ancestry can provide improved complex trait mapping in multi-ethnic populations, such as African Americans and Hispanic/Latino populations, who have admixed ancestry derived from multiple continents. We have developed mixed effects models for admixture mapping that incorporate both local and global ancestry for the identification of genetic loci influencing complex traits will be presented. The proposed mixed model admixture mapping methods have been developed for continuous and dichotomous traits and are completely applicable to large-scale whole genome studies with multi-ethnic samples from a variety of study designs. We demonstrate the utility of leveraging genetic ancestry for complex trait mapping in applications to the Hispanic Community Health Study / Study of Latinos (HCHS/SOL) to elucidate the genetic determinants of phenotype traits associated to chronic kidney disease. Our proposed mixed model admixture approach identifies novel loci that are not identified via genetic association mapping.

46 | Discovery of Pleiotropic Variants Associated with Multiple Sclerosis and Migraine

Mary K. Horton[1,3]*, Sarah C. Robinson[1], Xiaorong Shao[1], Hong Quach[1], Diana Quach[1], Kalliope H. Bellesis[2], Terrance Chinn[2], Catherine A. Schaefer[2], Lisa F. Barcellos[1,3]

[1]Genetic Epidemiology and Genomics Laboratory, Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley, CA, United States of America; [2]Kaiser Permanente Division of

Research, Oakland, CA, United States of America; [3]Computational Biology Graduate Group, University of California, Berkeley, Berkeley, CA, United States of America.

Risk factors and symptoms of multiple sclerosis (MS) and migraine often overlap, and up to 69% of MS patients experience migraine. We investigated whether genetic risk variants previously identified from GWAS of either condition could be contributing to both conditions (i.e., whether any variant exerted a pleiotropic effect). Data from 1,073 MS cases and 12,000 controls (white, non-Hispanic) from Kaiser Permanente Northern California were utilized. Migraine status was obtained through self-report and a validated electronic health record algorithm. Genotyping used Illumina microarrays on saliva or blood. MS or migraine SNPs were identified from prior GWAS, and after quality control, 902 SNPs with MAF > 1% were available for analysis. A method by Lutz et.al. (2017) was used to identify pleiotropic SNPs in which observed *P* values were compared to genotype permuted *P* values for both phenotypes. Significant SNPs were subsequently used in logistic regression models to estimate the association between each variant and phenotype, adjusting for ancestry. To account for ascertainment bias from using a case-control study secondary phenotype, the migraine model adjusted for a propensity score representing the probability of case-control status given covariates. Preliminary results showed five SNPs were significantly associated with MS and migraine. Three were protective for MS and all increased odds of migraine. Implicated genes include CD58 which modulates regulatory T-cells and several immune genes within the 6q23 chromosomal region. Results illuminate the shared genetics of MS and migraine and, because several variants increase risk of migraine but decrease risk of MS, implications for targeted therapies.

47 | Causal Inference for Highly Pleiotropic Biomarkers Using Mendelian Randomization and Bayesian Networks

Richard Howey*, Heather J. Cordell

Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, NE1 3BZ, United Kingdom.

Mendelian randomization (MR) is a popular tool for performing causal inference in genetic epidemiology, but it can have limitations for evaluating simultaneous causal relationships in complex data sets. We consider a scenario that may arise when analyzing biomarker data as generated from modern "omics" technologies which is highly pleiotropic and violates one of the required assumptions of MR. However, Bayesian network analysis (BN) offers an alternative approach. In BN, the conditional dependencies and independencies of variables are described by a graphical model (a directed acyclic graph) and its accompanying joint probability, which may be estimated using individual-level data.

We perform computer simulations to investigate the utility of BN in this situation. We simulate data using previously estimated real effect sizes of 150 genetic variants on 12 biomarkers which are assigned to have either an effect, no effect or a reverse effect. As well as BN and MR we evaluate several other recently-proposed causal inference methods: multivariable MR based on Bayesian model averaging (MR-BMA), a multi-SNP mediation intersection-union test (SMUT) and a latent causal variable (LCV) test. Our results showed that BN outperformed all other methods in terms of high power to detect an effect in the correct direction, while maintaining low type I error. As expected, MR had high power but also very high type I error due to pleiotropy.

We conclude that BN is a useful complementary approach to existing methods for performing causal inference in complex data sets such as those generated from modern "omics" technologies.

48 | GWAS of the Postprandial Triglyceride Response Yields Common Genetic Variation in Hepatic Lipase (*LIPC*)

Dorina Ibi[1,2]*, Raymond Noordam[3], Jan Bert van Klinken[1,2], Ruifang Li-Gao[4], Renée de Mutsert[4], Dennis O. Mook-Kanamori-[4,5], Frits R. Rosendaal[4], Martijn E.T. Dollé[6], Patrick C.N. Rensen[2,7], Ko Willems van Dijk[1,2,7]

[1]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [2]Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; [3]Department of Internal Medicine, Leiden University Medical Center, Leiden, The Netherlands Division of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; [4]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands; [5]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands; [6]National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands; [7]Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands.

Aim: Increased serum triglyceride (TG) concentrations in response to a meal are considered a risk factor for cardiovascular disease. We aimed to elucidate the genetic background of postprandial TG response through genome-wide association studies (GWAS).

Methods: Participants of the Netherlands Epidemiology of Obesity (NEO) study (n = 5,630) consumed a liquid mixed meal after an overnight fast. GWAS on fasting TG and postprandial serum TG at 150 minutes were performed. To identify genetic variation of postprandial TG independent of fasting TG, we defined the postprandial TG response as the residuals of a nonlinear regression that predicted postprandial TG concentration at 150 minutes as a function of fasting TG concentrations. Using the identified variants as determinants, we additionally performed linear regression analyses on the residuals of the postprandial response of 149 nuclear magnetic resonance (NMR)-derived metabolic measures.

Results: We identified rs7350789-A (allele frequency = 0.36), mapping to hepatic lipase (*LIPC*), to be associated with a lower TG response (P value = 5.1×10^{-8}). Furthermore, rs7350789-A was associated with responses of 23 metabolic measures (P value $<1.4 \times 10^{-3}$), mainly with decreased response of the TG component in almost all high density lipoprotein (HDL) sub-particles (largest effect on HDLTG, P value = 4.5×10^{-7}), increased response of HDL diameter (P value = 2.8×10^{-4}) and decreased response of most components of very low density lipoproteins (VLDL) sub-particles (largest effect on VLDLC, P value = 2.7×10^{-6})

Conclusion: GWAS on the TG response identified variants mapped to *LIPC* as the main contributor to postprandial TG metabolism. Additionally, this variant reduces HDLTG and decreases VLDLC response.

49 | Best Practices to Integrate Transcriptome Data with Gwas Studies to Understand the Biology of Complex Traits

Alvaro N. Barbeira, Rodrigo Bonazzola, Eric R. Gamazon, Yanyu Liang, Yoson Park, GTEx Consortium, Christopher D. Brown, Ayellet Segre, Xiaoquan Wen, Hae Kyung Im*

GWAS have been very successful in identifying genetic loci that are robustly associated with complex traits. However, the majority of the loci are located in non-coding regions rendering the downstream mechanism difficult to pinpoint.

The complex trait genetics field come to the general consensus that many of these loci are acting through the regulation of gene expression traits, including total mRNA levels and splicing. Enrichment of eQTLs (variants associated with expression levels), and other molecular traits related to open chromatin added support to an important role of transcription.

To assign target genes of GWAS discoveries using transcriptome reference data, several methods have been developed. These can be classified into association, colocalization, and proximity-based methods.

I will cover association methods such as PrediXcan, SMR, and TWAS; colocalization methods such as enloc, coloc, and ecaviar. Using data from the latest freeze of the GTEx consortium and thousands of GWAS studies, I will provide a comparison of the relative advantages of different approaches and propose a set of best practices for integration of transcriptome data into GWAS studies. Best practices will necessarily evolve as new methods and datasets become available. I will also provide a list of limitations of existing approaches which should motivate future method development opportunities.

50 | A Recall-by-genotype Pilot Study to Assess the Effects of Common *TMPRSS6* Variants on Oral Iron Absorption

Momodou W. Jallow[1,2]*, Susana Campino[2], Andrew Prentice[1,2] and Carla Cerami[1]

[1]The Nutrition Theme, Medical Research Council Unit The Gambia at London School of Hygiene & Tropical Medicine, Atlantic Road Fajara, P.O. Box 273 Banjul, The Gambia; [2]Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom.

Background: Anaemia is a global health problem, and iron supplementation is the routine treatment and prevention strategy. However, this is often ineffective in low- and middle-income countries. SNPs in the *TMPRSS6* gene have been associated susceptibility to anaemia, but the impact of these SNPs on poor iron response has not been prospectively investigated.

Objectives: To investigate the effects of common *TMPRSS6* SNPs on oral absorption in healthy individuals.

Methods: Participants were recruited based on rs4820268 and rs2235321 as follows: 1) homozygous for the variant allele at each SNP, 2) homozygous for wild type alleles, and 3) heterozygous at the two SNPs simultaneously (double heterozygotes). Each participant was bled at baseline, received 400 mg ferrous sulfate, and was bled at 1-hour, 2-hour, 5-hour and 24-hours post supplementation. Plasma hepcidin and iron biomarkers were measured and the effect of genotype group on iron absorption was assessed.

Results: Significant differences were observed between double heterozygotes and wild types for TSAT (transferrin saturation) at 2 hr ($p = 0.0131$) and 5 hr ($p = 0.0035$), serum iron at 2 hr ($p = 0.025$) and 5 hours ($p = 0.007$), and UIBC at 2 hr ($p = 0.008$) and 5 hr ($p < 0.0001$). No significant differences were observed between the two extreme genotypes of individual SNPs at different time points.

Conclusion: Our results suggest that individuals carrying multiple risk alleles might not respond efficiently to supplementation. Analysis of a larger number of individuals and more genetic markers in the *TMPRSS6* gene other iron-related genes may provide a more robust insight into the effect of genetic variations on iron absorption.

51 | Cohort Study of Serum Bisphenol A, Polygenetic Risk Score, and Thyroid Cancer in Korea

Keum Ji Jung, Sun Ha Jee*

Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea.

This study investigated the relationship between serum Bisphenol A (BPA) levels and thyroid cancer in a blood-based cohort study and whether the relationship was modified by genetic risk. Data for both case-cohort and case-control designs came from prospective cohort, which were built in 2004–2013. Thyroid cancer was confirmed by December 2016 in conjunction with the national cancer registry database. The number of samples used in the final analysis was 152 thyroid cancer and 354 healthy controls. All subjects were genotyped for thyroid related 48 SNPs. Serum BPA was divided into three groups and five groups. The lowest group was used as a reference and the association of each group with thyroid cancer was examined. In case-cohort analysis, the hazard ratio (95% CI) of the thyroid cancer at the highest tertiary of serum BPA was 2.12 (1.42–3.16), and increased to 3.16 (1.71–5.84) when the initial 3-year follow-up was removed. In the analysis of serum BPA divided by quintiles, reverse causation was observed, but when the initial 3-year follow-up was removed, the reverse causation was not significant and the hazard ratio (95% CI) of the highest thyroid cancer was 2.83 (1.30–6.16) times higher. Serum BPA increased the risk of thyroid cancer, and this association was modified by the genetic risk of thyroid disease. *This research was supported by a grant (18162MFDS121) from Ministry of Food and Drug Safety in 2018

52 | Hierarchical Modeling Framework for Mendelian Randomization and Transcriptome-wide Association Approaches for Correlated SNPs and Intermediates

Lai Jiang*, Shujing Xu, David V. Conti

Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America.

Previous research has demonstrated the usefulness of hierarchical modeling for incorporating a flexible array of prior information in genetic association studies. When this prior information consists of effect estimates from association analyses of SNP-intermediate or SNP-gene expression, the hierarchical model is equivalent to a two-stage instrumental or transcriptome-wide association studies (TWAS) analysis, respectively. Here, we propose to extend our previous approach for the joint analysis of marginal summary statistics (*JAM*) to incorporate prior information via a hierarchical model (*hJAM*). In this framework, the use of appropriate effect estimates as prior information yields an analysis similar to Mendelian Randomization (MR) and TWAS approaches such as S-PrediXcan. *hJAM* is applicable to multiple correlated SNPs and multiple correlated intermediates to yield conditional estimates of effect for the intermediate on the outcome, thus providing advantages over alternative approaches. We investigate the performance of our model in comparison to existing MR approaches (e.g. inverse-variance weighted MR, multivariate MR, and MR with Egger regression) and existing TWAS approaches (e.g. S-PrediXcan) for effect estimation, type I error and empirical power. Across numerous causal simulation scenarios, *hJAM* is unbiased, maintains correct type-I error and has increased power. We apply *hJAM* to two applied analyses: 1) estimation of the conditional effects of body mass index (BMI), asthma, smoking, and type 2 diabetes on myocardial infarction; and 2) investigation of the impact of gene expression on prostate cancer.

53 | Genome-wide Analysis of Non-completion of Controlled Exercise Trials in Sedentary Adults

Rong Jiang[1]*, Kim M. Huffman[2,3], Elizabeth R. Hauser[2,4], Monica J. Hubal[5], Cris A. Slentz[2,3], Johanna L. Johnson[2], Michael Babyak[1], Redford B. Williams[1], Ilene C. Siegler[1], William E. Kraus[2,3]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America; [2]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, United States of America; [3]Department of Medicine, Duke University Medical Center, Durham, North Carolina, United States of America; [4]Department of Biostatistics, Duke University Medical Center, Durham, North Carolina, United States of America; [5]Department of Kinesiology, Indiana University Purdue University Indianapolis, Indianapolis, IN, United States of America.

As a mainstay of clinical lifestyle medicine, exercise offers a wide-range of health benefits. Despite recognizing its benefits, many individuals discontinue exercise programs. We posit this behavior has biological underpinnings. Here, we used a GWAS approach to

identify putative genetic variants in STRRIDE (Study of a Targeted Risk Reduction Intervention through Defined Exercise) subjects initiating a well-characterized exercise training trial (n = 603). Non-completion occurred when a subject withdrew from further participation in the study. Non-completion was associated with a cluster of SNPs with the top hit of rs722069 (C/T, risk allele = T) (unadjusted $P = 2.2 \times 10^{-7}$, OR = 2.23) in a linkage disequilibrium block on chromosome 16. Rs722069 was an eQTL of the *EARS2, COG7* and *DCTN5* genes in skeletal muscle tissue in GTEx. In subsets of the STRRIDE sample with available muscle expression (n = 37) and metabolic data (n = 82), the T allele was associated with lower muscle expression of *EARS2* and *COG7* ($P < 0.036$), and lower concentrations of C2- and C3-acylcarnitines ($P < 0.047$) - incomplete oxidation products of fatty acid and amino acid metabolism. Lower *EARS2* and *COG7* expression was also related to the lower muscle C2- and C3-acylcarnitine concentrations ($P < 0.05$). Our results imply that non-completion is genetically moderated through alterations in gene expression and metabolic pathway in skeletal muscle. Impaired mitochondrial energetics and Golgi function in skeletal muscle may be partly responsible for non-completion. Individual genetic traits may allow development of a biomarker-approach to identify individuals that would benefit from more intensive counseling to maintain exercise programs.

54 | Genome-wide Association Study in Multiplex Consanguineous Pakistani Pedigrees with Schizophrenia and Bipolar Disorder

Jibin John[1,2]*, Amelie M. Johnson[2], Qin He[2], Mehtab Christian[2], Lynn E. Delisi[3], Ridha Joober[1], Marie-Pierre Dubé[4], Guy A. Rouleau[5], Lan Xiong[1,2,5]

[1]Department of Psychiatry and Douglas Mental Health University Institute, McGill University, Verdun, QC, Canada; [2]Centre de Recherche, Institut Universitaire en Santé Mentale de Montréal et Université de Montréal, Montréal, QC, Canada; [3]VA Boston Healthcare System, Department of Psychiatry, Harvard Medical School, Brockton, MA, United States of America; [4]Research Center of Montreal Heart Institute, Université de Montréal, Montreal, QC, Canada; [5]Department of Neurology and Neurosurgery and Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada.

Introduction: Genome wide association studies (GWASs) in additional diverse populations are essential to test the generalizability of the previous GWAS findings, to extend our understanding of the disease etiologies, and to identify the population-specific genetic risk in addition to common factors. Particularly no such studies have been performed in South Asian populations, one of the largest parts of the world population with distinctive population history and genomic characteristics.

Samples and Methods: Ten large consanguineous pedigrees (124 affected with SCZ or BPD and 151 unaffected individuals) and 34 unrelated healthy individuals were recruited from Sindh Province of Pakistan. Genotyping was performed using Illumina HumanOmniExpress BeadChip. After the standard QC and genotype phasing and imputation, we performed family-based GWAS using DFAM implemented in PLINK1.07 and FUMA for post-GWAS annotations.

Results: We have identified one locus (chr2:11, 237,350, rs541513649; *P value* $5.52e^{-08}$) associated with the broad psychiatric phenotype and 14 other independent loci showing suggestive statistical evidence of association (*P value* $< 10^{-5}$). Out of 15 loci, 10 also showed nominal association (*P* value $< 1.00e^{-3}$-$2.31e^{-11}$) in the previously reported psychiatry and/or related GWASs, e.g., a locus on chr7: 68,977,045–69,825,163, (rs12698811, P value $8.75e^{-07}$) was previously reported associated with cognition (rs12112638, *P value* $2.31e^{-11}$; Lee et al., 2018) and intelligence (rs12698891, P value $1.22e^{-09}$, Savage et al., 2018). Functional annotation of these loci using FUMA have also identified a number of interesting disease relevant genes, including *KCNF1, ROCK2, GABRA4, GABRB2, AUTS2, IGF1* etc.; which showed supporting evidence from the previous genetics and functional studies.

Invited Abstract

55 | Identifying Drug Targets Using Human Genetics at Scale

Toby Johnson*

Human Genetics, GlaxoSmithKline (GSK), Stevenage, SG1 2NY, United Kingdom.

Drug efficacy and safety are definitively tested in phase III trials, completed 10–15 years after "Commit to Target" (C2T) decisions. Notwithstanding high failure rates throughout the drug discovery, three-quarters of drugs with novel targets fail in phase III. This implies many *therapeutic hypotheses* (modulating target X to treat disease Y) proposed at C2T are wrong. Selecting target-disease pairs using human genetic studies can increase the probability of success in drug development. This is essentially a

Mendelian randomization (MR) argument, and is supported by retrospective analyses of target-disease pairs that succeeded or failed. With well powered Genome-Wide Association Studies (GWAS) for thousands of human diseases, traits, and -omic phenotypes, it nonetheless remains both challenging and necessary to infer the causal genes, in a robust and high throughput manner.

I demonstrate data and compute infrastructure, and key inferential tools, developed within GSK to systematically evaluate genetic support for every target-disease pair, and applied to C2T decisions. I focus on two areas, where I show that commonly used inferential approaches have high risk of mis-inference. For MR and Phenome-Wide Association Study (PheWAS) approaches, it is important to have good tools to evaluate the "genomic context" of the instrument(s). For GWAS-expression colocalization approaches, it is important to have tools and data to evaluate the extent of "molecular pleiotropy" across genes and tissues (and ideally cell types and conditions). Underlying the novel tools and visualizations are some conceptual advances, which could be broadly applied in "post-GWAS" research to generate more robust target-disease therapeutic hypotheses.

56 | eMERGE Phenome-wide Association Study of Biogeographic Ancestries Predicts Ocular, Immune System, Renal, Cardiometabolic, Gynecological, and Vector-borne Disease Risk

Jacob M. Keaton[1,2,3,4]*, Jacklyn N. Hellwege[2,5], Eric S. Torstenson[1,2], Ky'Era Actkins[2,6], Rachel Knevel[7,8,9], Lea Davis[2,5], Joshua C. Denny[2,4,11], Dan M. Roden[2,4,10,11], Todd L. Edwards[1,2,3], Digna R. Velez Edwards[2,3,4,12], on behalf of the eMERGE Network

[1]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America; [3]Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [5]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America; [6]Department of Microbiology, Immunology, and Physiology, Meharry Medical College, Nashville, Tennessee, United States of America; [7]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands; [8]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America; [9]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America; [10]Department of Pharmacology, Vanderbilt University, Nashville, Tennessee, United States of America; [11]Department of Medicine, Vanderbilt University, Nashville, Tennessee, United States of America; [12]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.

Racial disparities may arise in part due to phenomena that increase the frequency of trait-increasing alleles in one geographic parental subpopulation relative to another. In admixed offspring populations, those differences manifest as association between proportions of genetically inferred ancestry and traits. To evaluate this, we estimated six geographic genetic ancestry proportions based on 1000 Genomes reference populations in 60,267 non-Hispanic white (NHW) and 10,168 non-Hispanic black (NHB) participants from the electronic MEdical Records and GEnomics (eMERGE) Network and performed a phenome-wide association study (PheWAS) stratified by race. Associations of ancestry proportion, ancestry-body mass index (BMI) interactions, and ancestry-ancestry interactions with PheWAS outcomes were adjusted for age, sex, and BMI. Among NHW participants, Northern European (NEUR) and Southern European (SEUR) ancestries were associated with eye disease (e.g. presbyopia; $p_{NEUR} = 9.67 \times 10^{-276}$, $p_{SEUR} = 2.20 \times 10^{-207}$) and immune system disorders ($p_{NEUR} = 2.80 \times 10^{-172}$, $p_{SEUR} = 3.29 \times 10^{-153}$). Among NHB participants, East African (EAFR) and West African (WAFR) ancestries were associated with rosacea ($p_{EAFR} = 6.73 \times 10^{-6}$), stage III chronic kidney disease (CKD; $p_{WAFR} = 2.51 \times 10^{-6}$), diabetic kidney disease ($p_{WAFR} = 1.27 \times 10^{-5}$), and hypertensive kidney disease ($p_{WAFR} = 4.29 \times 10^{-6}$). Additionally, protective effects for WAFR ancestry were observed for pregnancy complications among NHB women (e.g. malposition and malpresentation of fetus or obstruction; $p_{WAFR} = 9.37 \times 10^{-8}$), which is consistent with a model of increasing admixture promoting reproductive success. Significant ancestry-BMI interactions were observed for psychiatric, metabolic, and immune diseases. Significant ancestry-ancestry interactions were observed for vector-borne, metabolic, and renal diseases. These results demonstrate the ability of geographic genetic origin to predict many types of disease risk. Replication analyses are underway in a large clinical biobank.

