DOI: 10.1002/gepi.22163

ABSTRACTS

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

The 2018 Annual Meeting of the International Genetic Epidemiology Society

1 | Comparison of normalization methods for RNA-Seq data

Farnoosh A. Aghababazadeh[1], Qian Li[1,2], Brooke L. Fridley[1]

[1]Department of Biostatistics & Bioinformatics, The Moffitt Cancer Center, Tampa, United States of America; [2]Health Informatics Institute, University of South Florida, Tampa, United States of America

Normalization of RNA-Seq data is essential to ensure accurate statistical inferences. The goal of this study was to assess the various methods for the normalization of RNA-Seq, including the assessment of known and unknown technical artifacts. Additionally, we assessed the impact of the reduction in degrees of freedom (DF) due to the normalization for known and latent technical artifacts which results in inflated type I error rates in the identification the differentially expressed (DE) genes. Specifically, we compared the remove unwanted variation (RUV), surrogate variable analysis (SVA "BE" and SVA "Leek"), and principal component analysis (PCA) methods for RNA-Seq data using both simulated and publicly available data from The Cancer Genome Atlas (TGCA) cervical cancer study (CESC). Using CESC data, a comparison between the top estimated latent factors between the methods showed that the SVA ("Leek") and RUV estimates were highly correlated ($r = -0.88$ and $p < 2.2e\text{-}16$), and the RUV and PCA methods produced more similar DE results compared to SVA ("Leek"). The simulation study found that the permutation procedure in SVA ("BE") to determine the number of significant surrogate variables outperforms other methods for correctly estimating the number of technical artifacts. As expected, ignoring the loss of DF due to normalization results in an inflated type I error rates. Hence, we recommend to include not only library size corrections but also the assessment of known and unknown technical artifacts, and if needed, across sample normalization. Lastly, we recommend that the known technical artifacts along with the primary factors of interest in the design matrix, as opposed to differential analysis on a post-normalized data set.

2 | Polygenicity of idiopathic pulmonary fibrosis

Richard J. Allen[1], Justin M. Oldham[2], Tasha E. Fingerlin[3,4], Rebecca Braybrooke[5,6], UK ILD Consortium, Carlos Flores[7,8,9], Imre Noth[10], David A. Schwartz[3,11,12], R. Gisli Jenkins[5,13], Martin D. Tobin[1,14], Louise V. Wain[1,14]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Department of Internal Medicine, University of California Davis, Davis, United States of America; [3]Center for Genes, Environment and Health, National Jewish Health, Denver, United States of America; [4]Department of Biostatistics and Informatics, University of Colorado, Denver, United States of America; [5]NIHR, Nottingham BRC, Nottingham, United Kingdom; [6]Division of Epidemiology and Public Health, University of Nottingham, Nottingham, United Kingdom; [7]Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [8]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain; [9]ITER, Santa Cruz de Tenerife, Spain; [10] Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, United States of America; [11]Department of Medicine, University of Colorado, Denver, United States of America; [12]Department of Immunology, University of Colorado, Denver, United States of America; [13]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [14]NIHR, Leicester Respiratory BRC, Leicester, United Kingdom

Idiopathic pulmonary fibrosis (IPF) is a rare lung disease of unknown cause with poor prognosis. Genome-wide association studies have identified 17 variants associated with IPF susceptibility. The majority of these have moderate effect sizes, however, one variant (rs35705950 in the *MUC5B* promoter) has a very large effect (OR ≈ 5 in Europeans). We aimed to identify whether genetic risk of IPF is explained by a small number of variants, or whether IPF is polygenic with many tens or hundreds of signals yet to be detected.

Using weights from a genome-wide meta-analysis of 1,153 IPF cases and 3,908 controls, risk scores were generated for individuals in an independent cohort (1,616 IPF cases and 4,863 controls). *P* value thresholds were used to vary the number of variants included in the score from including just the *MUC5B* variant to including all variants genome-wide.

The strongest association with IPF was observed when just including rs35705950 in the risk score ($P = 5.83 \times 10^{-121}$, Nagelkerke's $R^2 = 0.126$). As more

variants were included in the score, the significance initially decreased before increasing again and plateauing with $P \approx 10^{-40}$ and Nagelkerke's $R^2 \approx 0.04$. Crucially, after excluding the most strongly associated variants by only including variants with $P > 0.05$, the score still showed an association with IPF.

Few variants have been reported as associated with IPF though previous analyses have comprised of small sample sizes. Our results suggest that IPF is highly polygenic (albeit with a subset of variants showing larger effect sizes), supporting the need for larger genome-wide analyses to be conducted.