57 | Epigenetic Loci for Blood Pressure are Associated with Hypertensive Target Organ Damage in an Older African American Population

Minjung Kho[1]*, Wei Zhao[1], Scott M. Ratliff[1], Farah Ammous[1], Thomas H. Mosley[2], Sharon L.R. Kardia[1], Xiang Zhou[3], Jennifer A. Smith[1]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Memory Impairment and Neurodegenerative Dementia (MIND) Center, University of Mississippi Medical Center, Jackson, Mississippi, United States of America; [3]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America.

Hypertension is a major modifiable risk factor for arteriosclerosis that can lead to target organ damage (TOD) of heart, brain, kidneys, and peripheral arteries. A recent epigenome-wide association study for blood pressure (BP) identified 13 significant DNA methylation CpG sites, but it is not known whether these sites are also associated with TOD. In 1,218 African Americans from the Genetic Epidemiology Network of Arteriopathy study, a cohort of hypertensive sibships, we evaluated the associations between these 13 CpG sites measured in peripheral blood leukocytes and TOD traits assessed approximately 5 years later. Ten significant associations were found after adjustment for age, sex, blood cell counts, time difference between CpG and TOD measurement, and 10 genetic principal components (FDR $q < 0.1$): 2 with estimated glomerular filtration rate (eGFR), 6 with urinary albumin-to-creatinine ratio (UACR), and 2 with left ventricular mass/height^2.7 (LVM). All associations with eGFR and 4 associations with UACR remained significant even after further adjustment for BMI, smoking, and diabetes. We further tested the interaction of the significant CpG sites with risk factors for arteriosclerosis including BMI, smoking, and diabetes on TOD measures, and found 1 CpG-by-BMI and 3 CpG-by-diabetes interactions on UACR (FDR $q < 0.1$). Mediation analysis showed that 4.7% to 38.1% of the relationship between CpG sites (cg19693031 and cg00574958) and two TOD measures (UACR and LVM) was mediated by blood pressure (Bonferroni-corrected $P < 0.05$). This study may lend insight into the role of DNA methylation in pathological mechanisms underlying target organ damage from hypertension.

58 | aNSAIDS and Colorectal Cancer: Results from Genomewide Gene Environment Interaction Scans

Andre E. Kim[1]*, David A. Drew[2], John P. Morrison[1], Stephanie A. Bien[3], Victor Moreno[4], Graham Casey[5], Ulrike Peters[3], Andrew T. Chan[2], W. James Gauderman[1]

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [2]Clinical and Translational Epidemiology Unit, and Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America; [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [4]Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain; [5]Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, United States of America.

Genomewide Interaction Scans (GWIS) can provide insight into novel susceptibility loci and biologically meaningful interactions. Colorectal cancer (CRC) is well suited for GWIS because it is influenced by both genetics and several modifiable lifestyle and pharmacological risk factors, including nonsteroidal anti-inflammatory drugs (NSAIDs). Major limitations of GWIS are large sample size requirements and multiple testing burden. We address these limitations by pooling epidemiological and imputed genetic data from over 40 studies (48,258 cases and 54,534 controls), employing innovative statistical methods to boost power, and incorporating experimentally informed functional scores to prioritize biologically relevant loci.

Epidemiological data harmonization was conducted by mapping study specific questionnaire and data dictionary items to common data elements (CDEs) in close collaboration with contributing study leads. CDEs were combined into a single dataset with consistent definitions, permissible values, and coding. Genotype data were imputed to the Haplotype Reference Consortium reference panel using the Michigan Imputation Server and pooled across imputation batches. Data management and analyses were performed using a suite of custom R packages, *BinaryDosage* and *GxEScanR*, that convert imputed genotypes to a binary format and efficiently implement several GWIS methods.

In joint 2 degrees of freedom testing of marginal and interaction terms we identified loci not previously reported for CRC that may warrant further investigation, including SNPs in HLA-B, TTC22 regions. Our GWIS of NSAIDs identified novel genomic regions associated with CRC, and forthcoming incorporation of functional scores may lead to identification of additional regions of interest.

59 | Entanglement Mapping: A Model-free Approach to Detecting Interactions Among Predictive Features

Daniel Kiser[1,2]*, Jeremy Sabourin[1], Anthony M. Musolf[1], Joan E. Bailey-Wilson[1], James D. Malley[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2]Renown Institute for Health Innovation, Desert Research Institute, Reno, Nevada, United States of America.

In many genome-wide association studies (GWAS), the proportion of variation in disease risk or quantitative trait measurement that can be attributed to individual significantly-associated genetic variants has tended to be small. One possible reason is that most GWAS look at only single risk factors or independent features, and do not examine potential interactions among all such features. That is, individual features, such as genotypes and environmental measures, may be only weakly predictive in themselves, but strongly predictive when part of a subgroup of other features.

We provide a scheme for identifying such *entangled communities* of interacting features and demonstrate its performance on simulated data.

This scheme for locating entangled features requires only a single pass through the full (potentially large) list of features. The method removes or randomizes a feature and determines which of the remaining features are also down-weighted in importance (using any reasonable feature importance method) when compared to the analysis with all features. All the features in the identified subgroup are *jointly* necessary for good prediction, but individually may be only weakly predictive. For each down-weighted subgroup, this entire set of features must be included for good prediction. Simulations show the algorithm can detect: (1) higher order interactions of more than two features, (2) interactions that do not follow the multiplicative model, (3) multiple sets of interacting features, and (4) interactions in the presence of noise features and main effects.

This is a model-free, parameter-free scheme for detecting jointly predictive features and works with any learning machine.

60 | Four Novel Signals Suggest Possible Genetic Component to Age-of-onset of Idiopathic Pulmonary Fibrosis

Luke M. Kraven[1]*, Richard J. Allen[1], Adam Taylor[2], Martin D. Tobin[1,3], Ian Sayers[4,5], Richard B. Hubbard[5], Toby M. Maher[6,7], Astrid Yeo[2], Gisli R. Jenkins[4,5], Louise V. Wain[1,3]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]GlaxoSmithKline, Stevenage, United Kingdom; [3]NIHR, Leicester Respiratory Biomedical Research Centre, Leicester, United Kingdom; [4]NIHR, Nottingham University Hospitals, Nottingham, United Kingdom; [5]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [6]NIHR Respiratory Biomedical Research Unit, Royal Brompton Hospital, London, United Kingdom; [7]National Heart and Lung Institute, Imperial College, London, United Kingdom.

Idiopathic pulmonary fibrosis (IPF) is a progressive lung disease with poor prognosis. Identifying genetic determinants of disease can improve our understanding of underlying mechanisms and highlight important pathways for future drug targets. Previous genome-wide association studies (GWAS) have identified 17 genetic signals associated with IPF susceptibility but ours is the first to investigate the age-of-onset.

We performed a discovery GWAS in 465 subjects with IPF from the PROFILE study, assuming an additive genetic model and adjusting for sex, smoking history and the first 10 genetic principal components. Genetic variants which showed suggestive statistical significance ($P < 1 \times 10^{-5}$) in the discovery analysis were followed-up in two additional independent cohorts; the Trent Lung Fibrosis study (n = 210) and UK Biobank (n = 98). The results from the three cohorts were then meta-analysed. Genome-wide significance in the meta-analysis was defined as $P < 5 \times 10^{-8}$.

The genome-wide analysis was performed on 10,858,143 genetic variants. There were 14 independent genetic signals that showed suggestive statistical significance in discovery. No genetic variants reached genome-wide significance in the meta-analysis, but four variants did maintain suggestive statistical significance ($P < 1 \times 10^{-5}$). The most significantly associated variant was rs75681116, which is located on chromosome 8 between the genes *LZTS1* and *RNU3P2,* where each copy of the risk allele was associated with a 9-month earlier age-of-onset of IPF (95% CI [6 months, 12 months], $P = 2.2 \times 10^{-7}$).

We will further investigate the potential effects of index event bias and choice of statistical model on our findings and seek additional support for the four novel signals in additional independent data sets.

61 | Comparison of Imputation Quality for an Arab Population Using Different References, GWAS Panels, and Methods

Khalid Kunji*, Mohamad Saad

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.

There has been a significant gap between research on European genetics and that of other ancestries. The ExAC dataset is 54.97% European and the Telenti et. al 10,000 deeply sequenced genomes are 78.55% European. There are efforts underway to address this data bias, e.g. the work of the Consortium on Asthma among African-ancestry Populations in the Americas, the Personal Genome Project's 10,000 Korean genomes goal, etc. Despite these, there has been a relative dearth of data and analysis of Middle Eastern populations. Here we attempt to identify which methods, reference panels and Genome-Wide Association Study (GWAS) SNP arrays that perform well for a Qatari cohort.

We used a public set of 108 Qatari Whole Genome Sequences (WGSs). We formed three GWAS arrays by keeping only the SNPS present in two GWAS panels from Illumina: (1) The Multi-Ethnic AMR/AFR-8 Kit (AMR_AFR), and (2) Infinium Multi-Ethnic EUR/EAS/SAS-8 Kit (EUR_EAS_SAS) panels, and (3) we also considered selecting SNPs at random. Two imputation methods were performed on each GWAS data, Minimac and the recent

Positional Burrows-Wheeler Transform (PBWT) using two reference populations, 1KGP3 (1000 Genomes Phase 3) and Haplotype Research Consortium (HRC). We find that our best results are achieved using the 1KGP3 reference with the EUR_EAS_SAS GWAS panel and Minimac. Minimac combined with the 1KGP3 reference performed particularly well, and Minimac was consistently the better imputation method. The EUR_EAS_SAS panel did slightly better than the AMR_AFR panel, but the random SNP selection performed worse than either panel. Rare SNPs were imputed markedly well.

62 | Leveraging External Repositories to Generate Calibrated Rare Variant Gene Risk Scores

Ricky Lali[1]*, Michael Chong[1,2], Arghavan Omidi[1], Pedrum Mohammadi-Shemirani-[1,3], Ann Le[1,3], and Guillaume Paré[1,2,3,4,5,6]

[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, ON, Canada; [2]Department of Biochemistry and Biomedical Sciences, McMaster University, Faculty of Health Sciences, Hamilton, Canada; [3]Department of Medical Sciences, McMaster University, Faculty of Health Sciences, Hamilton ON, Canada; [4]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, ON, Canada; [5]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton ON, Canada; [6]Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton ON, Canada.

Rare variants (minor allele frequency < 0.001) individually confer modest effects but are collectively numerous and may account for a considerable proportion of complex disease risk. Identifying rare variant associations is challenging due to the need for large sample sizes, technical artefacts, and population structure. We propose a novel method that leverages summary level data from a large public exome sequencing database (gnomAD) as controls and calibrates rare variant burden at the individual and gene levels to circumvent biases in population substructure and mutation burden, respectively. The method was applied to a European coronary artery disease (CAD) cohort (N = 5921) where the *bonafide* CAD gene, low-density lipoprotein receptor, reached exome-wide significance (OR = 2.34; 95% CI, 1.76–3.18; $P = 1.90 \times 10^{-13}$). A rare variant genetic risk score was generated using the top 1040 discovery genes (RVGRS1040) and found to associate with CAD in: 1) UK Biobank Europeans (N = 45850) (OR = 1.08 per SD; 95% CI, 1.04–1.12; $P = 5.50 \times 10^{-5}$) and 2) Pakistan Risk of Myocardial Infarction South Asians (N = 6655) (OR = 1.07; 95% CI, 1.02–1.12; $P = 0.009$). Furthermore, RVGRS1040 was independent of both common variant genetic risk score and clinical risk factors (OR = 1.07; 95% CI, 1.03–1.10; $P = 0.0002$), and significantly improved classification of CAD events (Net Reclassification Index = 0.0554; $P = 0.001$). Our method improved CAD risk prediction by accounting for the aggregate effect of rare variants through a polygenic model of inheritance, which is robust to population genetic effects.

63 | An Extended Expression Prediction Approach for Twas Leveraging the cis-mediator *trans*-eQTL Paradigm

Nicholas B. Larson[1]*, Shannon K. McDonnell[1], Zachary Fogarty[1], Stephen Thibodeau[2]

[1]Department of Health Sciences Research, Mayo Clinic College of Medicine and Science, Rochester, Minnesota, United States of America; [2]Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine and Science, Rochester, Minnesota, United States of America.

With the increasing availability of large tissue-specific gene expression datasets, transcriptome-wide association studies (TWAS) have become an effective integrative strategy for disease gene discovery. One limitation of existing TWAS methods is restriction to cis-acting variation, as robust trans-eQTL identification requires very large sample sizes. Another limitation is that weak gene dysregulation signals may mediate risk in aggregate via downstream known and/or latent molecular network effects on key driver disease genes, which may go undetected. These initial genes are known as cis-mediators, and growing evidence indicates trans-eQTLs are heavily enriched for cis-eQTLs of other genes. Here, we propose to extend the standard cis-focused TWAS strategy for expression prediction to accommodate trans-eQTLs under the cis-mediator paradigm via a simple two-step approach. We first train standard cis-variation expression prediction models using elastic net. In the second step, we aggregate predicted expression values transcriptome-wide as features in a LASSO model for every gene, treating the original expression predictions as an offset. To illustrate improvement in expression prediction as well as inferred transcriptome regulatory connectivity, we apply our approach to a large prostate tissue eQTL dataset of N = 471 samples with available RNA-Seq and imputed genome-wide genotypes.

Among 6,676 protein-coding genes with significant cis-heritability, we observed a median 98 genes (IQR = [48,177]) selected as additional predictors in our second LASSO step, with a median model R2 increase of .40

(IQR = [.24,.56]) over cis-only models. Finally, we outline a TWAS strategy leveraging these extended expression prediction models and apply it to GWAS summary statistics for prostate cancer risk.

64 | A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-onset Alzheimer's Disease Using Whole Genome Sequence Data

Linhai Zhao[1], Zongxiao He[1], Di Zhang[1], Gao T. Wang[2], Alan E. Renton[3], Badri N. Vardarajan[4], Michael Nothnagel[5,6], Alison M. Goate[3,7], Richard Mayeux[4], Suzanne M. Leal[1,4,8]*

[1]Center for Statistical Genetics, Baylor College of Medicine, Houston, TX, United States of America; [2]Department of Human Genetics, The University of Chicago, Chicago, IL United States of America; [3]Department of Neuroscience and Ronald M. Loeb Center for Alzheimer's Disease, Icahn School of Medicine at Mount Sinai, New York, NY, United States of America; [4]Department of Neurology, Taub Institute on Alzheimer's Disease and the Aging Brain, and Gertrude H. Sergievsky Center, Columbia University, New York, NY, United States of America; [5]Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics University of Cologne, 50931 Cologne, Germany; [6]University Hospital Cologne, 50937 Cologne, Germany; [7]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY, United States of America; [8]Center for Statistical Genetics, Columbia University, New York, NY, United States of America.

To analyze family-based whole genome sequence (WGS) data for complex traits, we developed a rare variant (RV) non-parametric linkage (NPL) analysis method, which has advantages to RV aggregate association methods. The RV-NPL differs from the NPL in that RVs are analyzed in aggregate and allele sharing amongst affected relative-pairs is only estimated for minor alleles. Analyzing families can increase power because causal variants with familial aggregation usually have larger effect sizes than those underlying sporadic diseases. Differing from association analysis, for NPL only affected individuals are analyzed, which can increase power, since unaffected family members can be susceptibility variant carriers. Unlike RV aggregate association methods, RV-NPL is robust to population substructure and admixture, inclusion of nonpathogenic variants, allelic and locus heterogeneity, and can readily be applied outside of coding regions. In contrast to analyzing common variants using NPL, where loci localize to large genomic regions e.g. >50 Mb, mapped regions are well defined for RV-NPL. Using simulation studies, we demonstrate that RV-NPL is substantially more powerful than applying traditional NPL methods to analyze RVs. The RV-NPL was applied to analyze 107 Late-onset Alzheimer's disease (LOAD) pedigrees of Caribbean Hispanic and European ancestry with WGS data and statistically significant linkage (LOD ≥ 3.8) was found with RVs in *PSMF1* and *PTPN21,* that were previously shown to be involved in LOAD etiology. Additionally, nominally significant linkage was observed with RVs in LOAD associated genes *ABCA7*, *ACE, EPHA1*, and *SORL1*. RV-NPL is an ideal method to elucidate the genetic etiology of complex familial diseases.

65 | Can Identity-by-descent Sharing Information Complement Population Based Imputation Algorithms?

Anthony F. Herzig[1], Marina Ciullo[2,3], and Anne-Louise Leutenegger[4]*

[1]Université de Brest, GGB, Inserm, Brest, France; [2]Institute of Genetics and Biophysics A. Buzzati-Traverso,- Naples, Italy; [3]Istituto di Ricovero e Cura a Carattere Scientifico, Neuromed, Pozzilli, Isernia, Italy; [4]Université de Paris, NeuroDiderot, Inserm, Paris, France.

In the context of population-based imputation, the intuitive benefits of including study specific haplotypes in one's reference panel have now been widely demonstrated. In some cases, such as in the study of isolated populations, it may be possible to include reference panel individuals who are closely related to target individuals; thus giving exceptionally high imputation accuracy.

Such findings suggest the potential importance of long matches between target and reference haplotypes, in other words regions that are likely to be shared identical-by-descent (IBD), whose detection is typically the basis of family-based imputation methods. However, in most circumstances, IBD-sharing information will typically not be sufficient to impute missing genotypes across the whole genome as identifiable shared regions will not cover entire chromosomes.

This has led to the idea of combining IBD and population-based methods. Whilst a full integration of the two has yet to be successfully demonstrated, two recent software have been put forward that provide two-step approaches: Ped-Pop and Kinpute. These methods seek to significantly improve upon the exactitude of imputed genotypes from population-based methods by overlaying inference from IBD-sharing information. Here, we review the software and perform tests on simulation data based on the structure of the known genetic isolates of Cilento in Southern Italy. This allows us to determine and provide examples of specific scenarios where such methods may be most or least successful.

66 | QC Measurements of Exome Chip Sequence Data in a Family-based Study

Qing Li[1], Stephen Wank[2], 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]Digestive Diseases Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, United States of America.

Quality control (QC) is an important step in sequence data analysis. In VCF files, many genotype calling quality measurements are reported at the variant level and the variant-and-sample level. Common practice is to drop variants based on pre-set thresholds of several key QC measurements, including QD, MQ, FS, SOR MQRankSum, and ReadPosRankSum at the variant level. In addition, genotypes at specific loci should be set to missing if the variant-and-sample level measurements are poor. Pedigree information can be used as the third step of data cleaning, eliminating genotypes causing Mendelian inconsistencies.

In a linkage analysis of small intestinal carcinoid tumors, we performed QC on whole exome sequence data from a pedigree of 34 individuals. In this work, we reported the summary statistics on the genotype calling QC measurements of 219,742 variants. We found that the QC measurements have a wide range of variation across different chromosomes. Using hard-filtering based on recommended thresholds, 182,132 (~83%) variants were kept. We found that the current QC thresholds cannot remove all the poorly typed genotypes. Among the kept variants, 21% variants incurred at least one Mendelian error. Based on the sequenced individuals, we can infer 18 child-parent(s) trios, ~1.9% with Mendelian errors (the total number of Mendelian errors divided by the product of number of variant (N=28,114) and 18 child-parent(s) trios). Extensive and iterative cleaning of Mendelian errors are needed after data filtering by QC measures alone. In conclusion, the current hard filtering thresholds are inadequate and are improved by Mendelian inconsistency checks.

67 | FamRVC Program for Family-based Rare Variant Association Tests for Censored Traits and its Applications to Age-at-onset of Alzheimer's Disease

Yi-Ju Li[1,2]*, Wenjing Qi[1,2], Michael A. Schmidt[3], Xuejun Qin[2], Susan H. Slifer[3], Andrew S. Allen[1,4], X. Raymond Gao[5], Eden R. Martin[3]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States of America; [2]Duke Molecular Physiology Institute, Duke University, Durham, NC, United States of America; [3]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States of America; [4]Center for Statistical Genetics and Genomics, Duke University, Durham, NC, United States of America[5]Department of Ophthalmology and Visual Science, Department of Biomedical Informatics, Division of Human Genetics, The Ohio State University, Columbus, OH, United States of America.

Analysis of genomic sequence data in families has an advantage over unrelated samples for rare-variant (RV) association due to higher chance of sampling causal RVs in a family. While many RV association tests have been developed, few have been designed for time-to-event outcomes, referred to as censored traits. We recently proposed a set of gene-based RV tests for censored traits in families, the FamBAC and FamKAC tests (Qi et al. 2019). The tests are based on the score statistics derived from the frailty model. Here, we implemented these two tests in a Family-based Rare Variant association tests for Censored traits (FamRVC) program written in R. FamRVC can handle nuclear families of multiple siblings, with or without parental information. Input files are similar to PLINK with some additions, including a.map file for SNP list,.raw or.dosage for genotypes,.fam for family specification,.maf for SNP population minor-allele frequency,.pheno for phenotype with censorship index,.cov for covariates, and a gene list file. We applied FamRVC to imputed SNP data in 505 families of Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE). We considered age-at-onset (AAO) of Alzhiemer's disease (AD) as a censored trait with AAO censored at age-at-exam for unaffecteds. Rare variants (MAF < 0.02) of GSTO1 and GSTO2 in chromosome 10 (Li et al. 2003) and APOE gene were analyzed. GSTO1 was confirmed (FamBAC p = 0.027) but not rare variants in APOE, implying that the well-known APOE effect on AD AAO mainly from two known common SNPs (*rs429358* and *rs7412)*.

68 | Fast and Powerful Method for eQTL and Fine-mapping Integrating Total and Allele-specific Expression

Yanyu Liang[1]*, Stephane E. Castel[2,3], GTEx Consortium, Tuuli Lappalainen[2,3], Hae Kyung Im[1]

[1]Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [2]Department of Systems Biology, Columbia University, New York, New York, United States of America; [3]New York Genome Center, New York, New York, United States of America.

Increasing number of studies measure RNA-seq with allele-specific (AS) reads. Although there have been

efforts to unify total and AS reads, they tend to be computationally intensive and prohibitive for GTEx-scale studies. In addition to QTL mapping, multi-site models for prediction and fine-mapping have become increasingly important for post-GWAS analyses such as PrediXcan and COLOC. To our knowledge, there is no multi-site method taking advantage of AS observations.

Here we propose a linear mixed effect model combining total counts and gene-level AS read counts obtained by phASER. We implement the single-site version of the model for QTL-mapping. To fit the multi-site model, we develop algorithm to train prediction model using all variants inside cis-window. To make use of existing fine-mapping tools, which assume independent observations, we perform a whitening procedure on the three observations per individual. Finally, we apply the methods to GTEx RNA-seq data to study cis-regulation of gene expression.

We applied our methods to GTEx data. Our eQTL mapping method (mixQTL) was over 10 times faster than existing methods such as WASP and RASQUAL. Compared to GTEx standard eQTL calling pipeline, mixQTL was more powerful while maintaining type I error calibration. Our multi-site method showed better performance in identifying causal variants both in simulations and real data.

We present here a suite of methods that are both fast and powerful for cis-eQTL mapping, fine-mapping, and prediction by taking full advantage of total and AS counts available in many studies.

69 | Independent Replication of Genetic Associations with Urinary Bladder Cancer Prognosis in the UK Biobank Using Hospital Record Data

Nadezda Lipunova[1,2,5]*, Anke Wesselius[2], Kar K. Cheng[3], Frederik-Jan van Schooten[4], Richard T. Bryan[1], Jean-Baptiste Cazier[1,5], Maurice P. Zeegers[1,2]

[1]Institute of Cancer and Genomic Sciences, University of Birmingham, United Kingdom; [2]Department of Complex Genetics, Maastricht University, The Netherlands; [3]Institute for Applied Health Research, University of Birmingham, United Kingdom; [4]Department of Pharmacology and Toxicology, Maastricht University, The Netherlands; [5]Centre for Computational Biology, University of Birmingham, United Kingdom.

Urinary bladder cancer (UBC) is a disease with great burden on healthcare systems and patients. Advances in genetic research has great potential to aid clinical management of already diagnosed cases; however, UBC prognosis is a difficult phenotype to study due to lack of routinely collected data on disease recurrence and progression. As such, independent validation of genetic associations with UBC are rarely carried out and prevent timely evidence transfer from hypothesis to action.

We have reviewed existing reports on SNP associations with UBC prognostic outcomes of recurrence, progression, and survival. A Principal Component Analysis (PCA) was carried out to investigate similarity between outcomes in the genetic context. To replicate existing associations, we have used UK Biobank data on Health Episode Statistics (HES) to create variables representative of UBC outcomes. We are using SNPtest for replication analyses, all associations currently being tested under additive model of inheritance.

There was considerable heterogeneity between genetic architectures of each studied outcome, recurrence being the most distinct phenotype. UK Biobank HES data showed to be a good data source to model prognostic variables that are otherwise not collected routinely. As such, external replication analyses are currently in development and will be presented at the IGES 2019 Meeting.

Our study indicates UBC prognosis is a highly complex outcome that should be studied with more narrowly defined endpoints. Moreover, we also show routinely collected HES data is a useful source for studying UBC prognosis and should be encouraged to use in both exploratory and confirmative research.

70 | Exome-wide Low-frequency Genetic Variants Contribute to Human Craniofacial Morphology

Dongjing Liu[1]*, Nora Alhazmi[2], Jacqueline T. Hecht[3], George L. Wehby[4], Lina M. Moreno[5], Carrie L. Heike[6], Peter Claes[7], Eric C. Liao[8], Seth M. Weinberg[9], John R. Shaffer[1]

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [2]Department of Oral Biology, Harvard School of Dental Medicine, Boston, Massachusetts, United States of America; [3]Department of Pediatrics, University of Texas McGovern Medical Center, Houston, Texas, United States of America; [4]Department of Health Management and Policy, University of Iowa, Iowa City, Iowa, United States of America; [5]Department of Orthodontics, University of Iowa, Iowa City, Iowa, United States of America; [6]Department of Pediatrics, Seattle Children's Craniofacial Center, University of Washington, Seattle, Washington, United States of America; [7]Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium; [8]Department of Surgery, Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America; [9]Department of Oral Biology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America.

The genetic basis of normal-range variation in human facial traits is still poorly understood. Most studies to

date have approached facial morphology via univariate measurements with genetic hypotheses centered on common genetic variants. Studies on other complex traits have identified fruitful associations with rare and low-frequency variants involved. To better understand the genomic architecture of facial morphology, we studied the influence of low-frequency coding variants on multi-dimensional facial shape phenotypes. A cohort of 2,329 European individuals were genotyped for approximately 245,000 coding variants on the Illumina Exome v1.2 array. Using three-dimensional facial images, we partitioned the full face into 31 hierarchically arranged modules to model global-to-local features, and generated multi-dimensional phenotypes representing the shape variation within each module. We used multivariate kernel regression (implemented in the MultiSKAT R package) to test the association between the multivariate facial phenotypes and exome-wide variants with frequencies <1% in a gene-based manner. After accounting for multiple tests, we identified eight significant genes (*AR*, *CARS2*, *FTSJ1*, *HFE*, *LOC108783645*, *LTB4R*, *TELO2*, *NECTIN1*) as influencing the morphology of the cheek, chin, nose and philtrum region. These genes displayed a wide range of phenotypic effects, with some impacting the full face and others affecting only localized regions. Notably, *NECTIN1* is a well-established craniofacial gene that underlies the etiology of a syndromic form of cleft lip and palate (MIM#225060). These results have expanded our understanding of the genetic basis of normal-range human facial morphology by implicating rare and low frequency coding variants in novel candidate genes.

71 | Testing Gene-environment Interactions Without Measuring the Environment

Jiacheng Miao[1], Yuchang Wu[1,2], Kunling Huang[1], Zheng Ni[1], Qiongshi Lu[1,2]*

[1]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America.

Both genetic variations and environmental factors play key roles in shaping the etiology of complex disease. Dissecting the interplay between genetics and environmental factors may provide insights into disease development and shed light on treatment strategies. However, studies focusing on gene-environment (GxE) interactions only had limited success, in part due to a lack of large datasets with both genetic data and robust environmental measures. Here, we seek a solution for a challenging problem: can we test GxE interactions without measuring the "E"? We propose a novel statistical framework to test GxE interactions by using polygenic scores (PGS) as proxies for the environmental factors. PGS has gained popularity in GxE research and many studies used PGS as the "G" component. Through theoretical and numerical analyses, we demonstrate that the inference of interaction remains valid after replacing the "E" component with its PGS. We applied our method to three large, independent genetic datasets for autism spectrum disorder (ASD; n = 7,805 probands) to investigate the interaction of genetics and birth weight on ASD risk. We used a PGS derived from the UK biobank (n = 205,475) as the genetic proxy for birth weight. Meta-analysis revealed a significant negative interaction between genetics and birth weight ($P = 4.6e\text{-}4$), suggesting that a higher birth weight may buffer the genetic risk of ASD. As a negative control, we also applied our approach to 3,243 healthy siblings of ASD probands and did not identify any interaction ($P = 0.65$). We believe this method has great potential for advancing our understanding of complex disease.

72 | Whole-genome Bisulfite Sequencing in Systemic Sclerosis Provides Novel Targets to Understand Disease Pathogenesis

Tianyuan Lu[1,2]*, Kathleen Oros Klein[1], Inés Colmegna[3], Maximilien Lora[3], Celia M. T. Greenwood[1,4,5,6], Marie Hudson[1,3]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Quantitative Life Sciences Program, McGill University, Montreal, Canada; [3]Division of Rheumatology, Department of Medicine, McGill University, Montreal, Canada; [4]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada; [5]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada; [6]Department of Human Genetics, McGill University, Montreal, Canada.

Background: Systemic sclerosis (SSc) is a rare autoimmune connective tissue disease whose pathogenesis remains incompletely understood. Increasing evidence suggests that both genetic susceptibilities and changes in DNA methylation influence pivotal biological pathways and contribute to the disease. The role of DNA methylation in SSc has not been fully elucidated, because existing investigations of DNA methylation predominantly focused on nucleotide CpGs within restricted genic regions, and were performed on samples containing mixed cell types.

Methods: We performed whole-genome bisulfite sequencing on purified CD4+ T lymphocytes from nine SSc patients and nine controls in a pilot study, and then profiled genome-wide cytosine methylation as well as genetic variations. We adopted robust statistical methods to identify differentially methylated genomic regions (DMRs). We then examined pathway enrichment associated with genes located in these DMRs. We also tested whether changes in CpG methylation were associated with adjacent genetic variation.

Results: We profiled DNA methylation at more than three million CpG dinucleotides genome-wide. We identified 599 DMRs associated with 340 genes, among which 54 genes exhibited further associations with adjacent genetic variation. We also found these genes were associated with pathways and functions that are known to be abnormal in SSc, including Wnt/b-catenin signaling pathway, skin lesion formation and progression, and angiogenesis.

Conclusion: The CD4+ T cell DNA cytosine methylation landscape in SSc involves crucial genes in disease pathogenesis. Some of the methylation patterns are also associated with genetic variation. These findings provide essential foundations for future studies of epigenetic regulation and genome-epigenome interaction in SSc.

73 | Comprehensive Analysis of Pulmonary Surfactant Metabolism Genes and Gene Expression Patterns Associated with Lung Cancer Risk

Jennifer Luyapan[1,2]*, Xiangjun Xiao[3], Younghun Han[3], The INTEGRAL Consortium[4], Todd A. MacKenzie[1,2], Christopher I. Amos[1,2,3,4]

[1]Quantitative Biomedical Science, Geisel School of Medicine at Dartmouth College, United States of America; [2]Department of Biomedical Data Science at Dartmouth College, United States of America; [3]Dan L. Duncan Comprehensive Cancer Center at Baylor College of Medicine, United States of America; [4]The Integrative analysis of lung cancer etiology and risk consortium.

Surfactant metabolism genes play a key role in maintaining the integrity of pulmonary airways by encoding for proteins involved in the function of pulmonary surfactant. Mutations in genes responsible for surfactant homeostasis have been associated with lung cancer development. We are analyzing genetic variants in the surfactant metabolism pathway to identify its effects on lung cancer risk. To further understand the relevance of surfactant genes in lung cancer development, we predicted gene expression levels by performing a transcriptome-wide association study (TWAS). We used gene expression panels measured in healthy lung, whole blood and lung tumors. Genome-wide association summary statistics were computed using a fixed-effect meta-analysis for genome-wide genotype data from 29,266 cases and 56,450 controls of European Ancestry imputed from the OncoArray, containing association statistics for 7,673,198 variants. We identified one SNP in Surfactant Protein B (*SFTPB*), three SNPs in Surfactant Associated Protein 2 (*SFTA2*), four SNPs in Cathepsin H (*CTSH*), and 22 SNPs in Telomerase reverse transcriptase gene (*TERT*) achieving Bonferroni corrected *P* values for multiple testing. These SNPs showed diversity in associations across lung cancer histological subtypes and according to differences in environmental exposures, with *SFTA2* variants associated with squamous cell carcinoma, small cell lung cancer, overall lung cancer, and smoking exposures; the *SFTPB* variant is associated with lung adenocarcinoma. TWAS associations that did not overlap a genome-wide significant variant (i.e., +/− 100 kilobase target gene region) include significant associations of *CTSH*. Further analyses will be conducted to identify additional variants in surfactant metabolism proteins.

74 | Network-based Identification of Key Master Regulators for Immunologic Constant of Rejection in Cancer

Raghvendra Mall[1]*, Mohamad Saad[1], Jessica Roelands[2], Wouter Hendrickx[2], Michele Ceccarelli[3], Davide Bedgonetti[2]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [2]Department of Immunology, Inflammation and Metabolism, Division of Translational Medicine, Research Branch, Sidra Medicine, Doha, Qatar; [3]Computational Biology, Computational Oncology and Immunology (CIAO), AbbVie Biotherapeutics Inc., Redwood City, California, United States of America.

Molecular alterations governing mechanisms leading to immune exclusion are largely unknown. Availability of large-scale biomedical data offers an opportunity to assess the effect of key cellular determinants for an observed phenotype. We develop a network-based approach to identify key transcription factors (TF) associated with poor immunologic responsiveness. A cancer phenotype displaying the coordinated expression of T-helper-1 (Th1) chemokine, interferon, and immune-effector function genes, is associated with favorable prognosis and responsiveness to immunotherapy. This disposition is summarized by a signature that we termed as Immunologic Constant of Rejection (ICR). Based on expression of ICR genes, cancers are classified as immune active (ICR4), or immune silent (ICR1).

We use The Cancer Genome Atlas (TCGA) RNA-seq data for 12 cancer types (~2,500 samples) to: (1) build gene regulatory networks via Regularized Gradient

Boosting Machines (RGBM); (2) determine each TF's regulon, which are sets of genes regulated by the TF; (3) determine activity matrix of TFs for all samples; and (4) run functional gene set enrichment analysis to identify the top TFs, named Master Regulators (MR), that discriminate ICR1 vs ICR4. MRs such as *L3MBTL1*, *HDAC11*, and *SALL2* were coherently associated with the immune-silent phenotype (ICR1) across 12 cancers. Downstream analysis of MRs specific to ICR1 resulted in identification of NOTCH signaling pathways, chromatin regulation, transcriptional regulation of oncogene TP53, and several cancer-related signaling pathways that can represent novel targets to reprogram the immune suppressive tumor microenvironment. In summary, this is the first report that identified MRs associated with an immune-excluded cancer phenotype.

75 | Analysis of Whole Exome Sequencing Data of Hereditary Lung Cancer Families Identifies Germline Copy Number Variations (Cnvs) in Multiple Genes

Diptasri Mandal[1]*, Dinh-Van Tran[1], Kirsten Termine[1], Anthony M. Musolf[2], Mariza de Andrade[3], Ann G. Schwartz[4], Susan M. Pinney[5], Christopher I. Amos[6], Ramaswamy Govindan[7], Joan E. Bailey-Wilson[2], for the Genetic Epidemiology of Lung Cancer Consortium (GELCC)

[1]Department of Genetics, LSU Health Sciences Center, New Orleans, LA, United States of America; [2]National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, United States of America; [3]Mayo Clinic, Rochester, MN, United States of America; [4]Karmanos Cancer Institute, Wayne State University, Detroit, MI, United States of America; [5]University of Cincinnati College of Medicine, Cincinnati, OH, United States of America; [6]Baylor College of Medicine, Houston, TX, United States of America; [7]Division of Oncology, Washington University School of Medicine, St. Louis, Missouri, United States of America.

About 10% of lung cancer (LC) cases (22,000 cases per year) in the U.S. have at least one first-degree relative affected with LC, and about 1% of cases have at least three first- or second-degree affected relatives. To identify susceptibility gene(s) for hereditary lung cancer (HLC) families (≥3 LC/family) collected by the Genetic Epidemiology of Lung Cancer Consortium, we conducted whole exome sequencing (WES) on eight highly aggregated families (≥4 LC cases/family). These families included samples from 65 individuals, 20 of whom were affected and 45 of whom were unaffected at the time of data collection. We used two CNV-specific algorithms as incorporated in CANOES (http://www.columbia.edu/~ys2411/canoes/) and XHMM (https://atgu.mgh.harvard.edu/xhmm/) to analyze these data in identifying germline copy number variations (CNVs). To confirm, CNVs were then visualized independently using Integrative Genomics Viewer (IGV).

We have identified CNVs in more than 25 genes that are deleted or duplicated in two or more individuals in two or more families. Furthermore, both CNV-specific algorithms identified four cancer related genes: GSTM1, RHD, CFHR3, and CFHR1 in more than five families in two or more affected individuals/family. Reports from other studies show CNVs in those genes in somatic LC samples. Previously, no germline specific CNVs have been reported in those genes in HLC families. This observation calls for more in-depth analysis to take place on the next generation sequencing data that might be useful for the development of HLC-specific biomarkers in the future.

76 | Efficient Estimation of Hidden Ancestry Structure Using Summary Genotype Frequency Data

Alexandria Ronco[1]*, Ian S. Arriaga-Mackenzie[1]*, Gregory M. Matesi[1]*, Ryan Scherenberg[2], Yinfei Wu[1], James Vance[1], Jordan R. Hall[1], Christopher R. Gignoux[3], Megan Null[1], Audrey E. Hendricks[1,3]

[1]Mathematical and Statistical Sciences, University of Colorado, Denver, Colorado, United States of America; [2]Business School, University of Colorado, Denver, Colorado, United States of America; [3]Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora CO, United States of America.

Genetic research is being transformed by large publicly available genotype frequency databases such as the genome aggregation database (gnomAD). This summary level data is used to prioritize possible causal variants in the study of rare diseases and, more recently, as controls in association studies. Some data provided has heterogeneous ancestry such as the African group in gnomAD, which contains African-Americans. Lack of precise ancestry information can lead to confounded associations and incorrect prioritization of putative causal variants.

We have developed a generative method that estimates the proportions of reference ancestries from genome-wide publicly available summary level data. We use sequential quadratic programming, an iterative minimization algorithm, to estimate the mixing proportions within seconds. The speed of our algorithm allows for error estimates using block bootstrapping, which maintains the linkage disequilibrium structure within the genome.

We use a reference panel that includes 1000Genomes superpopulations (non-Finnish European, South Asian, East Asian, and African) and Native American ancestry. Our method is easily generalizable to other reference data

and populations. We evaluate our method in hundreds of simulation scenarios achieving estimates of ancestry proportions within 0.1% accuracy. Finally, we apply our method to the African, Latino, and Other groups in gnomAD to identify the proportion of ancestries within these heterogeneous groups. We find ~85% African ancestry within the African gnomAD group, which is consistent with African-Americans within this group. Our method provides accurate ancestry proportion estimates for publicly available genotype frequency data enabling better use of these and other summary level datasets.

77 | Genetic Correlations and Exploration of Uterine Fibroid Clinical Phenome in Black and White Women

Brian S. Mautz[1,2]*, Sarah H. Jones[3], Eric S. Torstenson[1,2], Jacklyn N. Hellwege[2,4], Todd L. Edwards[1,2,3], Digna R. Velez Edwards[2,3,5,6]

[1]Division of Epidemiology; [2]Vanderbilt Genetics Institute; [3]Institute for Medicine and Public Health; [4]Division of Genetic Medicine; [5]Division of Quantitative Sciences, Department of Obstetrics and Gynecology; [6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.

Uterine fibroids affect up to 70% of women by menopause, disproportionately impacting black females. Prior studies have investigated a limited set of clinical factors associated with fibroid risk. Moreover, the genetic relationship underpinning fibroids and other diagnoses are unexamined. Electronic health records (EHR) and publicly available genetic datasets provide a unique opportunity for a comprehensive, agnostic investigation of the fibroid phenome and genetic correlations. Utilizing the Vanderbilt University Medical Center EHR database, cases and controls in black (N = 3,568 cases; 12,521 controls) and white women (N = 7,577; 60,296) were identified using a validated method. First, we conducted a "phenome-wide association study" (PheWAS) to test for associations between fibroids and all diagnoses across EHRs in black and white women. Second, in initial analyses, we identified publicly available GWASs for top PheWAS hits and known fibroid risk factors. Using LDScore Regression and our own fibroid GWAS summary statistics, we estimated the genetic correlation between traits. Across racial groups the most significant PheWAS associations were previously identified fibroid symptoms (e.g. excessive menstruation, $P < 1.0 \times 10^{-274}$; dysmenorrhea, $P < 2.58$ X 10^{-152}). We also detected numerous novel associations with neoplasms, endometriosis, and diverticulosis. We didn't detect any significant fibroid-trait genetic correlations in either race, though there was evidence of a large, non-significant negative relationship of fibroids with type 2 diabetes in blacks ($r_g = -1.21$, $P = 0.09$). These results provide novel insight into the fibroid phenome. More thorough analyses of genetic correlations are underway.

78 | A Bayesian Model to Estimate Microbiome Network Changes with Respect to a Covariate Profile

Kevin D.J. McGregor[1,3]*, Aurélie Labbe[2], Celia M. T. Greenwood[1,3,4]

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Canada; [2]Département de sciences de la decision, HEC Montréal, Montréal, Canada; [3]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Canada; [4]Gerald Bronfman Department of Oncology, McGill University, Montréal, Canada.

The human microbiota is the collection of microorganisms colonizing the human body and plays an integral part in human health. A growing trend in microbiome analysis is to construct a network to estimate the co-occurrence patterns among taxa though correlation matrices, as these patterns are known to reflect metabolic interaction and competition for resources between taxa. Though methods have been developed to estimate differences in microbiome networks between two groups, there is currently no way to explore how networks associate with multiple covariates, given the compositional nature of microbiome data.

We propose a new model to estimate network changes with respect to a covariate profile. The counts of individual taxa in the samples are modelled through a multinomial distribution whose probabilities depend on a latent Gaussian term. The covariance matrix of the Gaussian term determines the taxa co-occurrence network and is parameterized to depend on an individual's covariate profile while preserving positive semi-definiteness. We propose a Gibbs sampler, similar to a model from Silverman et al. (2019), to obtain model parameter and interval estimates. We perform a simulation study to assess whether covariance patterns are correctly estimated over the covariate profiles of individuals. Additionally, we run an application of the model on intestinal microbiome 16 S sequencing data from individuals under 18 years with Crohn's disease. We find that our model outperforms a naive Gaussian-based model in the simulation study. We also outline important family level network changes with respect to age in the Crohn's dataset.