3 | Making the most of exome sequencing data from family trios with probands affected by very rare birth defects

Lynn M. Almli[1,2], Jessica X. Chong[3], Elizabeth Blue[3], Stuart K. Shapira[2], Faith Pangilinan[4], Jennita Reefhuis[2], James C. Mullikin[5], Deborah A. Nickerson[3], Lawrence C. Brody[4], Michael J. Bamshad[3], Mary M. Jenkins[2], NIH Intramural Sequencing Center, University of Washington Center for Mendelian Genomics, and the National Birth Defects Prevention Study

[1]Carter Consulting Inc., Atlanta, United States of America; [2]National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, United States of America; [3]University of Washington, Seattle, United States of America; [4]Gene and Environment Interaction Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, United States of America; [5]Comparative Genomics Analysis Unit, National Human Genome Research Institute, National Institutes of Health, Rockville, United States of America

Structural birth defects are a heterogeneous group of disorders, the etiologies of which remain largely unknown. For the last 20 years, the Centers for Disease Control and Prevention has conducted population-based case-control studies to identify risk factors for birth defects in the United States, including the National Birth Defects Prevention Study (NBDPS). To discover genes that influence birth defect risk, exome sequencing is being carried out on NBDPS DNA samples from parents and their affected children born between 1997 and 2011. Finding causal variants will be challenging as the sample is ethnically diverse, and the biological bases of these defects are unknown. We present preliminary results from a pilot study of two very rare intestinal anomalies; DNA samples from nine children with colonic and four children with multiple intestinal atresia, and their parents were sequenced at the National Institutes of Health Intramural Sequencing Center. Whole exome sequence data were analyzed using a pipeline developed at the University of Washington's Center for Mendelian Genomics. After sample- and variant-level quality control, the observed genetic variation among seven colonic (three non-Hispanic white and four Hispanic) and two multiple intestinal (both Hispanic) atresia cases were analyzed under different inheritance models. While several putatively deleterious genetic variants were identified, no families shared candidate variants in the same gene. Nonetheless, sequencing these samples from NBDPS trios will create a unique birth defect research resource as a large amount of additional data are available from these families (e.g., pregnancy exposure information, birth defects clinical information, and family histories).

4 | The genotype by environment interaction (G×E) in inbred lines and hybrids of maize

David Almorza[1], María V. Kandús[2], Juan C. Salerno[2], Arturo Prada[3]

[1]Department of Statistics and Operational Research, University of Cádiz, Cádiz, Spain; [2]INTA-Instituto de Genética, Hurlingham, Argentina; [3]Department of Human Anatomy and Embryology, University of Cádiz, Cádiz, Spain

Maize breeding is based on the production of single hybrids using inbred lines of low performance and selecting the best hybrid combinations. The objectives of this study was to assess the genotype by environment interaction (G×E) in inbreds and hybrids.

The single hybrids and the parental inbreds were tested in comparative yield trials in three locations during two years and we discovered that the non-additive and environmental effects had a greater influence on yield. In relation to G×E interaction, stable lines were detected, whereas in the inbreds, these stable lines were poorly adapted. No defined trend was found between the hybrid type and its differential behavior (stability or specific adaptation).

In a previous paper, we presented that the additive main effects and multiplicative interaction (AMMI) model is the most convenient technique in the analysis of multi-environmental testing. However, the generalized estimating equation (GEE) model facilitates the visualization and the identification of genotypes in each environment, and this should be taken into account when selecting the model to use.

The AMMI model combines analysis of variance (ANOVA) for the main effects of genotype and environment and only uses principal component analysis (PCA) for interaction GxE.

In This study, we proved that stable hybrids and others with high specific adaptation were identified with AMMI and site regression (SREG) models, although AMMI showed a more defined pattern of the GxE interaction than SREG.

5 | Addressing the missing data issue in multi-phenotype genome-wide association studies

Mila D. Anasanti[1], Marika Kaakinen[1,2], Marjo-Riitta Jarvelin[3,4,5,6,7], Inga Prokopenko[1]

[1]Department of Genomics and Common Disease, Imperial College London, London, United Kingdom; [2]Department of Medicine, Division of Experimental Medicine and Toxicology, Imperial College London, London, United Kingdom; [3]Department of Epidemiology and Biostatistics, Imperial College London, 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) suffers from missingness in phenotype values, leading to inefficiency of the standardly implemented complete case (CC) analysis. We investigated properties of missing data imputation techniques and compared them with the full data and CC analyses. We used Bayesian method single and multiple-imputation (SI/MI), 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 (three/nine/30/120) for these individuals in R. We randomly selected common, low-frequency and rare variants to be significantly ($P < 5 \times 10^{-8}$) associated with simulated phenotypes. We considered different proportions of missing data (1/5/20/50%) under three mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The Root Mean Squared Errors (RMSE) of coefficient regression from the analyses regressing each SNP on the linear combination of the phenotypes, after applying the selected missing data methods, were compared to the values from the full data analysis. RF and MI from the simulation studies outperformed and were implemented into SCOPA for the Northern Finland Birth Cohorts (NFBC1966, $N = 4955$ and NFBC1986, $N = 2687$) for anthropometric, fasting blood glucose/insulin measurements and 149 serum metabolite levels. We observed a number of additional loci identified in MP-GWAS after RF and MI phenotype imputation; more precise, on average, parameter estimates for significant association and compared to CC analyses. We proposed new user-friendly high-performing solutions for phenotype imputation in highly-dimensional omics data analyses.