79 | Nearest-neighbor Projected-distance Regression to Detect Network Interactions and Control for Confounders, Population Structure and Multiple Testing

Trang T. Le[1], Bryan A. Dawkins[2], Marziyeh Arabnejad[3], Brett A. McKinney[2,3]*

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; [2]Department of Mathematics, University of Tulsa, Tulsa, Oklahoma, United States of America; [3]Tandy School of Computer Science, University of Tulsa, Tulsa, Oklahoma, United States of America.

Efficient machine learning feature selection is needed to detect complex interaction network effects in complicated modeling scenarios in high dimensional data, such as GWAS or gene expression for case-control or quantitative traits. A challenge for many machine learning feature selection methods is to detect interactions while also computing statistical significance of features and controlling for potential confounders from demographic data or population structure.

To address these challenges, we propose a new feature selection technique called Nearest-neighbor Projected-Distance Regression (NPDR) that uses the generalized linear model to perform regression between nearest-neighbor pair distances projected onto predictor dimensions. NPDR detects interaction structure using local nearest-neighbor information in the full space of predictors, which may be SNPs or expression levels. The method handles both case-control and quantitative traits, and the regression formalism allows for adjustment for multiple testing, correction for covariates and regularization.

Using realistic simulations with main effects and network interactions, we show that NPDR outperforms standard Relief-based methods and random forest at detecting functional variables while also enabling covariate adjustment and multiple testing correction. Using RNA-Seq data from a study of major depressive disorder, we show that NPDR with covariate adjustment removes spurious associations due to confounding by sex. We compare NPDR for a GWAS of lupus with and without adjustment for principal components.

80 | Genome-wide Gene-smoking Interaction Analysis of Lung Function in UK Biobank

Carl A. Melbourne[1]*, Nick Shrine[1], Chiara Batini[1], Ian P. Hall[2,3], SpiroMeta Consortium, Martin D. Tobin[1,4], Louise V. Wain[1,4]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR, Nottingham University Hospitals, Nottingham, United Kingdom; [3]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [4]NIHR, Leicester Respiratory Biomedical Research Centre, Leicester, United Kingdom.

Chronic obstructive pulmonary disease (COPD), characterized by severe airflow obstruction, is a leading cause of mortality worldwide. Smoking is the biggest risk factor, however there is a genetic component, with 279 signals identified to date for association with lung function and COPD. These signals account for modest proportions of lung function heritability, and not all smokers develop COPD, so we must consider other contributors such as gene-smoking interactions.

We present the largest genome-wide gene-smoking (ever/never smoker) interaction analysis in lung function (FEV_1, FVC, FEV_1/FVC and PEF) to date using 303,612 unrelated European individuals from UK Biobank. Phenotypes were inverse normalised adjusting for age, sex and height, with 10 principal components and genotype array adjusted for during analysis. Interaction effect was determined using a gene-smoking interaction term in a linear regression model. Replication was sought using the SpiroMeta consortium (22 studies and 71,067 individuals).

Analysis of 8,647,748 variants identified 53 independent genetic signals at, none of which have previously been implicated for lung function or COPD. Twenty six signals were imputed with high quality in the SpiroMeta consortium (effective sample size >50,000), of which 11 had consistent direction of effect in ever and never-smokers. None reached a Bonferroni corrected threshold of for replication.

These 53 signals may influence estimates of relative and absolute genetic risk for poor lung function and COPD, and aid in the development of personalised medicine based on smoking behaviour. Larger sample sizes with denser imputation are however required to establish these signals as true positives.

81 | Germline Mutations in the *BRCA1* Gene are Associated with Increased Risk For Additional Cancers Including Female Reproductive System Cancers

Candace D. Middlebrooks[1]*, Kenzhane Pantin[2], Mark Stacey[3], Carrie Snyder[3], Trudy Shaw[3], Marc Rendell[4], Peter Silberstein[5], Murray Joseph Casey[3,6], Joan E. Bailey-Wilson[1], Henry T. Lynch[3]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America; [3]Hereditary Cancer Center, Creighton University, Omaha, Nebraska, United States of America; [4]The Rose Salter Medical

Research Foundation, Newport Coast, California, United States of America; [5]Department of Hematology/Oncology, Creighton University, Omaha, Nebraska, United States of America; [6]Department of Obstetrics and Gynecology, Creighton University, Omaha, Nebraska, United States of America.

Mutations within the *BRCA1* gene have been linked to up to an 80% lifetime risk of breast cancer as well as increased risk for ovarian, pancreatic and melanoma cancers. In this study we examined families with known germline mutations in *BRCA1* after long-term follow-up to determine whether carriers experience higher rates of other cancers that have not yet been associated with germline mutations in the *BRCA1* gene.

We studied 127 Hereditary Breast and Ovarian Cancer (HBOC) syndrome families (N = 23,078 individuals who have been followed at Creighton University) in which a causal mutation in the *BRCA1* gene was identified. We performed survival analysis and a mixed effects cox regression with age at follow-up or cancer event as our time variable and presence or absence of *BRCA1*-related or other cancers (separate analyses) as our indicator variable. The survival curves showed a significant age effect with carriers having a younger age at cancer onset for *BRCA1*-related (as expected) as well as other cancers than that of non-carriers. The cox regression models were also highly significant (P value = 1.77E-37 and P value = 1.04E-07 for the *BRCA1*-related and other cancers, respectively). Of the cancers with enough samples to do stratified analyses, cervix, uterine, skin, lymphoma and colon cancers occurred at higher rates and at earlier ages in mutation carriers.

These analyses support the hypothesis that the *BRCA1* mutations carriers of HBOC syndrome have increased risk for early onset of several additional cancer types, especially cancers that arise in estrogen-influenced tissues.

82 | Genetic Association Testing with Multivariate Outcomes: Methods Comparison with Application to Cognition and Eye Disease

Zahra Montazeri[1]*, Fahimeh Moradi[1], Joycelyne E. Ewusie[1], Kelly Burkett[2], Ellen E Freeman[1,3,4], Marie-Hélène Roy-Gagnon[1]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada; [2]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada; [3]Maisonneuve-Rosemont Hospital, Montreal, Canada; [4]Department of Ophthalmology, University of Ottawa, Ottawa, Canada.

There is growing evidence that some complex traits are caused by overlapping biological mechanisms and thus under common genetic influences. In addition, complex traits are often measured by multiple correlated phenotypes. Analyzing the different related traits in multivariate analyses could provide increased power to detect genetic associations but is especially challenging with both discrete and continuous phenotypes. Retinal diseases (like age-related macular degeneration (AMD) and glaucoma) and cognitive decline are thought to share biological mechanisms. In addition, cognitive function can be measured by several related cognitive tests. To study the effect of different genetic variants on retinal diseases and cognition, we analyze data from 312 participants recruited from the Ophthalmology clinics of the Maisonneuve-Rosemont Hospital (Montreal, Canada). Variables collected included 6 continuous cognitive phenotypes, 2 discrete eye-disease phenotypes (AMD or Glaucoma vs. normal vision), 39 candidate genetic variants and covariates like age and sex. We first used univariate and multivariate regression to assess genetic associations within subsets of multivariate outcomes. For example, we found SNPs rs1061170 and rs10808746 to be associated with glaucoma based on the univariate logistic regression analyses. These two SNPs also showed nominal significance with cognitive phenotypes in multivariate linear regression analyses. Using simulations based on our data, we assess the performance of different methods that can accommodate both discrete and continuous phenotypes in multivariate analyses, such as GAMuT (Gene Association with Multiple Traits; Broadaway *et al*., 2016) and apply these methods to our data.

83 | Highly Aggregated Lung Cancer Families Show Significant Linkage to Chromosome 12q23.3 for Cancer Risk

Anthony M. Musolf[1]*, Claudio W. Pikielny[2], Diptasri Mandal[3], Richard K. Wilson[4], Ann G. Schwartz[5], Susan M. Pinney[6], Christopher I. Amos[7], Ramaswamy Govindan[8], Joan E. Bailey-Wilson[1] for the Genetic Epidemiology of Lung Cancer Consortium (GELCC)

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2]Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, United States of America; [3]Department of Genetics, Louisiana State University Health Science Center, New Orleans, Louisiana, United States of America; [4]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, United States of America; [5]Karmanos Cancer Institute, Wayne State University, Detroit, Michigan, United States of America; [6]Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, Ohio, United States of America; [7]Baylor College of Medicine, Houston, Texas, United States of America; [8]Division of Oncology,

Washington University School of Medicine, St. Louis, Missouri, United States of America.

Lung cancer (LC) kills more people than any other cancer in the United States. It is known that lung cancer is a complex trait caused by both environmental and genetic factors, yet the genetic etiology of lung cancer remains poorly understood. In this study, we have performed whole exome sequencing (WES) on 262 individuals from 28 extended families that have a strong history of LC and are highly aggregated for the phenotype. The WES was recalled with PICARD/GATK and standard quality controls were performed, leaving approximately 500,000 SNVs and indels for analysis.

Parametric genetic linkage analysis was performed on these families using two distinct models – the lung cancer only (LCO) model, where only lung cancer patients were coded as affected, and the all inherited cancers (AIC) model, where other inherited cancers were coded as affected as well (pedigrees averaged 1–2 additional non-lung cancer patients). All unaffected individuals were given an "unknown" phenotype. Both models assumed an autosomal dominant mode of inheritance with a disease allele frequency of 1% and a penetrance of 80% for carriers and 1% for non-carriers. The AIC model yielded a genome-wide significant result at rs61943670 in the RNA polymerase III gene *POLR3B* at 12q23.3. *POLR3B* has been implicated somatically in lung cancer but this germline finding is novel. Interesting genome-wide suggestive haplotypes were also found within individual families, particularly near *SSPO* at 7p36.1. Functional work on *POLR3B* and several of the best candidates for the individual family signals is planned for future studies.

84 | Metasubtract: An R-package to Analytically Produce Leave-one-out Meta-analysis Summary Statistics

Ilja M. Nolte*

Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.

Background: Consortia for many common diseases exist that made the summary statistics from a meta-analysis of multiple genome-wide association studies (GWAS) freely available. These meta-GWAS summary statistics could for instance be used for constructing polygenic risk scores. However, if the summary statistics are used for validation in one of the cohorts that was included in the meta-analysis, this will yield too optimistic results. For unbiased results the validation cohort needs to be independent from the meta-GWAS results. Usual practice is to contact the consortium and to ask them for meta-GWAS results with the validation cohort left out. I developed the R package "MetaSubtract" to subtract the results of the validation cohort from the meta-GWAS results analytically. For this package it is sufficient to have the meta-GWAS results and the cohort's GWAS results that have been contributed.

Methods: The statistical formulas a meta-analysis were inverted to compute corrected summary statistics of a meta-GWAS leaving one cohort out. These formulas have been implemented in MetaSubtract for different meta-analyses methods (fixed effects inverse variance, sample size or z-score weighted). It can take into account if single or double genomic control correction was applied. It can be used for whole GWAS, but also for a limited set of genetic markers.

Results: Results obtained by MetaSubtract are identical to those calculated using meta-analysis leaving the validation cohort out.

Discussion: MetaSubtract allows researchers to compute meta-GWAS summary statistics that are independent of the GWAS results of the validation cohort without interference of the corresponding consortium.

Invited Abstract

85 | The Future of Genomic Studies Must be Globally Representative

Kari E. North

Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America.

The past decade has seen a revolution in human genetics that has empowered investigations into the biology of complex traits. Although these discoveries rely on genetic variation, association studies have overwhelmingly been performed in populations of European descent. Given the differential genetic architecture that is known to exist across populations, such bias in representation can exacerbate disparities and impact clinical guidelines and drug development. Critical variants will be missed if they are low frequency or absent in European populations. Additionally, effect sizes and their derived risk prediction scores derived in one population may not accurately extrapolate to other

populations. Here we demonstrate the value of diverse, multi-ethnic participants in large studies by providing an overview of strategies to improve global representation in research and highlighting the successes of individual studies and consortia, for example PAGE, TOPMed, and CCDG, which have provided unique knowledge. Specifically, we will outline best practices for performing genetic epidemiology in multiethnic contexts, to identify effect heterogeneity and improve fine mapping, and to demonstrate how limiting investigations to single populations impairs findings in the clinical domain and for risk prediction. We argue that lack of representation of diverse populations in genetic research will result in inequitable access to precision medicine and advocate for continued, large genome-wide efforts in diverse populations to maximize genetic discovery and reduce health disparities.

86 | Prostate Cancer Risks For Male *BRCA1* and *BRCA2* Mutation Carriers: Prospective Analysis of the EMBRACE Study Cohort

Tommy Nyberg[1]*, Debra Frost[1], Daniel Barrowdale[1], D. Gareth Evans[2], Marc Tischkowitz[3,4], EMBRACE collaborators, Douglas F. Easton[1], Antonis C. Antoniou[1]

[1]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Manchester Regional Genetics Service, Central Manchester University Hospitals NHS Foundation Trust, Manchester, United Kingdom; [3]Department of Medical Genetics, University of Cambridge, Cambridge, United Kingdom; [4]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Trust, Cambridge, United Kingdom.

BRCA1 and *BRCA2* mutations have been associated with prostate cancer (PCa) risk but a wide range of risk estimates has been reported, based mostly on retrospective studies. We used a prospective cohort of unaffected male *BRCA1* (N = 376) and *BRCA2* carriers (N = 447) identified through clinical genetics centres in the UK and Republic of Ireland (median follow-up: 5.6 years), to estimate age-specific incidences, standardised incidence ratios (SIRs) relative to population incidences, absolute risks of PCa, and modification of these risks by family history and mutation position. Sixteen *BRCA1* and 26 *BRCA2* mutation carriers were diagnosed with PCa during follow–up. *BRCA2* carriers had a SIR of 4.45 (95% CI 2.99–6.61), and absolute risk of PCa of 60% (95% CI 43%–78%) by age 85. For *BRCA1* carriers, the overall SIR was 2.35 (95% CI 1.43–3.88), with higher SIR at ages <65 (SIR = 3.57, 95% CI 1.68–7.58). However, the *BRCA1* SIR was not consistently statistically significant in sensitivity analyses that assessed potential screening effects in this cohort. PCa risks for *BRCA2* carriers increased with family history (HR per affected relative = 1.68, 95% CI 0.99–2.85). Mutations in the ovarian cancer cluster region of *BRCA2* (c.2831–c.6401), showed weaker association with PCa risk (HR = 0.37, 95% CI 0.14–0.96) compared to mutations outside this region. For *BRCA2* carriers, the association was stronger with Gleason score ≥7 (SIR = 5.07, 95% CI 3.20–8.02) than Gleason score ≤6 PCa (SIR = 3.03, 95% CI 1.24–7.44). The results confirm the high risk of aggressive PCa for *BRCA2* carriers and give some support for a weaker association in *BRCA1* carriers.

87 | Risk Prediction for Colorectal Cancer Based on Extended Family History and Body Mass Index

H.M. Ochs-Balcom[1]*, P. Kanth[2], J.M. Farnham[3], S. Abdelrahman[4], L.A. Cannon-Albright[3,5,6]

[1]Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, University at Buffalo, Buffalo, NY, United States of America; [2]Gastroenterology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States of America; [3]Genetic Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States of America; [4]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States of America; [5]George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT, United States of America; [6]Huntsman Cancer Institute, Salt Lake City, UT, United States of America.

Family history is a well-known risk factor for colorectal cancer (CRC), as is BMI. The joint effects of body mass index (BMI) and CRC family history on risk for CRC are not well-described. Linked genealogy data, self-reported height and weight from driver's licenses, and a decades old SEER cancer registry for Utah was analyzed to estimate CRC risk based on extended family history of CRC and BMI. Even in the presence of a positive family history for CRC, which increases risk, high midlife BMI also contributes significantly to risk for CRC. Increasing number of first degree relatives (FDR) is associated with higher RR for CRC for overweight/obese probands but not for under/normal weight probands (for probands with two CRC-affected FDRs, ignoring second (SDR) and third degree relatives (TDR) RR = 1.91 (95% CI: 0.52, 4.89) for under/normal weight probands and RR = 4.83 (95% CI: 2.86, 7.64) for overweight/obese probands. In the absence of CRC-affected FDRs, any number of CRC-affected SDRs did not significantly increase CRC risk for under/normal weight probands, but for overweight/obese probands with at least three CRC-affected SDRs the RR = 2.68 (95% CI: 1.29, 4.93). In the absence of CRC-affected

FDRs and SDRs, any number of CRC-affected TDRs did not increase risk in under/normal weight probands, but significantly elevated risk for overweight/obese probands with at least two CRC-affected TDRs was observed; RR = 1.41 (95% CI: 1.12, 1.74). For non-syndromic CRC, maximum midlife BMI affects risk prediction based on family history and should be taken into account for CRC risk prediction when possible.

88 | LDScore Regression Identifies Novel Associations Between Glioma and Auto-immune Conditions

Quinn T. Ostrom[1]*, Jacob Edelson[2], Jinyoung Byun[1,2], Younghun Han[1,2], Kyle M. Walsh[3,4], Christopher I. Amos[1,2], Melissa L. Bondy[1], GLIOGENE Consortium

[1]Department of Medicine, Section of Epidemiology and Population Sciences, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America; [2]Institute for Clinical and Translational Research, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America; [3]Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, United States of America; [4]Department of Neurosurgery, Duke University School of Medicine, Durham, North Carolina, United States of America.

Prior epidemiological studies in glioma have identified 25 germline risk variants, as well as risk associations with exposure to ionizing radiation (which increases risk) and history of allergies and aspirin use (which decrease risk). In this analysis we used LDscore regression, which leverages single SNP associations and known patterns of linkage disequilibrium (LD) to estimate the genetic correlation between phenotypes, to confirm prior associations as well as attempt to identify novel phenotype associations for traits not previously assessed that may improve genetic prediction for glioma. Summary statistics for glioma were obtained from a prior meta-analysis. Summary statistics for 13 immune- and atopy-related traits were obtained from the prior case-control studies and the UK Biobank. Data were filtered to include only SNPs with imputation INFO value >0.7, and minor allele frequency >0.01, excluding the HLA region. Pairwise genetic correlation (rg) between traits was generated using LDSC. Associations were considered significant at P value < 0.05. Significant negative correlations were identified between glioma and ulcerative colitis (rg = −0.4039, P value = $4.91\text{x}10^{-10}$), celiac disease (rg = −0.2028, P value = 1.18×10^{-4}), lupus (rg = −0.0956, P value = 0.0083), and multiple sclerosis (rg = −0.5755, P value = 4.46×10^{-9}). Associations were generally consistent in both GBM and non-GBM. There was a significant correlation between both self-reported (rg = −0.102, P value = 0.0233) and doctor diagnosed (rg = −0.116, P value = 0.0305) hayfever/allergic rhinitis and GBM only. This analysis confirms the previously identified protective effect of allergic rhinitis and identifies novel associations between multiple auto-immune traits and glioma. Further studies are necessary in order to identify the mechanism through which increased immune activity may lower risk of glioma.

89 | A Flexible Copula-based Approach for the Analysis of Secondary Phenotypes in Ascertained Samples

Karim Oualkacha[1]*, Fodé Tounkara[2], Geneviève Lefebvre[1], Celia M.T. Greenwood[3,4]

[1]Department of Mathematics, Université du Québec À Montréal, Montreal, Québec, Canada; [2]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada; [3]Lady Davis Research Institute, Jewish General Hospital, Montreal, Québec, Canada; [4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Québec, Canada.

Data collected for a genome-wide association study 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, naive analyses of secondary phenotypes may induce spurious associations in non-randomly ascertained samples.

Previously, retrospective likelihood-based methods have been proposed to correct for sampling biases arising in secondary trait association analyses. However, most such methods have been introduced to handle studies with a case-control design and hence a binary primary phenotype. As such, these methods are not directly applicable to more complicated study designs such as multiple-trait studies where the sampling mechanism also depends on the secondary phenotype, or extreme-trait studies, where individuals with extreme primary phenotype values are selected.

To accommodate these more complicated sampling mechanisms, only a few prospective likelihood approaches have been proposed. These approaches assume a normal distribution for the secondary phenotype (or the latent secondary phenotype) and a bivariate normal distribution for the primary-secondary phenotype dependence.

In this work, we propose a unified copula-based approach to appropriately detect genetic variant/secondary phenotype association in the presence of selected sampling schemes. We use both prospective and retrospective likelihoods to account for the sampling mechanism and use a copula model to allow for potentially different dependence structures between the primary and

secondary phenotypes. We demonstrate the effectiveness of our approach through simulation studies and by analyzing data from the Avon Longitudinal Study of Parents and Children cohort.

91 | Genome-wide Association Study of the Cerebrospinal Fluid Metabolome

Daniel J. Panyard[1]*, Burcu F. Darst[2], Qiongshi Lu[3], Corinne D. Engelman[1]

[1]Department of Population Health Sciences, University of Wisconsin, Madison, Wisconsin, United States of America; [2]Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America; [3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, United States of America.

An important preliminary step in understanding the cerebrospinal fluid (CSF) metabolome in neurodegenerative disorders, such as Alzheimer's disease (AD), is to elucidate the genetics of CSF metabolite levels in healthy, older individuals.

We conducted a genome-wide association study (GWAS) of 338 CSF metabolite levels in 156 cognitively healthy adults (age 55.4–85.5) from the Wisconsin Alzheimer's Disease Research Center (W-ADRC) study. Associations were tested using linear models for log-transformed, baseline CSF metabolite levels, adjusting for age, sex, and five principal components to adjust for population stratification. Of the 12 CSF metabolites with at least one Bonferroni-corrected (P value $< 5x10^{-8}$ / 338), statistically significant SNP-metabolite association, 10 replicated in an independent cohort of 136 cognitively healthy adults (age 45.5–74.1) from the Wisconsin Registry for Alzheimer's Prevention (WRAP). We compared our findings for six of these CSF metabolites that were also studied in a GWAS of blood metabolites (Long et al., 2017). Three of the associations in CSF overlapped with those found in blood, further validating the CSF results. The other three CSF associations did not overlap with those in blood, potentially indicating distinct genetic mechanisms operating in the CSF. Finally, a meta-analysis of W-ADRC and WRAP samples revealed a total of 16 CSF metabolites with statistically significant genetic associations (P value $< 7.3 \times 10^{-11}$).

Our results suggest that distinct genetic mechanisms likely influence CSF metabolites. These genetic-metabolomic associations may be used to impute CSF metabolite levels into large-scale GWAS summary statistics of AD-related phenotypes to investigate the role of metabolites in AD.

92 | An Adjusted Survival Tree Model in Search of Genetic Polymorphisms Predictive for Oxaliplatin Treatment in Colorectal Cancer

Hanla Park[1]*, Petra Seibold[1], Axel Benner[2], Lina Jansen[3], Federico Canzian[4], Michael Hoffmeister[3], Hermann Brenner[3,4,6], Jenny Chang-Claude[1,7]

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany; [3]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [4]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany; [6]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany; [7]Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

Oxaliplatin is a platinum drug often given in combination with other anticancer drugs to treat colorectal cancer (CRC). Several candidate gene studies have been conducted to identify susceptibility loci that influence the efficacy of oxaliplatin treatment. However, results have been inconsistent so that predictive genetic markers are currently not available for clinical practice. Therefore, we conducted a genome-wide association (GWA) analysis to identify novel predictive genetic variants associated with differential prognosis in CRC patients receiving oxaliplatin-based chemotherapy vs. others.

In total 1,400 stage II-IV patients that received primary chemotherapy in German population-based study (DACHS) were included, of which, ~38% of patients received oxaliplatin treatment. The analysis consisted of two steps, firstly multivariable Cox proportional hazards models were used to detect single-nucleotide polymorphisms (SNPs) that are associated with differential overall survival according to the type of chemotherapy (oxaliplatin-based vs. others). The selection of SNPs for a second step was based on the False Discovery Rate adjusted ($P < 0.2$) for the interaction term between SNPs and the types of the treatment. For the second step, model-based random forests will be applied with the selected SNPs to identify SNP-based patient subgroups with differential treatment outcome.

The GWA analysis identified SNPs that showed differential overall survival of patients receiving oxaliplatin-based chemotherapy compared to patients receiving other types of chemotherapy in 7p21.2, 14q12, 14q32.2, and 14q23.3. Analyses are ongoing, and findings from model-based random forests will be presented at the conference.

93 | Polygenic Risk Scores Accounting for LD: Estimation and Model Selection Based on GWAS Summary Statistics

Jack Pattee*, Wei Pan*

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America.

Polygenic risk scores are used to quantify the genetic risk associated with certain diseases or phenotypes. Polygenic scores can help predict disease risk for individuals and can be used infer the genetic architecture of complex polygenic phenotypes. There has been recent interest in developing methods to construct polygenic risk scores after accounting for LD among SNPs using GWAS summary statistic data. We propose a method to construct polygenic risk scores via penalized regression using summary statistic data and published genotypic reference data. Our method bears many similarities to existing method LassoSum but extends their framework to the Truncated Lasso Penalty (TLP), a non-convex penalty with better finite-sample performance and theoretical properties for true sparse models. We show via simulations that the TLP can produce sparser effect size estimates as compared to the LASSO penalty. To facilitate model selection, we propose a method of estimating model fitting criteria AIC and BIC using only GWAS summary data. These methods approximate the AIC and BIC in the case where we have a polygenic risk score estimated on summary statistic data, but no individual=level data as validation data. Additionally, we propose a so-called quasi-correlation metric, which quantifies the predictive accuracy of a polygenic risk score applied to out-of-sample data for which we have only GWAS summary data. In total, these methods facilitate estimation and model selection of polygenic risk scores on GWAS summary statistic data, and the application of these polygenic risk scores to out-of-sample data for which we have only summary statistic information. We demonstrate the utility of these methods by applying them to GWAS studies of lung cancer and height, respectively.

94 | Integrating Germline and Somatic Genetics to Identify Genes Associated with Lung Cancer

Jack W. Pattee[1]*, Xiaowei Zhan[2], Guangua Xiao[2], and Wei Pan[1]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America.

Genome-wide association studies (GWAS) have successfully identified many genetic variants associated with complex traits. However, GWAS experience power issues, resulting in the failure to detect certain associated variants. Additionally, GWAS are often unable to parse the biological mechanisms driving associations. An existing gene-based association test framework, the Transcriptome-Wide Association Study (TWAS), leverages gene expression data to increase the power of association tests and illuminate the biological mechanisms by which genetic variants modulate complex traits. We extend the TWAS methodology to incorporate somatic information from tumors. By integrating germline and somatic data we are able to leverage information from the nuanced somatic landscape of tumors. Thus, we can augment the power of TWAS-type tests to detect germline genetic variants associated with cancer phenotypes. We use somatic and germline data on lung adenocarcinomas from The Cancer Genome Atlas in conjunction with a meta-analyzed lung cancer GWAS to identify novel genes associated with lung cancer.

95 | Modeling Heterogeneity of Complex Traits Using Mixture Models and Secondary Phenotypes

Subrata Paul[1]*, Stephanie A. Santorico[1,2,3]

[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, Denver, Colorado, United States of America; [3]Biostatistics and Informatics, Colorado School of Public Health, Denver, Colorado, United States of America.

Most common human diseases and complex traits are etiologically heterogeneous. Genome-wide association studies (GWAS) aim to discover common genetic variants that are associated with complex traits, typically without considering heterogeneity. Previously, we incorporated heterogeneity in a case-control design using a finite mixture of binomial distributions built off ancestry clusters. A simulation study showed that, in the presence of heterogeneity, the binomial mixture model estimates the odds ratio with less bias compared to logistic regression while having comparable statistical power to detect association.

Through analysis on a vitiligo case-control study, the mixture model did not provide results aligned with an underlying subtype previously discovered through secondary phenotype data. We extend our prior work by considering the use of secondary phenotypes and propose

a mixture of factor analysis that also accounts for differential allele frequency between cases and controls. Multivariate quantitative phenotypes are modeled as functions of latent variables. The model simultaneously identifies if there are multiple underlying latent dimensions of the trait and detects if a genetic variant is associated with at least one of the subgroups. A simulation study will be performed with comparison to standard multivariate methods such as minP and TATES, and a case-control design.

96 | Interaction Analyses of *MUC5B* Risk Allele Status and the HLA Region for Idiopathic Pulmonary Fibrosis Susceptibility

Megan L. Paynton[1]*, Richard J. Allen[1], Tasha Fingerlin[2,3], Rebecca Braybrooke[4,5], UK ILD Consortium, Philip Molyneaux[6,7], David Schwartz[2,8,9], R. Gisli Jenkins[10,11], Edward J. Hollox[12] and Louise V. Wain[1,13],

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Center for Genes, Environment and Health, National Jewish Health, Denver, United States of America; [3]Department of Biostatistics and Informatics, University of Colorado, Denver, United States of America; [4]Division of Epidemiology and Public Health, University of Nottingham, Nottingham, United Kingdom; [5]National Institute for Health Research, Nottingham Biomedical Research Centre, Nottingham University Hospitals, Nottingham, United Kingdom; [6]NIHR Respiratory Clinical Research Facility, Royal Brompton Hospital, London, United Kingdom; [7]National Heart and Lung Institute, Imperial College, London, United Kingdom; [8]Department of Medicine, University of Colorado Denver, Denver, United States of America; [9]Department of Immunology, University of Colorado Denver, Denver, United States of America; [10]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [11]National Institute for Health Research, Nottingham Biomedical Research Centre, Nottingham University Hospitals, Nottingham, United Kingdom; [12]Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom; [13]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom.

Idiopathic pulmonary fibrosis (IPF) is a rare interstitial lung disease believed to be the result of an abnormal wound healing response leading to scarring in the lungs. The SNP rs35705950 near *MUC5B* is the largest genetic risk factor for IPF; each copy of the risk allele is associated with a five-fold increase in odds of IPF. The human leukocyte antigen (HLA) region, which encodes genes important in inflammatory processes, has also been implicated in IPF susceptibility. We investigated the association of HLA variation with IPF susceptibility in those with and without the *MUC5B* risk allele.

The complexity of the HLA region necessitates bespoke imputation strategies that capture SNPs, HLA gene allele and amino acid variation. We combined the latest SNP imputation panel (Haplotype Reference Consortium) and an HLA variation imputation panel (T1DGC) to test for an interaction effect between rs35705950 and 35,455 SNPs, 424 HLA alleles and 1,276 amino acid changes in the HLA region on IPF susceptibility. Analyses were performed using 2,127 IPF cases and 8,049 controls adjusting for sex and 10 principal components.

No signals passed a Bonferroni corrected threshold of $P < 2.8 \times 10^{-6}$, however three SNP interactions were suggestively significant in IPF susceptibility ($P < 5 \times 10^{-3}$). These SNPs were not in linkage disequilibrium with any HLA alleles and were in or near the genes, *HLA-DOA*, *LOC100294145* and *PPP1R10*.

We present the first interaction analysis between rs35705950 and the HLA region providing further understanding of the genetic aetiology of IPF susceptibility.

97 | Uterine Leiomyomata Polygenic Risk Score (PRS) Confers Novel Relationships in the Clinical Phenome

Jacqueline A. Piekos[1]*, Jacklyn N. Hellwege[1,2], Eric S. Torstenson[1,3], Yanfei Zang[4], Sarah A. Pendergrass[5], Electronic Medical Records and Genomics (eMERGE) Network, Dan Roden[1,6], Josh C. Denny[1,7], Todd L. Edwards[1,3], Digna R. Velez Edwards[1,7,8]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United states of America; [2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United states of America; [3]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United states of America; [4]Genomic Medicine Institute, Geisinger, Danville, Pennsylvania, United states of America; [5]Biomedical and Translational Informatics, Geisinger, Danville, Pennsylvania, United states of America; [6]Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center Nashville, Tennessee, United states of America; [7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United states of America; [8]Division of Quantitative Science, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, Tennessee, United states of America.

Uterine leiomyomata (UL) are common pelvic tumors in women, with prevalence up to 77%. Previous Genome-Wide Association Study (GWAS) have associated ~30 loci with UL but the etiology is still relatively unknown. We constructed a polygenic risk score (PRS) for UL from predominately European Ancestry (EA) individuals and used a Phenome-Wide Association Study (PheWAS) approach to gain understanding about the shared genetic contribution across many clinical phenotypes. PRS was constructed in imaging confirmed UL GWAS data

(N = 2,651 cases, 4,326 controls) in a subset of individuals from the Electronic Medical Records and Genomics (eMERGE) network and optimized in an independent set from BioVU (N = 5,179). Using a *P* value threshold of $P < 0.001$, 4,448 variants were included in the PRS. PheWAS analyses were performed in eMERGE, excluding samples used for PRS construction, race stratified (N = 53,116 EA, N = 5,583 African American, AA), using the PRS as the predictor for clinical disease phenotypes (N = 1,738) adjusted for sex, age, BMI, and 10 principal components. UL was the most significant phenotype for non-Hispanic EAs ($P = 9.58 \times 10^{-167}$). In a sex combined analysis and female only analysis, we detected 40 and 47 ($P < 2.7 \times 10^{-5}$) significant and 16 and 19 suggested significant associations, respectively. Associated phenotypes fell into categories of genitourinary, neoplasms, and sense organs. The sense organ phenotypes were comprised of various eye diseases. For the non-Hispanic AA group, no phenotypes reached genome wide significance, likely due to power. Our results indicate that UL may share genetic architecture with other diseases that is yet to be characterized.

98 | A Novel Statistical Test Identifies Eight Loci Associated with Two Non-syndromic Orofacial Cleft Subgroups in GWAS of Multi-Ethnic Case-Parent Trios

Debashree Ray[1,2]*, Sowmya Venkataraghavan[1], Wanying Zhang[1], Jacqueline A. Bidinger[1], Elizabeth J. Leslie[3], Margaret A. Taub[2], Mary L. Marazita[4,5,6], Terri H. Beaty[1]

[1]Department of Epidemiology, Johns Hopkins University School of Public Health, Baltimore, Maryland, United States of America; [2]Department of Biostatistics, Johns Hopkins University School of Public Health, Baltimore, Maryland, United States of America; [3]Department of Human Genetics, Emory University School of Medicine; Atlanta, Georgia, United States of America; [4]Department of Oral Biology, Center for Craniofacial and Dental Genetics, University of Pittsburg School of Dental Medicine, Pittsburg, Pennsylvania, United States of America; [5]Department of Human Genetics, University of Pittsburg School of Public Health, Pittsburg, Pennsylvania, United States of America; [6]Department of Psychiatry, University of Pittsburg School of Medicine, Pittsburg, Pennsylvania, United States of America.

Based on epidemiologic and embryologic patterns, non-syndromic orofacial clefts (OFCs) are commonly categorized as cleft lip with or without cleft palate (CL/P) and cleft palate alone (CP). Several genes have been identified from linkage, candidate gene and genome-wide studies of CL/P and CP separately, however, combined these explain only one-fourth of the estimated total heritability of risk to OFCs. Some evidence of shared genetic risk in *IRF6*, *GRHL3* and *ARHGAP29* regions exists but no gene outside *FOXE1* has approached genome-wide significance in GWAS of CL/P and CP together. To identify genetic variants influencing risk of both CL/P and CP, we used a new statistical method, metapleio2. Although originally designed to detect pleiotropic effects of genetic variants on two seemingly unrelated phenotypic traits, metapleio2 can be used to detect variants with non-null effects (may or may not be in the same direction) on the two common subgroups of OFCs. When applied to a combined multi-ethnic GWAS of 2,847 CL/P and 611 CP case-parent trios, we identified six loci with compelling candidate genes: *PAX7* ($p = 7.3 \times 10^{-8}$), *IRF6* ($p = 4.6 \times 10^{-12}$), *DLG1* ($p = 5.6 \times 10^{-7}$), *LIMCH1* ($p = 5.2 \times 10^{-7}$), *SHROOM3* ($p = 8.5 \times 10^{-7}$) and near *NOG* ($p = 6.3 \times 10^{-8}$) that each appear to increase risk for one cleft subgroup but decrease risk for the other. Additionally, we replicated known variants in *FOXE1* ($p = 1.8 \times 10^{-7}$) and identified one locus in *RAB8A* ($p = 7.1 \times 10^{-7}$) that influence risk for both CL/P and CP. In summary, we confirm some candidate genes and find evidence for new genetic regions either exerting shared risk or with opposite effects on CL/P and CP.

99 | Imputation of Missing Genotypes and Estimation of Relatedness Between Subjects Without Genetic Data Across Pedigrees

Mohamad Saad[1]*, Ehsan Ullah[1], and Ellen M. Wijsman[2]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [2]Division of Medical Genetics, Department of Medicine, and Department of Biostatistics, University of Washington, Seattle, USA.

Family-based imputation allows better imputation of rare variants compared to population-based imputation. With whole genome sequencing becoming inexpensive, there is an opportunity to investigate rare variation. In the search for disease-associated rare variation, family-based designs have again become common, because rare, highly-penetrant genotypes can segregate in pedigrees.

In pedigrees, for subjects without genotype data, phenotype data can be available. Family-based imputation has the strength of imputing genotypes on such subjects using the genotypes of the available relatives. This likely increases the sample size and therefore the statistical power if imputation is accurate. In association testing, when related subjects are included, a kinship matrix must be used to account for relatedness. Relationships may be known from pedigree structure, or can be inferred using observed genotypes. For subjects without genotype data, however, no approach is able to infer relatedness, which is crucial to control the type 1 error.

In this work, we assess the performance of family-based imputation (GIGI2) for imputing subjects with no genotype data. We also propose a solution for inferring relatedness between such subjects by incorporating posterior probabilities of missing genotypes in an Expectation-Maximization approach. Through simulation, we obtained an average correlation of 0.6 between imputed and observed subjects. Correlation reached one for some subjects. Imputation performance decreased with size of imputation reference panel. Our approach succeeded in inferring many relationship types. The average kinship estimates were 0.19, 0.1267, 0.0641, and 0.03068 for underlying kinships of 0.25, 0.125, 0.0625, and 0.03125, respectively, for pairs of subjects without genotype data.

100 | Association Analyses of Handgrip Strength Leveraging Longitudinal and Sequence Data from the Trans-omics for Precision Medicine (TOPMed) Program

Chloé Sarnowski[1]*, Han Chen[2,3], Mary L. Biggs[4], Sylvia Wassertheil-Smoller[5], Jan Bressler[2], Marguerite Irvin[6], Jeffrey R. O'Connell[7], Kathleen A. Ryan[7], Joanne Murabito[8,9], Kathryn L. Lunetta[1] for the TOPMed Longevity and Healthy Aging Working Group

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [3]Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [4]Cardiovascular Health Unit, Department of Medicine and Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [5]Department of Epidemiology and Population Health, Albert Einstein College of medicine, Bronx, New York, United States of America; [6]Department of Epidemiology, University of Alabama at Birmingham School of Public Health, Birmingham, Alabama, United States of America; [7]Program for Personalized and Genomic Medicine, Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, United States of America; [8]Framingham Heart Study, Framingham, Massachusetts, United States of America; [9]Section of General Internal Medicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, United States of America.

Background: Handgrip strength is a widely used proxy of muscular fitness and a marker of frailty. Low or declining handgrip strength predicts a range of morbidities and all-cause mortality. We compared different strategies to perform genome-wide association study on longitudinal handgrip measurements using sequence data from the Trans-Omics for Precision Medicine (TOPMed) program.

Methods: A total of 12,342 participants from six cohort studies (Amish, ARIC, CHS, FHS, HyperGEN, and WHI) had between two and six handgrip measures per exam from one to nine separate exams over time, totaling 32,266 observations. We selected the maximum observation per participant at each exam and used per participant: 1) all exams (ALL), 2) one exam (ONE), or 3) the mean of all exams (MEAN). We conducted association analyses with GMMAT using linear mixed models adjusted for age, sex, height, BMI, study, age×sex, BMI×sex, study×sex, and 11 ancestry principal components, with random effects for sex, study and kinship. In addition, for the ALL analysis, we included a random effect for participant to account for correlation across exams.

Results: Leveraging multiple measures per individual resulted in a 5–10% increase in effective sample size. The $-\log(P)$ from the three analyses were highly correlated ($r_{ALLvsONE} = 0.85$, $r_{ALLvsMEAN} = 0.97$, $r_{ONEvsMEAN} = 0.88$). The genomic control inflation factors for the three analyses were similar; the ALL analysis had the lowest lambda for low frequency and rare variants.

Conclusion: When available, the use of multiple observations per individual using mixed effect models can increase effective sample size and thus power while controlling type I error.

101 | Identifying Risk Factors Involved in the Common Versus Specific Liabilities to Substance Abuse: a Genetically Informed Approach

Tabea Schoeler[1], Eleonora Iob[2], Charlotte M. Cecil[3], Esther Walton[4], Andrew McQuillin[5], Jean-Baptiste Pingault[1,6]

[1]Division of Psychology and Language Sciences, University College London, United Kingdom; [2]Department of Behavioral Science and Health, University College London, United Kingdom; [3]Department of Child and Adolescent Psychiatry, Erasmus University Medical Center, Netherlands; [4]MRC Integrative Epidemiology Unit, Bristol Medical School, Population Health Sciences, University of Bristol, United Kingdom; [5]Division of Psychiatry, University College London, United Kingdom; [6]Social, Genetic and Developmental Psychiatry Centre, King's College London, United Kingdom.

The co-occurrence of abuse of multiple substances is thought to stem from a common liability that is partly genetic in origin. Genetic risk may indirectly contribute to a common liability through genetically influenced individual vulnerabilities and traits. To

disentangle the aetiology of common versus specific liabilities to substance abuse, polygenic scores (PGS) can be used as genetic proxies indexing such risk and protective individual vulnerabilities or traits. In this study, we used genomic data from a UK birth cohort study (ALSPAC, N = 4218) to generate 18 PGS indexing mental health vulnerabilities, personality traits, cognition, physical traits, and substance abuse. Common and substance-specific factors were identified based on four classes of substance abuse (alcohol, cigarettes, cannabis, other illicit substances) assessed over time (age 17, 20, and 22). In multivariable regressions, we tested the independent contribution of selected PGS to the common and substance-specific factors. Our findings implicated several genetically influenced traits and vulnerabilities in the common liability to substance abuse, most notably risk taking ($b_{\text{standardized}}$ = 0.14; 95%CI: 0.10,0.17), followed by extraversion ($b_{\text{standardized}}$ = −0.10; 95%CI: −0.13,−0.06), and schizophrenia risk ($b_{\text{standardized}}$ = 0.06; 95%CI: 0.02;0.09). Educational attainment (EA) and body mass index (BMI) had opposite effects on substance-specific liabilities such as cigarettes ($b_{\text{standardized-EA}}$ = −0.15; 95%CI: −0.19,−0.12; $b_{\text{standardized-BMI}}$ = 0.05; 95%CI: 0.02,0. 09), alcohol ($b_{\text{standardized-EA}}$ = 0.07; 95%CI: 0.03,0.11; $b_{\text{standardized-BMI}}$ = −0.06; 95%CI: −0.10,−0.02), and other illicit substances ($b_{\text{standardized-EA}}$ = 0.12; 95%CI: 0.07,0.17; $b_{\text{standardized-BMI}}$ = −0.08; 95%CI:−0.13,−0.04). This is the first study based on genomic data that clarifies the aetiological architecture underlying the common versus substance-specific liabilities, providing novel insights for prevention and treatment of substance abuse.

102 | Testing for Multiple Shared Variants in Two Traits with Summary Genetic Association Data

Simon M. Schoenbuchner[1]*, Chris Wallace[1,2], Paul J. Newcombe[1]

[1]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom; [2]Department of Medicine, University of Cambridge, Cambridge, United Kingdom.

The integration of results from multiple genetic association studies can contribute to our understanding of the molecular basis of disease by identifying genetic variants that simultaneously influence multiple traits. Existing colocalization tests require patient-level data from both studies or use summary statistics, but are typically applied under an assumption that a region of interest can contain at most one causal variant for each trait.

We view colocalization as a problem of simultaneous fine mapping of two traits. We therefore built on our existing summary statistic fine mapping approach, which places no restrictions on the number of causal signals for either trait, to develop a new formal colocalization test. A Bayesian stochastic model search is used to efficiently explore the joint causal configuration space, using estimates from the UK Biobank to account for linkage disequilibrium. An appealing feature of our algorithm is that it provides model-averaged posterior probabilities of colocalization, accounting for uncertainty over the causal configuration space.

We present simulation results showing that our method can identify shared variants that are missed by an existing summary data colocalization test and illustrate the method with an application to inflammatory bowel disease, using published gene expression data and summary disease association statistics.

103 | Change in Ancestry-related Assortative Mating in the United States: Implications for Genetic Diseases

Ronnie Sebro[1,2,3,4]*, Sarah Murray[4]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, United States of America; [2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, United States of America; [3]Department of Orthopedic Surgery, University of Pennsylvania, Philadelphia, United States of America; [4]Department of Radiology, University of Pennsylvania, Philadelphia, United States of America.

Introduction: The United States of America (USA) is a cosmopolitan, multiracial, multiethnic country with over 800 different ancestries. The aim of this study was to evaluate the rate of the change of endogamy across several generations in the United States and to evaluate how this affects genetic diseases.

Materials and Methods: The data were extracted from the 2000 and 2010 decennial USA Censuses and the 2011 to 2016 American Community Survey (ACS) Census. Data collected included individual age, sex, race, ancestry, and marital status. Spouse-pairs were considered to be from the Silent Generation (born 1928–1945), Baby Boomer (born 1946–1964), Generation X (born 1965–1981; or Millennials (born 1982–2000). Phi-coefficient estimates of endogamy in the 50 most common ancestries were calculated. McNemar's odds ratio was used to evaluate asymmetry in assortative mating. Statistics were all two-sided and P values <0.05 were considered statistically significant.

Results: Over 3 million spouse-pairs were available for analysis (1.5 million spouse-pairs from Census 2000, and over 1.5 million spouse-pairs from Census 2010). The data show strong ancestry-related assortative mating, which decreased across each generation. The ancestry-related assortative mating was weakest among European-derived populations ranging from 0.24 ($P < 0.05$) (between Danish and Germans). The largest overall asymmetry was seen between Dutch and the Taiwanese (17) – Dutch men were 17 times more likely to marry Taiwanese women than Taiwanese men were to marry Dutch women.

Conclusion: Ancestry-related assortative mating decreased across each generation consistent with decreased endogamy. Significant asymmetric mating was noted, which has implications for diseases affected by genetic imprinting.

104 | Smoothed Moving Landmark Analysis for the Age-dependent Effects of DNA Methylation on the Risk of Coronary Heart Disease

Bin Shi[1,2]*, Ziqiao Wang[1,3], Xuelin Huang[1], Peng Wei[1]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America; [2]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America' [3]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, Texas, United States of America.

Coronary heart disease (CHD) is the most common type of heart disease, which is the leading cause of mortality among adults in the US. Numerous evidence has been accumulated to show that DNA methylation plays important roles in the development of CHD. This effect is believed to vary by age. However, due to the economic burden, in most studies, DNA methylation is measured only once, at a different age for each subject, which prohibits estimating age-dependent effects of methylation on CHD risk by conventional statistical methods. Here we provide two novel approaches to tackle this analytical challenge, namely separate models and super models. By separate models, we divide all subjects into different groups by their age of methylation measurement, and then estimate the methylation effects in these groups separately using landmark analysis. These estimates along the follow-up ages dynamically approximate the effect profile of methylation on CHD onset risk. On the other hand, the super models combine the above separate landmark analyses together and borrow information between adjacent age intervals to construct a smoother and more stable profile of the methylation effects on CHD risk. Simulation studies confirm these advantages of the super models. We apply both approaches to the Framingham Heart Study of 1540 subjects with CHD event data and methylation profiling at a single clinical visit, for illustration and demonstration of the comparisons between them. Our novel analysis reveals that methylation levels at several CpG sites show declining age-dependent effects on the CHD risk.

105 | Comparison of Multiple Phenotype Association Tests Using Summary Statistics in Genome-wide Association Studies

Colleen M. Sitlani[1]*, Antoine R. Baldassari[2], Heather M. Highland[2], Chani J. Hodonsky[2], Barbara McKnight[3], Christy L. Avery[2]

[1]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America.

A majority of complex phenotypes have a large genetic component, underscored by genome-wide association studies identifying thousands of loci. Although single-phenotype studies are common, more researchers are exploring the benefits of assessing multiple phenotypes simultaneously. A number of statistical methods have been proposed for analyzing multiple phenotypes using summary statistics. Leveraging publicly available summary statistics enables studies of shared genetic effects, while avoiding challenges associated with individual-level data sharing. One challenge in choosing amongst multi-phenotype tests is that there is no uniformly-most-powerful test. Researchers rarely know the underlying structure of the associations between the multiple phenotypes and a variant, making it difficult to choose the ideal test. We focus here on adaptive tests that maintain power against multiple alternatives simultaneously: an adaptive sum of powered scores (aSPU) test, a unified score association test (metaU-SAT), an adaptive test in a mixed-models framework (mixAda), and two principal-component-based adaptive tests (PCAQ and PCO). These tests cannot determine true pleiotropy but can identify loci where more explicit evaluation of pleiotropy is warranted.

Our simulations highlight practical challenges that arise with low minor allele frequencies, non-normal multivariate distributions of test statistics, and singular or nearly singular covariance matrices across phenotypes. The aSPU test relies least on asymptotic and distributional assumptions, leading to less type I error in these scenarios. However, the more reliable performance sometimes comes at the cost of decreased power and/or ability to detect certain phenotype-variant relationship patterns. We illustrate these tradeoffs with multi-phenotype analyses of six quantitative electrocardiogram traits.

106 | The Association Between Common Risk Factors for Age-related Disease and DNA Methylation Clocks in an African American Population

Wei Zhao[1], Farah Ammous[1], Scott Ratliff[1], Thomas H. Mosley[2], Sharon L.R. Kardia[1], Jennifer A. Smith[1,3]*

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Memory Impairment and Neurodegenerative Dementia (MIND) Center, University of Mississippi Medical Center, Jackson, Mississippi, United States of America; [3]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, United States of America.

DNA methylation clocks (DNAmAge), biomarkers of cellular aging which represent the difference between epigenetic age and chronological age (DNAmAge acceleration), are associated with age-related chronic diseases and all-cause mortality. Multiple DNAmAge estimators, built upon methylation levels from different loci, have been developed to capture cellular aging. Examining the relationship between risk factors for age-related diseases and multiple DNAmAge estimators can increase understanding of how risk factors contribute to aging at cellular level. This study explored the association between common risk factors for age-related diseases and four DNAmAge estimators, including intrinsic (IEAA) and extrinsic epigenetic age acceleration (EEAA), PhenoAge acceleration and GrimAge acceleration, in African Americans from the Genetic Epidemiology Network of Arteriopathy (GENOA). We performed both cross sectional analyses (N = 1100) and longitudinal analyses (N = 266) using clocks calculated from the Illumina Infinium HumanMethylationEPIC BeadChip. In cross sectional analyses, gender, education, BMI, smoking and alcohol consumption were all independently associated with GrimAge acceleration ($p < 0.05$). Only smoking and BMI were associated with PhenoAge acceleration. Gender, education and former smoking were associated with IEAA, whereas gender and education were associated with EEAA. The effect of smoking and education on GrimAge varied by gender. Longitudinal analysis suggests that age and BMI continued to increase GrimAge acceleration, and that age and current smoking continued to increase PhenoAge acceleration. In conclusion, common risk factors for age-related disease were associated with many DNAmAge estimators. However, the contribution of each risk factor to DNAmAge varies by clock, which suggests different methylation clocks may capture different components of cellular aging.

107 | Accounting for Covariates in Tiled Regression Analysis of Complex Traits

Alexa J.M. Sorant*, Jeremy A. Sabourin, Alexander F. Wilson

Computational and Statistical Genomics Branch, National Human Genome Research Institute, Baltimore, Maryland, United States of America.

Complex traits typically have both genetic and non-genetic components. Tiled regression, which determines an additive model with stepwise regression applied in stages, can incorporate covariates with several methods. The traditional method is to regress the trait on the covariates and then analyze the residuals as an "adjusted" trait considering genetic predictors. It is also possible to force covariates into each considered model or to test them as additional potential predictors. Forcing inclusion produces a predictive model similar to that of pre-adjustment, except that the covariate coefficients are not fixed, but determined in combination with the genetic components considered at each stage. When covariates are tested for inclusion, they may or may not be retained in the model. In this study, four ways of handling covariates were considered: (1) ignoring them, (2) pre-adjusting the trait, (3) forcing inclusion in every model, and (4) testing covariates for inclusion in the final model. Tiled regression was performed with each method for several previously studied metabolic phenotypes from the Trinity Student Study and for several simulated traits using the same genotypes. Analysis of methylmalonic acid, serum vitamin B12 and holo-transcobalamin concentrations (log-transformed), considering covariates age, sex, body mass index and B6 supplementary intake, produced similar predictive models with all analysis methods. With 200 replicates of simulated phenotypes based on five independent SNPs and two covariates, all

analysis methods had similar ability to detect causal SNPs, with pre-adjustment or keeping all considered covariates usually having slightly higher power and type I error rate.

108 | Multi-omic Analysis of Discordant and Concordant Sib-pairs with Inflammatory Bowel Disease

Andrew B. Stiemke[1], Steven R. Brant[2], Claire L. Simpson[1]

[1]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, United States of America; [2]Department of Medicine, Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick and Piscataway, and Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine and Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, United States of America.

Inflammatory bowel disease (IBD) is an immune-mediated chronic intestinal disorder that is typically divided into two distinct types; ulcerative colitis (UC) and Crohn's disease (CD). Over 240 IBD susceptibility loci have been identified, however, much of the etiology remains unexplained and, even within families, the disease can have a heterogeneous clinical presentation. Here we present a study attempting to discern the cause of that heterogeneity in discordant and concordant sib-pairs by comparing whole exome sequencing, epigenetic differences using DNA methylation analysis, and gene expression using next generation RNA sequencing (RNAseq).

A total of 96 discordant and concordant sib-pairs were included in the study. Genomic DNA and RNA were extracted from peripheral blood mononuclear cells that were isolated using cell sorting from whole blood. DNA methylation values were measured using the Illumina EPIC array. Whole exome sequencing was performed using the Illumina TruSeq Exome kit, and RNAseq was performed using the Illumina RNA Kit v2. Of 5000 initial variants, five genes— NOTCH1; SRRM2; ZNF276; GATAD2A; and DPH1—remain after cross-matching preliminary exome and methylation analyses. RNAseq analysis is currently underway and results will be presented.

Combining data from multiple technologies is an important next step for interpreting the results of genome-wide association studies (GWAS) and there is currently no agreement on the best approach for integrating these data. Given the known genetic and clinical heterogeneity in IBD, using discordant and concordant sib-pairs attempts to leverage the expected sharing between the sibs as supporting information to inform future studies.

109 | A Novel Method to Estimate the Distribution of Ancestral DNA Sequence

Jianping Sun*

Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, North Carolina, United States of America.

Large international genomic projects, such as the HapMap Project and 1000 Genomes Project, have collected data from thousands of individuals in various populations. These freely accessible databases offer great opportunities for scientists to study human evolution history by constructing hierarchical trees from current descendant DNA sequences. Such hierarchical tree plays a critical role in genome research, because it helps people finding the evolutionary history, identifying and mapping the potential genome which is susceptible to cause disease.

Due to the biological complexities and computation burden, most developed methods suffer from the limitations of strong assumptions or usages of approximations. In this talk, I will present a novel statistical method to estimate ancestral distribution by using mixture models and considering realistic biology complexities such as mutation and recombination. In addition, a composite likelihood based on Markov Chain property will be used for model inference in order to reduce computation burden. Finally, the performance of estimator will be examined via simulation studies.

110 | Polygenic Risk Scores and Epistatic Components for Alzheimer's Disease Prediction

Rui Sun[1,2]*, Xiaoxuan Xia[1,2], Maggie H. Wang[1,2]

[1]Division of Biostatistics and Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China; [2]Division of Biostatistics and Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China CUHK Shenzhen Research Institute, Shenzhen, China.

Genome-wide association studies (GWAS) have made great achievements in the past decade and have identified many common variants by single-marker detection for Alzheimer's disease (AD). However, a vast of missing heritability for AD exists. Possible reasons for the missing heritability include the existence of polygenic and epistatic components. In this study, we applied a novel method for pure interaction effect selection and then integrated polygenic risk scores and

epistasis components for Alzheimer's disease risk prediction in a real GWAS dataset. Prediction accuracy of models with and without epistasis was compared to evaluate the contribution of epistatic variables to risk prediction.

111 | Multi-ancestry Genome-wide Meta-analysis Accounting for Gene-education Interactions in Up to 227,850 Individuals Identifies Several Novel Lipid Loci

Yun J. Sung[1]*, Karen L. Schwander[1], Amy R. Bentley[2], Paul S. de Vries[3], Tuomas Kilpeläinen[4], Raymond Noordam[5], Solomon K. Musani[6], Hugues Aschard[7], Lisa de las Fuentes[1,8], on behalf of the CHARGE Gene-Lifestyle Interactions Working Group

[1]Division of Biostatistics, Washington University School of Medicine, St Louis, Missouri, United States of America; [2]Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [3]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas, Houston, Texas, United States of America; [4]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [5]Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; [6]Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, United States of America; [6]Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur, Paris, France; [8]Department of Medicine, Cardiovascular Division, Washington University School of Medicine, St Louis, Missouri, United States of America.

Introduction: High- and low-density lipoprotein cholesterol and triglycerides are influenced by genetic and lifestyle factors. Educational attainment is among the most widely-used indices of socioeconomic status. We hypothesize that gene-education interactions will help identify new lipid loci.

Methods: We conducted a multi-ancestry gene-education interaction study with "Some College" (yes/no) and "Graduated College" (yes/no). In Stage 1 (N = 110,319), genome-wide analyses were performed at ~15 million imputed variants. All suggestive variants ($P < 1 \times 10^{-6}$) in Stage 1 were followed up in Stage 2 (N = 117,531). We used a one degree of freedom (DF) interaction test and a joint 2DF test of genetic main and interaction effects.

Results: Combined meta-analyses identified 16 novel loci associated with lipids ($P < 5 \times 10^{-8}$). Ten loci were identified in African ancestry and one in Asian ancestry. Four loci were identified because association significantly differed by educational attainment, showing interaction evidence. *MBOAT4,* identified in Africans, plays roles in fatty acid metabolism and cognition via Ghrelin activation. Other novel loci include *PTPRE* implicated in high-fat diet-induced obesity, leptin sensitivity, and glucose homeostasis. *SLC1A3* and *NPTX2* are linked with excitatory brain neurotransmitter signaling. *GRIN2B* plays a role in mediating a crosstalk between fat and memory in the hippocampus.

Conclusions: Identified lipid loci may elucidate the role of an underlying mechanism for interaction of lifestyle factors captured by educational attainment in the genetic regulation of lipid levels. We also show the importance of including diverse populations, particularly in studies of interactions with lifestyle factors, where genomic and lifestyle differences by ancestry may contribute to novel findings.

112 | MOPower: a Web Application and Reporting Tool for the Simulation and Power Calculation of Multi-omics Study Data

Hamzah Syed[1,2]*, Georg W. Otto[1,2], Daniel Kelberman[1,2], Sergi Castellano[1,2]

[1]GOSGene, Genetics and Genomic Medicine, UCL GOS Institute of Child Health, University College London, London, United Kingdom; [2]NIHR Great Ormond Street Hospital Biomedical Research Centre, London, United Kingdom.

Multi-omics studies seek to uncover the underlying mechanisms of clinical phenotypes by integrating information from the genome, transcriptome, epigenome, metabolome and proteome. Correct integration of multiple data types is an increasingly difficult feat to achieve, with method development still in its early stages. This issue becomes more complicated in rare disease studies where the sample size is small, and the analysis of single-omics is hugely underpowered. Scenarios involving the power of multi-omics studies have not been extensively researched. Therefore, we have developed the interactive R-shiny web application and Python reporting tool MOPower, which provides essential information on the optimal sample size and choice of integration model. MOPower simulates data using statistical distributions or real study/reference data via bootstrapping for realistic disease-specific simulation. Several different omics features are simulated such as SNP's, gene expression and DNA methylation for binary, longitudinal or time-to-event outcomes. Each data replicate can be analysed using 15 of the latest multi-omics integration models such as factor, mediation and cluster analyses. The power is interpreted as the correlation between omics features or the power to

detect an association with the phenotype of interest through analysis of *n* number of replicates. The output from the power analysis is produced as a detailed HTML or PDF report, that includes power and false discovery rate plots with a statement of results interpretation. Initial findings show that increased dispersion of gene expression data is a crucial contributor to decreased power and analysis using similarity-network fusion is beneficial with small sample sizes.

113 | Comparison of Pathway Guided Random Forests Approaches for the Integration of Biological Knowledge and Omics Data

Stephan Seifert, Silke Szymczak*

Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany.

High-throughput technologies allow comprehensive characterization of individuals on many molecular levels. However, training prediction models on omics data is challenging. A promising solution is the integration of external knowledge about structural and functional relationships into the modelling process. We compared four published random forest based approaches using two simulation studies and nine experimental data sets.

In the first simulation study with many associated pathways, synthetic features (SF) and prediction error (PE) showed high empirical power across different pathway sizes, degrees of association and correlation patterns, whereas Hunting and Learner of Functional Enrichment (LeFE) were only able to detect large pathways with strong signal. In the complete null scenario, Hunting and LeFE falsely detected pathways with strong pairwise correlation, while SF had increased false discovery rates for all pathways.

The second simulation study with a single associated pathway and realistic correlation patterns showed that LeFE, Hunting and SF, in contrast to PE, had high empirical power. In the complete null scenario, SF was the only method with elevated false discovery rates.

In the experimental data sets PE and SF always identified the target pathway but additionally selected almost all other pathways. Hunting and LeFE had lower detection frequencies but rarely selected additional pathways.

The self-sufficient PE approach should be applied when large numbers of relevant pathways are expected. Competing methods (Hunting and LeFE), however, should be used when low numbers of relevant pathways are expected or the most strongly associated pathways are of interest.

114 | Epigenetic Aging of the Placenta: Sexually Dimorphic Influence on Fetal Growth and Risk of Low Birth Weight

Fasil Tekola-Ayele[1]*, Tsegaselassie Workalemahu[1], Gezahegn Gorfu[2], Deepika Shrestha[1], Ronald Wapner[3], Cuilin Zhang[1], Germaine Buck Louis[4]

[1]Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, United States of America; [2]Department of Clinical Laboratory Science, College of Nursing and Allied Health Sciences, Howard University, Washington, District of Columbia, United States of America; [3]Department of Obstetrics and Gynecology, Columbia University, New York, New York, United States of America; [4]College of Health and Human Services, George Mason University, Fairfax, Virginia, United States of America.

Fetal growth of males than females is more adversely affected by *in-utero* insults. Placental epigenetic mechanisms may underlie sexually dimorphic influences on fetal growth. We examined sex-specific associations of placental epigenetic aging with fetal growth measures, neonatal anthropometry, and risk of low birth weight. DNA methylation was measured on placenta samples (n = 152 males and 149 females) obtained as part of the NICHD Fetal Growth Studies-Singletons. Placental DNA methylation age was estimated using 62 methylation markers. Placental epigenetic age acceleration (PAA) was defined as the difference between DNA methylation age and gestational age in weeks. Associations of PAA with longitudinal trajectories and weekly measures (at 13–40 weeks gestation) of fetal growth (fetal weight, abdominal and head circumferences, humerus and femur lengths), and risk of low birth weight were tested using regression models. Longitudinal analyses reflected that PAA was associated with reduced fetal weight among males ($P = 0.04$) but increased all fetal growth measures among females ($P < 0.05$). In cross-sectional analyses, PAA was inversely associated with fetal weight, abdominal circumference, and biparietal diameter at 32–40 weeks among males ($P < 0.05$) but positively with all fetal growth measures among females beginning at week 13 ($P < 0.05$). A 1-week increase in PAA was associated with 1.95-fold (95% CI 1.20, 3.16) increased odds for low birth weight among male fetuses. In conclusion, we found that fetal growth was significantly reduced in males but not females exposed to a rapidly aging placenta. Epigenetic aging of the placenta may underlie observed sex differences in fetal and neonatal outcomes.

115 | SNP-based Epistasis Detection – a Lost Cause?

Kristel Van Steen[1,2*], Junior Ocira[1], Diane Duroux[1], Jason Moore[3]

[1]BIO3 - GIGA-R Medical Genomics, University of Liège, Liège, Belgium; [2]WELBIO researcher, University of Liège, Liège, Belgium; [3]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, United States of America.

Although an increasing number of efforts are being made to develop analytics that have increased power to identify epistasis using GWAS panels of SNPs, disappointing progress has been made in obtaining replicable results with functional or actionable repercussions. For some, this may be a reflection of the weak contributions of epistasis to genetic trait variability for most human complex traits. For others, it may be a trigger to develop a generic strategy that encompasses different analytic viewpoints, allowing the user to draw robust, replicable and "ensemble" conclusions. To achieve the latter goals, we dissect the components leading to heterogeneous epistasis results and present a critical and insightful evaluation of statistical epistasis analysis workflows. We present an epistasis detection protocol, that not only takes advantage of the heterogeneity in analysis strategies, but at the same time gives literature-based cautionary notes related to pragmatic choices. It includes elements to move from localization to function, to help elucidating the molecular mechanisms playing a synergetic role in human complex diseases. We furthermore show the relevance of epistasis research: over a decade of lessons learned impact the modelling of regulatory effects at the level of gene expression, the development of risk scores, and the assessment of individual contributions to statistical or biological networks, among others.

116 | Allele-specific QTL Fine-mapping with Plasma

Austin T. Wang[1,2,3]*, Anamay Shetty[3,4], Edward O'Connor[3], Connor Bell[3], Mark M. Pomerantz[3], Matthew L. Freedman[3,5,6], Alexander Gusev[3,7]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America; [4]Cambridge University, Cambridge, United Kingdom; [5]The Eli and Edythe L. Broad Institute, Cambridge, Massachusetts, United States of America; [6]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America; [7]Brigham & Women's Hospital, Division of Genetics, Boston, Massachusetts, United States of America.

Although quantitative trait locus (QTL) associations have been identified for many molecular traits such as gene expression, it remains challenging to distinguish the causal nucleotide from nearby variants. In addition to traditional QTLs by association, allele-specific (AS) QTLs are a powerful measure of cis-regulation that can be less susceptible to technical/environmental noise. We introduce PLASMA (PopuLation Allele-Specific MApping), a novel, LD-aware method that integrates QTL and asQTL information to fine-map causal regulatory variants. We demonstrate through simulations that PLASMA successfully detects causal variants over a wide range of genetic architectures. We apply PLASMA to RNA-Seq data from 524 kidney tumor samples and show that over 13 percent of loci can be fine-mapped to within 5 causal variants, compared to less than 2 percent of loci using existing QTL-based fine-mapping. Furthermore, PLASMA achieves greater power at 50 samples than QTL fine-mapping does at over 500 samples. PLASMA achieves a 6.4-fold reduction in median 95% credible set size compared to QTL-based fine-mapping. We apply PLASMA to H3K27AC ChIP-Seq from 28 prostate tumor/normal samples and demonstrate that PLASMA is able to prioritize markers even at small samples, with PLASMA achieving a 1.4-fold reduction in median 95% credible set sizes over QTL-based fine-mapping. Variants in the PLASMA credible sets for RNA-Seq and ChIP-Seq were more significantly enriched for open chromatin and chromatin looping (respectively) than credible variants from existing methods. Our results demonstrate how integrating AS activity can substantially improve the detection of causal variants from existing molecular data and at low sample size.

117 | Gene-based Rare Variant Association Tests for Ancestry-matched Case-control Data

Chaolong Wang[1]*, Baoluo Sun[2], Shanshan Cheng[1], Zengmiao Wang[3], Minghua Deng[3,4,5], Han Chen[6,7]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical School, Huazhong University of Science and Technology, Wuhan, Hubei, China; [2]Department of Statistics and Applied Probability, National University of Singapore, Singapore; [3]Center for Quantitative Biology, Peking University, Beijing, China; [4]LMAM, School of Mathematical Sciences, Peking University, Beijing, China; [5]Center for Statistical Sciences, Peking University, Beijing, China; [6]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [7]Center for Precision Health, School of Public Health & School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America.

With an increasingly large amount of human sequencing data available, analysis incorporating external controls becomes a popular and cost-effective approach to boost statistical power in disease association studies. To prevent spurious association due to population stratification, it is important to carefully match the ancestry backgrounds of cases and external controls. However, popular rare variant association tests based on logistic regression models, including the burden test, sequence kernel association test (SKAT) and the mixed effects score test (MiST), which is a hybrid version of burden and SKAT, are conservative for matched case-control samples when all matched strata have the same case-control ratio and might become anti-conservative when case-control ratio varies across strata. To account for the matching structure, we propose gene-based tests based on a conditional logistic regression (CLR) model, namely CLR-burden, CLR-SKAT, and CLR-MiST. We show that the CLR model coupled with ancestry matching is a general approach to control for population stratification. Through extensive simulations of population stratification and matching schemes, we demonstrate that both CLR-burden and CLR-SKAT are more powerful than standard burden test and SKAT respectively in ancestry-matched data while robustly controlling the type 1 error, and CLR-MiST is most powerful under a wide range of scenarios. Furthermore, because CLR-based tests allow for different case-control ratios across strata, a full-matching scheme can be employed to fully utilize available cases and controls to accelerate the discovery of disease genes.

118 | Estimation of Mediating Effect in a Mediation Model with a Censored Mediator in a Case-control Study

Jian Wang[1]*, Jing Ning[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.

A mediation model assesses the direct and indirect effects of an initial variable on an outcome by including one mediator. In practice, mediation models can involve a mediator with censored data. The current approaches for mediation analysis with a censored mediator are focused on models with continuous outcomes. However, the mediation models can involve binary outcomes based on data from a case-control study. Such models will result in biased estimations for the initial variable-mediator association if using standard approaches directly. In this study, we proposed an approach to analyze the mediation model with a censored mediator given data from a case-control study, based on the semiparametric accelerated failure time model along with a pseudo-likelihood function. We adapted the measures for assessing the indirect and direct effects using counterfactual definitions. We conducted simulation studies to investigate the performance of the proposed approach and compared it to those of the naïve approach and the complete-case approach. The proposed approach was applied to the mediation study of genetic variants, a woman's age at menopause, and type 2 diabetes based on a case-control study of type 2 diabetes, and the results showed that there is no mediating effect from the age at menopause on the association between the genetic variants and type 2 diabetes.

119 | Detecting Tumor-immunity-specific Expression QTL in Cancer

Xuefeng Wang[1], Sarah Urbut[2], Gao Wang[2], and Xiaoqing Yu[1]

[1]H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, United States of America; [2]University of Chicago, Chicago, Illinois, United States of America.

Immunotherapy has produced promising results in treating cancers. Recent studies that aim to predict which patients can benefit from an immune checkpoint blocker have been largely focused on tumor molecular profiling such as tumor mutation burden and T cell infiltrations. However, whether the immunity landscape in tumor can be favorably or adversely affected by polymorphisms carried in the germline remains unknown and understudied. Here we conduct the largest investigation of tumor-immunity-specific expression QTL to date, or tis-eQTL, to systematically identify germline genetic variants that affect immune landscape in tumor. Based on recently developed multivariate adaptive shrinkage (MASH) test, we analyzed genomic data from 10,380 cancer patients from 33 cancer types to reveal interactions between genetic variants (derived from germline SNP array and WES data) and immune- phenotypes derived from tumor RNAseq data. These phenotypes include immune gene expression (such as CD3E, GZMA, CXCR3 and PRF1), T-cell receptor (TCR) clonality, and antigen presenting scores (APM). We observed that stratifying patients by the prioritized pathogenic germline polymorphisms exposed distinct tumor immune landscape, implicating new prognostic factors or risk genes of cancer. In support of these findings, we show that the

top-ranked SNPs are associated with treatment responses to anti-PD-1 therapy in multiple cancers. This study creates a unique and pioneering resource of prognostic germline variants with the potential to modulate immune therapy response and cancer risk, opening new avenues for developing next generation therapeutics and for personalizing risk assessment.

120 | A Two-stage Epigenome Wide Association Study Identifies Novel Pancreatic Cancer Susceptibility Loci by Leveraging Public Controls

Ziqiao Wang[1]*, Yue Lu[2], Myriam Fornage[3], Li Jiao[4], Donghui Li[5], Peng Wei[1]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America[2]Department of Epigenetics & Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Smithville, Texas, United States of America[3]Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America[4]Department of Gastroenterology, Baylor College of Medicine, Houston, Texas, United States of America[5]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America.

We present the first epigenome-wide association study (EWAS) in whole blood of pancreatic cancer. Here we propose a novel statistical strategy by leveraging public controls in a case-control study with limited sample sizes, in order to boost the statistical power and reduce the cost of biological experiments. In our two-stage EWAS study of 44 pancreatic cancer cases/20 controls (discovery stage), and 23 cases/22 controls (validation stage) at the MD Anderson Cancer Center (MDACC), we increased the number of controls in the discovery stage from 20 to 556 by integrating public data, the Framingham Heart Study, with the MDACC controls. We successfully removed the batch effects between the two datasets from different sources, as shown by the visualization of unsupervised learning. We discovered and replicated six significantly differentially methylated CpG probes (DMPs) and three regions (DMRs) in the discovery and validation stages. Furthermore, one of the three DMRs, which is a non-coding RNA region, was replicated using another external validation prospective cohort, the Women's Health Initiative study with 13 incident cases and 26 matched controls. By performing causal inference using bidirectional Mendelian randomization analysis, we found evidence of directional relationships of the associations between DMPs and pancreatic cancer. RNA-sequencing analysis also illustrates the functional consequences of DMPs/DMRs on the cancer risk.

121 | Identification of Trans-eQTLs Using Mediation Analysis with Multiple Mediators

Nayang Shan[1], Zuoheng Wang[2]*, Lin Hou[1]

[1]Center for Statistical Science, Tsinghua University, Beijing, China; [2]Department of Biostatistics, Yale university, New Haven, Connecticutt, United States of America.

Mapping expression quantitative trait loci (eQTLs) has provided insight into gene regulation. Compared to cis-eQTLs, the regulatory mechanisms of trans-eQTLs are less known. Previous studies suggest that trans-eQTLs may regulate expression of remote genes by altering the expression of nearby genes. Trans-association has been studied in the mediation analysis with a single mediator. However, prior applications with one mediator are prone to model misspecification due to correlations between genes. Motivated from the observation that trans-eQTLs are more likely to associate with more than one cis-gene than randomly selected SNPs in the GTEx dataset, we developed a computational method to identify trans-eQTLs that are mediated by multiple mediators. We proposed two hypothesis tests for testing the total mediation effect (TME) and the component-wise mediation effects (CME). We demonstrated in simulation studies that the type I error rates were controlled in both tests despite model misspecification. Multiple mediator analysis had increased power to detect mediated trans-eQTLs, especially in large samples. In the HapMap3 data, we identified 11 more mediated trans-eQTLs in the African populations. Moreover, the mediated trans-eQTLs in the HapMap3 samples are more likely to be trait-associated SNPs. This study demonstrated that trans-eQTLs are more likely to associate with multiple cis-genes than randomly selected SNPs. Mediation analysis with multiple mediators improves power of identification of mediated trans-eQTLs, especially in large samples.

122 | Implementing Pharmacogenomics in Clinical Practice: Challenges and Realities to Managing Gene-drug Pair Information in the Electronic Medical Record

Stephen C. Waring[1]*, Bret E. Friday[2], David J. Sperl[3], Dianne L. Witten[3], Paul J. Schillo[3], Zachary T. Rivers[4], Jacob T. Brown[5], David D. Stenehjem[5]

[1]Essentia Institute of Rural Health, Duluth, Minnesota, United States of America; [2]Essentia Health Cancer Center, Duluth, Minnesota, United States of America; [3]Essentia Health Pharmacy Services, Duluth, Minnesota, United States of America; [4]University of Minnesota College of Pharmacy, Minneapolis, Minnesota, United States of America; [5]University of Minnesota Duluth, College of Pharmacy, Duluth, Minnesota, United States of America.

Preemptive pharmacogenomic testing to inform treatment for cancer, cardiac, psychiatric, and many other conditions is becoming more and more common. Advances in pharmacogenomics affords testing a broader panel of genetic variants than those of interest for a specific treatment. However, how best to manage this additional information for conditions that may not yet be present or diagnosed in an individual creates unique legal, ethical, social, financial, and practice issues that must be taken into account. Only then can truly informed practice parameters, protocols, information management systems, and educational programs meet expectations of scalability and sustainability for general practice. We are currently leading a demonstration project in a large integrated health care delivery system in the upper Midwest as part of a multi-stakeholder collaboration to implement pharmacogenomics testing statewide. Our project is being piloted in a primary care practice setting with the goal of accelerating learning to inform next steps for system-wide implementation. Critical components of this effort are 1) educational needs assessment and development of educational programs for a broad audience including patients; 2) developing an electronic health record integration prototype and clinical decision support systems to facilitate delivery and interpretation of test results; 3) determining feasibility for eventual statewide implementation; and 4) creating a foundation for pharmacoepidemiologic and clinical epidemiologic research to further refine our understanding of how to deliver the right drug at the right dose to the right individuals. This presentation will highlight barriers, methods, processes, selected drug-gene pairs, and lessons learned.

123 | Dyslexia Associated Functional Variants in Europeans are not Associated with Dyslexia in Chinese

Lingfei Liu[1], Huaiting Gu[1], Fang Hou[1], Xinyan Xie[1], Xin Li[1], Bing Zhu[2], Jiajia Zhang[3], Wen-Hua Wei[4]*, Ranran Song[1]

[1]Department of Maternal and Child Health and Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China; [2]Hangzhou Center for Disease Control and Prevention, Hangzhou, China; [3]Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, United States of America; [4]Department of Women's and Children's Health, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand.

Developmental dyslexia (DD) is a neurodevelopment disorder characterized by impaired reading acquisition in spite of adequate neurological and sensorial conditions, educational opportunities and normal intelligence. DD affects ~7% school aged kids globally with clear differences in prevalence across ethnic groups. Genome-wide association studies (GWAS) of DD often used European samples and identified only a handful associations with moderate or weak effects. This study aims to identify DD functional variants by integrating the GWAS associations with tissue-specific functional data and test the variants in a Chinese DD study cohort, Tongji Reading Environment and Dyslexia (READ). We colocalized associations from nine DD related GWAS with expression quantitative trait loci (eQTL) derived from brain tissues and identified two eSNPs rs349045 and rs201605. Both eSNPs had supportive evidence of chromatin interactions observed in human hippocampus tissues and their respective target genes *ZNF45* and *DNAH9* both had lower expression in brain tissues in schizophrenia patients than controls. In contrast, an eSNP rs4234898 previously identified based on eQTL from the lymphoblastic cell lines of dyslexic children had no chromatin interaction with its target gene *SLC2A3* in hippocampus tissues and *SLC2A3* expressed higher in the schizophrenia patients than controls. We genotyped the three eSNPs in the READ cohort of 372 cases and 354 controls and discovered only weak associations in rs201605 and rs4234898 with three DD symptoms (P *value* < 0.05). The lack of associations could be due to low power in READ but could also implicate different etiology of DD in Chinese.

124 | Incorporating Admixed Samples in Meta-analysis Methods of Genome-wide Association Studies

Emileigh L. Willems[1]*, Stephanie A. Santorico[1,2,3]

[1]Department of 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, Denver, Colorado, United States of America; [3]Department of Biostatistics & Informatics, University of Colorado Denver, Denver, Colorado, United States of America.

Meta-analysis methods are commonly used to combine summary statistics from genome-wide association studies (GWAS) to improve power by increasing effective sample size. Trans-ethnic meta-analysis methods improve upon the two standard meta-analysis methods (fixed and random effects models) by modeling correlation between studies using genetic distance. This better generalizes genetic associations across diverse populations. Previously, it has been shown that power is higher on average in admixed studies compared to non-admixed studies when the

effect size of the causal variant is similar across all populations, especially when minor allele frequencies differ. Because GWAS of admixed samples are becoming more common, we seek to find which existing meta-analysis method best incorporates information from admixed studies. We compare the following four meta-analysis methods via data simulated from 1000 Genomes Phase 3 reference data: Fixed Effects (FE), Random Effects (RE), TransMeta, and MR-MEGA. Simulation scenarios are designed to assess the Type I error, power, and localization ability of each meta-analysis method under homogeneous, independent, and trans-ethnic effects across ancestries. The trans-ethnic scenarios considered vary both the number of admixed studies and the admixed studies' ancestral populations, thus allowing a better understanding of how incorporating admixed studies affects each method. We also consider different genetic distance metrics tailored to better incorporate admixture when modeling correlation between studies in the two trans-ethnic meta-analysis methods. These simulations will be the first comparison of the two trans-ethnic methods, TransMeta (2016) and MR-MEGA (2017).

125 | Improving Power and Avoiding Pitfalls in Gene-environment Interaction Scans

Thomas W. Winkler[1]*, Felix Günther[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, Lausanne, Switzerland; [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

Previous studies demonstrated that gene-environment (GxE) interactions can successfully be identified by individual genome-wide association studies (GWAS) or meta-analyses of GWAS. However, there are various pitfalls when attempting to identify GxE interactions in GWAS. We exemplify potential pitfalls in GxE analysis in UK Biobank and GIANT consortium data. We highlight that the joint two-degree of freedom (*2DF*) test is not ideal as a test for interaction, but rather a test accounting for potential interaction. Judging on whether an observed joint effect derives from significant interaction requires testing the significant *2DF* variants for interaction in an independent sample. Second, we demonstrate that testing for interaction in GWAS should not be limited to using a genome-wide significance level (alpha = 5e-8). Instead, 2-step methods are beneficial that involve filtering variants for some statistical test result in the first step and testing filtered variants for interaction (at reduced alpha=0.05/#variants) in the second step. We compared various filtering approaches empirically and provide recommendations on choosing the optimal filter for the specific study design. Third, we show that observed GxE effects can be confounded by improper outcome transformation or model misspecification. For example, applying an inverse-normal transformation to the outcome can lead to spurious interaction results or diminish real interaction effects. We propose a sensitivity analysis method to investigate potential confounding by improper outcome transformation and provide practical solutions to avoid model misspecifications in GWAS. Our investigation illustrates potential pitfalls in GWAS screens for interactions and may guide future GxE studies.

126 | Genome-wide Meta-analysis Identifies Deletions or Excess Homozygosity Implicated in Head and Neck Cancer Susceptibility

Chih-Chieh Wu[1]*, Chien-Hsiun Chen[2], Robert Yu[3], Sanjay Shete[3]

[1]Department of Environmental and Occupational Health, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [2]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan; [3]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America.

Copy number variation (CNV) of DNA sequences involves losses (deletions) and gains (duplications) of large segments of the genome and accounts for the largest component of structural variation. CNV encompasses more polymorphic base pair than SNPs by an order of magnitude and is considered as an intermediate size. A recent meta-analysis for an updated comprehensive human CNV map that complied 23 published studies on healthy individuals across various ethnics estimates approximate 4.8–9.5% of the human genome to involve CNVs, depending on the stringency level. Here, we performed two distinct whole-genome scans of associations between deletion variants or excess homozygosity and head and neck squamous cell carcinoma (HNSCC) susceptibility and a meta-analysis that synthesizes the association summary statistics from these two whole-genome scans. We used 733,202 SNPs in 2185 patients with HNSCC from the MD Anderson Cancer Center in Houston, Texas, recruited from 1998 to 2012. We used a logistic regression framework extension of the genome-wide statistical method that we developed previously,

permitting to adjust for covariates. The two whole-genome scans of disease-associated deletions include batch 1 with 1,154 cases and 1,542 controls and batch 2 with 1,031 cases and 2,965 controls, respectively. We detected a 1.1-kb segment on chromosome 3p in batch 1 and 3 distinct 0.45–14 kb segments on chromosomes 1p, 5q, and 11q in batch 2. They were all significant at a corrected 0.05 nominal significance level, adjusted for multiple comparison procedures. The corresponding meta-analysis identified 5 distinct segments on chromosomes 5p, 6p, 9q, and 12q (2 segments).

127 | A Bayesian Method to Integrate Multi-omics Data for Disease Prediction

Xiaoxuan Xia[1,2]*, Rui Sun[1,2], Maggie Haitian Wang[1,2]

[1]The Jockey Club School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, Hong Kong SAR; [2]The Chinese University of Hong Kong Shenzhen Research Institute, Shenzhen, China.

The development of technologies allows us to collect multi-types of biological information of individuals. However, the complexity of datatypes and the large number of biological variables make it difficult to effectively integrate multi-omics data for prediction. In this study, we develop a new Bayesian model to integrate multi-types of biological data to predict disease status. For real data application, mRNA gene expression, DNA methylation, microRNA and copy number alterations data are integrated to predict breast invasive carcinoma cancer survival.

128 | Blood Lipoprotein Cholesterols Cause Coronary Artery Disease from Multivariate Mendelian Randomization Analysis

Hongyan Xu*

Department of Population Health Sciences, Augusta University, Augusta, Georgia, United States of America.

It is critical to use appropriate genetic variants as instrumental variables for valid and efficient Mendelian randomization (MR) analysis of a disease. We developed a genetic variant selection method from the empirical selection approach of Do et al. (2013). Using this method, we selected 338 single-nucleotide-polymorphisms (SNPs) from a meta-analysis data set of genome-wide association study (GWAS) of lipid and coronary heart disease (CAD) and 363 SNPs from another joint GWAS meta-analysis data. From the multivariate MR analysis, we found that low density lipoprotein cholesterols (LDL-c) was strongly associated with increasing risk for CAD, high density lipoprotein cholesterols (HDL-c) was strongly associated with decreasing risk for CAD, and triglycerides (TG) were not associated with risk for CAD. We performed a simulation study accounting for the linkage disequilibrium among the selected SNPs and the pleiotropy among lipid components. The results from simulation study confirmed the result from the multivariate MR analysis for the causal relationship of lipoprotein cholesterols and CAD.

129 | An Improved Maximum Information Coefficient Approach to Uncover Relationships of Variables in Big Datasets

Junzhao Xu[1]*, Zhaogong Zhang[1]*, Xuexia Wang[2]

[1]School of Computer Science and Technology, Heilongjiang University, Harbin, Heilongjiang, China; [2]Department of Mathematics, University of North Texas, Denton, Texas, United States of America.

By the year 2020, approximately 1.7 megabytes of new information will be created per second for every human being on the planet. The detection of the relationship among variables in a big dataset is becoming more and more common, especially in the fields of genetics or genomics. There is an urgent need to develop effective methods to uncover relationships in a big dataset. The Maximum Information Coefficient (MIC) is an effective tool for exploring such data relationships. The MIC exhausts all partitioning when meshing a pair of variables. This process is computationally complex for large datasets. We make a significant improvement for MIC by proposing a novel approximation algorithm CDMIC (Cluster Division Maximum Information Coefficient). First, we utilize a fast clustering method to produce center nodes of similar data, which represent closely related points and forms clusters. Second, we calculate each region's MIC. Finally, we estimate the sum of weighted MIC values, where the weight is estimated based on the amount of data in each region. Results from simulations and a real data reveal that the CDMIC retains the advantages of MIC and can accurately identify the existence of relevant data pairs. The CDMIC algorithm is far superior to MIC algorithm in the time-effective aspect.

130 | Adaptive Test for Meta-analysis of Rare Variant Association Studies

Tianzhong Yang[1]*, Junghi Kim[1], Chong Wu[2], Yiding Ma[3,4], Peng Wei[4], Wei Pan[1]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Statistics, Florida State University, Tallahassee, Florida, United States of America; [3]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America.

Single genome-wide studies may be underpowered to detect trait-associated rare variants with moderate or weak effect sizes. As a viable alternative, meta-analysis is widely used to increase power by combining different studies. On one hand, the statistical power of meta-analysis to detect associations critically depends on the underlying association patterns and heterogeneity levels, which are unknown and vary from locus to locus. On the other hand, existing methods mainly focus on one or only few combinations of the association pattern and heterogeneity level, thus may lose power in many situations. To address this issue, we propose a general and unified framework by combining a class of tests including and beyond some existing ones, leading to an adaptive test that achieves high power across a wide range of scenarios. Our test is applicable to summary statistics from each of the studies to be meta-analyzed, dramatically easing the data-sharing and analysis process. We demonstrate that the proposed test is more powerful than some existing methods in simulation studies, then show the performance of all methods with the NHLBI Exome-Sequencing Project (ESP) data. One novel gene was found by our proposed test, but not by others, to be statistically significantly associated with plasma triglyceride among African-ancestry subjects, whereas it was previously reported to be associated with coronary artery disease among European-ancestry subjects. We also showcase the performance of the methods in a combined analysis of the ESP exome-sequencing data and the whole genome-sequencing data of the UK10K Project.

131 | Effects of Mitochondrial DNA Variants on Blood Biomarkers

Ekaterina Yonova-Doing[1*], Claudia Calabrese[2], Lingyan Chen[1], Patrick F. Chinnery[2,3], Joanna M.M. Howson[1]

[1]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]MRC Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, United Kingdom; [3]Department of Clinical Neurosciences, Cambridge Biomedical Campus, University of Cambridge, Cambridge, United Kingdom.

Mitochondria (MT) are key players in cellular energy production and MT dysfunction leads to generation of reactive oxygen species, apoptosis, premature aging and various age-related diseases. Using UK Biobank data, we showed that mitochondrial single nucleotide variants (mtSNVs) are associated with blood cell traits (BCTs). Here, we aim to further study the role mtSNVs play in determining blood related parameters by assessing the association between mtSNVs and 34 blood biomarker in 358,916 UK Biobank unrelated individuals of European ancestry.

We found associations at MT genome-wide significance level ($P < 5 \times 10^{-5}$) as follows: 21 mtSNVs were associated with aspartate transaminase (AST, $P = 5.9 \times 10^{-15}$ at rs28358275), 8 mtSNVs with alanine amino-transferase (ALT, $P = 2.8 \times 10^{-7}$ at rs2853504), 8 mtSNV with creatinine (rs869183622 at $P = 1.1 \times 10^{-6}$), 7 mtSNVs with cystatin ($P = 2.7 \times 10^{-7}$ at rs3928306), and rs2853504 was associated with urea ($P = 3.7 \times 10^{-5}$). The associated molecules are biomarkers of liver (AST, ALT) and kidney dysfunction (creatinine, urea) or of metabolic syndrome (cystatin); all of these are processes where mitochondria play a role. When exploring additional liver and kidney related parameters, we found that of the 8 mtSNV associated with creatinine, 7 were also association with eGFR ($P < 5 \times 10^{-5}$) but we did not observe association with possible acute alcoholic hepatitis (AST/ALT > 2.0, 14,380 cases). To further understand the observed association, we cross-reference with metabolites and protein quantitative trait loci and perform Mendelian randomisation analyses using mtSNVs to assess the effect of the molecular phenotypes on the blood biomarkers.

132 | High-dimensional Regularized Regression for Identifying Gene-environment Interactions Incorporating External Information

Natalia Zemlianskaia*, Juan Pablo Lewinger, Jim Gauderman

Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America.

Reliable identification of gene-environment (GxE) interactions remains a challenging problem because it

requires very large sample sizes. Incorporating external information relevant to GxE interactions, for example, gene expression data has the potential to increase the power to detect GxE interactions, but few current methods have that capability. In addition, most approaches to detect GxE interactions to date only examine one-SNP at a time. We propose a regularized high-dimensional regression model for identifying GxE interactions, where feature selection is guided by external information. The model handles large numbers of SNPs jointly and incorporates constraints that enforce a "main effect before interaction" hierarchical structure in a convex optimization framework. Through extensive simulations, we evaluated the performance of our approach in scenarios where cis-eQTL data is informative for GxE interactions. Our results show that the model is able to select non-zero interactions with high accuracy (AUC up to 0.9) when expression data is informative and performs on par with standard feature selection methods for high-dimensional data (LASSO, hierNet) when expression data is non-informative.

133 | Genome-wide Association Study of Longitudinal Executive Functions

Bernadette Wendel[1]*, Urs Heilbronner[2], Monika Budde[2], Janos L. Kalman[2], Fanny Senner[2], Till F.M. Andlauer[3], Ashley L. Comes[2], Sergi Papiol[2], Heike Bickeböller[1], Thomas G. Schulze[2]

[1]Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Göttingen, Germany; [2]Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Germany; [3]Department of Neurology, University Hospital rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany.

Executive functions affect the control and coordination of mental processes and are often impaired in psychosis. The Verbal Digit Span Test backwards (VDS) and Trail-Making-Test Part B (TMT) are tests assessing working memory and set-shifting components of executive functions. We performed a genome-wide association study (GWAS) of the longitudinal course (LC) of VDS and TMT using linear mixed models (LMM).

The ongoing multi-center PsyCourse study investigated the LC of psychosis with measurements every six months for up to 18 months. Of 1,338 genotyped individuals, 1,080 were affective or psychotic patients and 258 controls; 1,297 probands had at least one measurement of VDS and 1,272 of TMT, 650 had three or four measurements. We considered age, gender, diagnosis, five ancestry components as covariates. PsychArray genotypes were imputed using SHAPEIT2/IMPUTE2. After quality control, only SNPs with a minor allele frequency (MAF) at least 1% were retained.

In the LMMs with outcomes VDS or log (TMT), a subject-specific time course was modeled, allowing for random intercept and slope. We included SNP, SNP × time, age, gender, diagnoses, and ancestry components as fixed effects and recruiting center as a random effect. We assessed the fixed effect of each SNP (SNP × time). For TMT, nine SNPs reached genome-wide significance, located on chromosome 5 within the same LD block ($r2 > 0.85$). For VDS, no significant SNP was found. For significant SNPs (MAF approximately 0.015) we further explored the corresponding linear mixed model.

134 | Incorporating SNP Data While Identifying Dna Methylation Changes Associated with Disease

Yixiao Zeng[1,2]*, Yi Yang[2], Celia Greenwood[1,3,4,5,6]

[1]Department of Quantitative Life Science, McGill University, Montreal, Canada; [2]Department of Mathematics & Statistics, McGill University, Montreal, Canada; [3]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada; [5]Department of Human Genetics, McGill University, Montreal, Canada; [6]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada.

Introduction: DNA methylation can be measured with a targeted custom capture bisulfite sequencing technique that captures regions of potential interest along with SNP data in the same regions. When testing for association between the methylation levels and a phenotype, bisulfite sequencing data frequently exhibits over-dispersion when using binomial regression models. This may arise as a consequence of DNA taken from mixed cell types in each sample, or if there is inadequate capture of the effects of SNPs on methylation. In this work, we particularly focus on how best to adjust for genetic effects, which can be challenging because there may be thousands of SNPs even in a small genomic region.

Methods: We implement penalized regression models incorporating SNP data into analysis of association between DNA methylation and a disease relevant phenotype. In simulations, we investigate the impact of incorporating genetic data on over-dispersion, type I error, and power. We also compare the behavior of different penalties and the performance associated with several transformations of the methylation data.

Results: Preliminary results indicate substantial reductions in over-dispersion and improved type I error associated with adjustment for genetic effects. Power varies across simulated scenarios, however, a square root arcsine transformation of the methylation proportion

values combined with a lasso penalty on the coefficients of genetic covariates shows promise.

Discussion: Adjusting for genetic effects may improve accuracy of results in analyzing bisulfite sequencing DNA methylation data. Explicit considerations of mediation or directions of causal action may also prove fruitful.

135 | Semiparametric Accelerated Failure Time Mixture Cure Model for Clustered Data

Dongfang Zhang*, Min Chen

Department of Mathematical Science, University of Texas at Dallas, Richardson, Texas, United States of America.

We propose a new estimation method for a semiparametric accelerated failure time (AFT) mixture cure model for clustered data. We consider a mixture population of two distinct subpopulations, the one of cured or long-term survivors who will not experience the event, and the other of susceptible subjects who will experience the event subject to right censoring. Further, the data in our study are clustered or correlated because some variables are shared by multiple observations, like common genetic or environmental factors. The proposed methods estimate the within-cluster dependency using generalized estimating equation, which uses a working covariance structure to specify the dependency. Large scale simulations are conducted to investigate the properties of the proposed estimators. The simulation results show higher efficiency could be achieved when the within-cluster dependence is strong compared with conventional mixture cure model ignoring the structure of clustering. The proposed method is applied to an experiment data with repeated measures of mouse behavioral changes after gene knockouts in a study of Alzheimer's disease.

136 | A Gene Based Association Test Utilizing an Optimally Weighted Combination of Multiple Traits

Jianjun Zhang[1]*, Qiuying Sha[2], Xuexia Wang[1]

[1]Department of Mathematics, University of North Texas, Denton, United States of America; [2]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America.

Pleiotropy is a widespread phenomenon in complex diseases for which multiple correlated traits are often measured. Existing methods for multiple traits association tests usually study each of the multiple traits separately and then combine the univariate test statistics or combine *P* values of the univariate tests for identifying disease associated genetic variants. However, ignoring correlation between phenotypes may cause power loss. Additionally, genetic variants in one gene are often viewed as a unit affecting the underlying disease because the basic functional unit of inheritance is a gene rather than a genetic variant. Thus, results from gene level association test can be more readily integrated with downstream functional and pathogenic investigation, whereas many existing methods for multiple trait association tests only focus on testing a single common variant rather than a gene. We propose a statistical method by Testing an Optimally Weighted Combination of Multiple traits (TOW-CM) to test association between multiple traits and a gene or a pathway. Extensive simulation studies demonstrate that the proposed method has controlled type I error very well and is either the most powerful test or comparable with the most powerful test. In addition, we illustrate the usefulness of TOW-CM by analyzing a whole-genome genotyping data from a COPDGene study.

137 | Modelling Covariate Effects in Bisulfite Sequencing-derived Measures of DNA Methylation, in the Presence of Overdispersion

Kaiqiong Zhao[1,2]*, Karim Oualkacha[3], Aurélie Labbe[4], Lajmi Lakhal-Chaieb[5], Celia Greenwood[1,2,6,7]

[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]Département de Mathématiques, Université du Québec à Montréal, Montreal, Canada; [4]Département des Sciences de la Décision, HEC Montreal, Montreal, Canada; [5]Département de Mathématiques et de Statistique, Université Laval, Quebec City, Canada; [6]Department of Human Genetics, McGill University, Montreal, Canada; [7]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada.

Introduction: Identifying disease-associated changes in DNA methylation can help us gain a better understanding of the underlying biological determinants of diseases. Bisulfite sequencing technology provides a powerful tool for measuring large-scale methylation at single nucleotide resolution of DNA. We have developed a method for estimating smooth covariate effects and identifying differentially methylated regions (DMRs) from Bisulfite sequencing data, which copes with experimental errors and variable read depths (https://github.com/GreenwoodLab/SOMNiBUS/). This method utilizes the binomial distribution to characterize the variability in the methylated counts at each site. However, bisulfite sequencing data frequently includes many low-count integers and can exhibit over or under dispersion relative to the binomial distribution. For example, overdispersion

can arise when the reads that are aligned to the same position come from cells with different cell-types.

Methods & Results: Using a small subsample from the CARTaGENE cohort in Quebec, we investigate the impact of dispersion on the discovery of DMRs and show that the inference for both univariate and regional differential methylation can be biased in the presence of dispersion. We therefore propose a quasi-likelihood-based regional testing approach which explicitly allows for varying strengths of dispersion across a region and provides correct inference for smooth covariate effects. We will apply this method to analyze genome-wide targeted bisulfite sequencing data on individuals sampled from the CARTaGENE cohort.

Conclusions: Overdispersion can lead to substantial bias, but smooth corrections for variable dispersion can provide accurate inference.

138 | Estimated Total Mediation Effects for Multiple Types of High-dimensional Omics in Over 3500 Individuals on Aging-related Variation in Blood Pressure

Yujie Zhao[1]*, Tianzhong Yang[2], Jinhao Zou[1], Ziqiao Wang[1], Jingbo Niu[3], Han Chen[4], Peng Wei[5]

[1]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, Texas, United States of America; [2]Division of Biostatistics, The University of Minnesota, Minneapolis, Minnesota, United States of America; [3]Section of Nephrology, Baylor College of Medicine, Houston, Texas, United States of America; [4]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [5]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America.

Environmental exposures can regulate intermediate molecular phenotypes, such as the methylome, transcriptome and metabolome, by different mechanisms and thereby lead to different health outcomes. It is of significant scientific interest to unravel the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposure and traits. Mediation analysis is an important tool for investigating such relationships. However, it has mainly focused on low-dimensional settings and there is a lack of a good measure of the total mediation effect. Here, we extend an R-squared (Rsq) effect size measure, originally proposed in the single-mediator setting, to the moderate- and high-dimensional mediator settings in the mixed model framework. Using extensive simulations we demonstrate appealing operating characteristics of the proposed Rsq measure of total mediation effect and the estimation procedure. By applying the proposed estimation procedure to the Framingham Heart Study (FHS) of 1,655 individuals, we found that 82% (95% confidence interval (CI) = [54%, 100%]), 55% ([33%, 90%]) and 41% ([10%, 83%]) of the aging-related variation in systolic blood pressure can be explained by the methylomics, mRNA expression and metabolomics profile, respectively, whereas the microRNA expression profile was not found to be a significant mediation mechanism between aging and blood pressure. Furthermore, these findings in the FHS were replicated in the Women's Health Initiative (WHI) study of 1,867 individuals. Finally, we have developed an R package "RsqMed" to implement the proposed novel mediation measure and estimation procedure.

139 | Comparison and Evaluation of Pathway and Gene-level Methods for Cancer Prognosis Prediction

Xingyu Zheng[1], Christopher I. Amos[1,2], H. Robert Frost[1]

[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America; [2]Department of Medicine, Baylor College of Medicine, Houston, Texas, United States of America.

Cancer prognosis prediction has become an important research goal. A limited number of gene-level analyses have led to clinically useful methods like Oncotype DX but these remain very difficult to develop and implement. A promising direction for improving the performance and interpretation of expression-based predictive models involves aggregating gene-level data into biological pathways. Although a few studies have used pathway-level predictors, a comprehensive comparison of pathway-level and gene-level prognostic models has not been performed.

To address this gap, we characterized the performances of penalized Cox proportional hazard models built using either pathway or gene-level predictors for the cancers profiled in The Cancer Genome Atlas (TCGA) and pathways from the Molecular Signatures Database (MSigDB). When analyzing the TCGA data, we found that pathway-level models are more parsimonious and easier to interpret than the gene-level models without a loss of predictive performance. For example, both pathway and gene-level models have an average Cox concordance index of 0.85 for the TCGA glioma cohort, however, the gene-level model has twice as many predictors on average and the predictor composition is less stable across cross-validation evaluations. In simulations, when the correlation structure of the real data is broken, the pathway-level models have greater predictive performance and superior interpretative power relative to the gene-level models. For example, the average concordance index of the pathway-level model is 0.88 while the gene-level model falls to 0.56 for the TCGA glioma cohort.

Invited Abstract

140 | Empirical Bayes Methods for Genetic Risk Prediction

Hongyu Zhao*

Department of Biostatistics, Yale University, New Haven, Connecticutt, United States of America.

Genetic risk prediction is an important problem in human genetics, and accurate prediction can facilitate disease prevention, diagnosis, and treatment. Calculating polygenic risk score (PRS) has become widely used due to its simplicity and effectiveness, where only summary statistics from genome-wide association studies are needed. Recently, several methods have been proposed to improve standard PRS by utilizing external information, such as linkage disequilibrium (LD) and functional annotations. In this presentation, we introduce empirical Bayes methods that leverage information from effect sizes, LD and other external sources to improve prediction accuracy. Compared to most existing genetic risk prediction methods, our methods do not need to tune parameters, and are computationally efficient. We demonstrate the effectiveness of our methods through their applications to a number of complex diseases in large population cohorts. This is joint work with Wei Jiang, Shuang Song, Yixuan Ye, Geyu Zhou, and others.

141 | Association Between Alzheimer's Disease Risk SNPs and Episodic Memory in South Asians from the LASI-DAD Study

Jennifer A. Smith[1], Wei Zhao[1]*, Miao Yu[1], Priya Moorjani[2], Andrea Ganna[3], A.B. Dey[4], Sharon L.R. Kardia[1], Jinkook Lee[5]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America; [2]Department of Molecular and Cell Biology, University of California, Berkeley, California, United States of America; [3]Institute for Molecular Medicine Finland, Helsinki, Finland; [4]Department of Geriatric Medicine, All India Institute of Medical Sciences, New Delhi, India; [5]Department of Economics, University of Southern California, Los Angeles, California, United States of America.

Genetic factors play an important role in Alzheimer's disease (AD) and cognitive aging. However, the majority of work in this area has been conducted in populations of European ancestry (EA), and it is unclear whether the identified risk loci are associated with AD or cognition in South Asians. We investigated the allelic distribution of 55 known AD risk SNPs identified from three EA GWAS in 937 South Asians from the Diagnostic Assessment of Dementia for the Longitudinal Aging Study of India (LASI-DAD). Participants were genotyped using the Illumina Global Screening Array and imputed to 1000 G Phase 3v5. We next assessed the association between each AD risk SNP, as well as genetic risk scores (GRS) created from each of the EA GWAS, with two episodic memory scores (total and delayed word recall) after controlling for age, gender, and population structure. The 55 AD risk alleles were at lower frequency, on average, in the LASI-DAD population. Although only a handful of SNPs were significantly associated with memory scores ($P < 0.05$), effect estimates from the AD GWAS and the LASI-DAD cognitive association tests showed moderate correlations (0.24 to 0.48) in the expected directions. GRSs were associated with each memory score in the expected direction, although percent variation explained was small (0.2–0.5%). Additional adjustment for educational attainment did not substantively change the results. This work justifies the need for a more comprehensive assessment of the genetic factors associated with AD and cognition in South Asian populations.

142 | Transcriptome-wide Association Study Identifies Novel Candidate Genes Associated with Osteoporosis

Muchun Zhu*, Peng Yin

Center for Biomedical Information Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

Osteoporosis (OP) is a polygenetic disease which is usually characterized by low bone mineral density (BMD). GWAS studies have identified hundreds of genetic loci associated with BMD. However, the causality of these loci remains elusive. To identify causal genes of the associated loci, we applied a Transcriptome-Wide Association Study (TWAS) that directly imputes gene expression effects from GWAS summary data, using a statistical prediction model trained on GTEx reference transcriptome data, restricting at OP, blood and muscles-skeletal tissues data. Preliminary analysis identified 279 TWAS-significant genes at a Bonferroni-corrected threshold of $P < 3.7E\text{-}6$ and colocalization probabilities of $PP3 > 0.8$ or $PP4 > 0.9$ using the COLOC method, including 210 protein-coding genes and 69 lncRNAs. Among them, 209 transcripts are located more than 500 kb away from known OP GWAS hits, representing potential novel causal genes. They were also validated by VarElect tool, indicating 80% of them are likely to be directly or indirectly involved in OP. Several novel candidate genes were enriched for differential expression genes in osteoblasts cells expression profiles (GSE35956

and GSE35959) from GEO database, including *IBSP*, affecting calcium and hydroxyapatite binding, and *CD44*, regulating alternative splicing of gene transcription, and *SPTBN1*, interacting with calmodulin in a calcium-dependent manner and candidate. PPI and pathway enrichment analysis detected several OP-associated pathways, including MAPK signaling pathway and Osteoclast differentiation, B cell receptor signaling pathway. Our findings will provide additional insight into the development of novel targeted therapeutics to treat OP and reduce the risk of fracture.

143 | Differentiate Horizontal Pleiotropy from Mediation Using GWAS Summary Statistics in Combining Mendelian Randomization Analysis

Xiaoyin Li, Xumin Ni, Xiaofeng Zhu

Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America.

The overall association evidence of a genetic variant with multiple traits can be evaluated by cross phenotype association analysis using summary statistics from individual trait GWAS. Dissecting the association pathways from a variant to multiple traits is extremely important but has not been well studied. In this study, we introduce a computationally efficient iterative approach using summary statistics from GWAS to differentiate horizontal pleiotropy from mediation in combining with Mendelian randomization analysis. Our extensive simulations suggest that the proposed iterative method has similar performance to the widely used MR-PRESSO for two-sample MR analysis but is substantially improved when using overlapped samples. Furthermore, our approach is computationally much faster than MR-PRESSO. Similar to ME-PRESSO, our proposed method leads unbiased estimates of causal effects when horizontal pleiotropy occurs in less than 50% instrumental variables. We applied our proposed method to the summary statistics from blood pressure (BP) and coronary artery disease (CAD) GWAS and detected multiple pleiotropy variants and significant causal relationships between BP and CAD.