6 | Genetic architecture of gene expression traits across diverse populations

Angela Andaleon[1,2], Lauren S. Mogil[1], Alexa Badalamenti[2], Scott P. Dickerson[3], Xiuqing Guo[4], Jerome I. Rotter[4], W. Craig Johnson[5], Hae Kyung Im[3], Yongmei Liu[6], Heather E. Wheeler[1,2,7,8]

[1]Department of Biology, Loyola University Chicago, Chicago, United States of America; [2]Program in Bioinformatics, Loyola University Chicago, Chicago, United States of America; [3]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, United States of America; [4]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics at Harbor-University of California Los Angeles Medical Center, Torrance, United States of America; [5]Department of Biostatistics, University of Washington, Seattle, United States of America; [6]Department of Epidemiology & Prevention, Wake Forest School of Medicine, Winston-Salem, United States of America; [7]Department of Computer Science, Loyola University Chicago, Chicago, United States of America; [8]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, United States of America

For many complex traits, gene regulation is likely to play a crucial mechanistic role. How the genetic architectures of complex traits vary between populations and subsequent effects on genetic prediction are not well understood, in part due to the historical paucity of Genome-wide Association studies (GWAS) in populations of non-European ancestry. We used genotype and monocyte expression data from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort, including African American (AFA, $n = 233$), Hispanic (HIS, $n = 352$), and European (CAU, $n = 578$), to characterize the genetic architecture of gene expression within and between diverse populations. We performed expression quantitative trait loci (eQTL) mapping in each population and show genetic correlation of gene expression depends on shared ancestry proportions. Using elastic net modeling with cross validation to optimize genotypic predictors of gene expression in each population, we show sparse genetic architecture of gene expression for most predictable genes. We found the best predicted gene, *TACSTD2*, was similar across populations with $R^2 > 0.86$ in each population. However, there are some genes that are well-predicted in one population but poorly predicted in another due to allele frequency differences between populations. Using genotype weights trained in MESA to predict gene expression in independent populations showed that a training set with ancestry similar to the test set is better at predicting gene expression in test populations, demonstrating an urgent need for diverse population sampling in genomics. Our predictive models in diverse cohorts are made publicly available for use in transcriptome mapping methods at http://predictdb.hakyimlab.org/.

7 | Genetics of the measure of physiological dysregulation: findings from the health and retirement study

Konstantin G. Arbeev[1], Olivia Bagley[1], Hongzhe Duan[1], Arseniy P. Yashkin[1], Alexander M. Kulminski[1], Irina V. Culminskaya[1], Svetlana V. Ukraintseva[1], Anatoliy I. Yashin[1]

[1]Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, United States of America

The statistical (Mahalanobis) distance measure (denoted as D_M here) was recently suggested in the literature for evaluating the level of "physiological dysregulation" in aging body based on measuring deviations of multiple biomarkers from baseline physiological state. This composite measure allows reducing high-dimensional biomarker space into a univariate measure which summarizes information about dynamics of various biomarkers. In this study, we constructed a D_M using measurements of multiple biomarkers from the Venous Blood Study (VBS) in 4,231 participants of European ancestry from the Health and Retirement Study (age in VBS: 40–100, 59% women). We selected biomarkers from VBS which were moderately correlated with age (absolute values of correlation exceeding 0.05) and constructed the D_M using sex-specific means and variance-covariance matrices for individuals younger than 60 years at the time of biomarker measurements in VBS, considered as the "reference" population (from which the distance is then computed). We performed a genome-wide association study of the D_M using a linear regression adjusted for sex, age in VBS, smoking, education, fasting, and principal components. One SNP (rs4446183 on chromosome 3) reached the genome-wide significance level ($P = 2.3E\text{-}8$) and several other SNPs on chromosomes 2 and 3 indicated a suggestive signal ($P < 1E\text{-}6$). We also investigated polygenic effects on this trait computing multiple polygenic risk scores from SNPs passing different P value thresholds. The results of this study illustrated possible genetic determinants of D_M and indicated that this composite and polygenic trait is associated with different aging-related outcomes.

8 | Amish families give evidence for rare variants linked to myopia

Joan E. Bailey-Wilson[1], Anthony M. Musolf[1], Claire L. Simpson[1,2], Laura Portas[3], Federico Murgia[3], Qing Li[1], Elise Ciner[4], Dwight Stambolian[5]

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

Myopia is a complex ocular disorder that affects about 25% of Americans. We performed linkage of myopia with exome enriched array genotypes for 349 patients from 43 extended Amish families. Affection status was based on mean spherical equivalent in Diopters (D): affected ($<= -1$D), unaffected ($>= 0$D) or unknown (<0D, > -1D). Two discrete types of two-point parametric linkage analysis were performed: variant-based and gene-based. Variant-based analysis tested for linkage between the phenotype and individual SNPs using TWOPOINTLODS while gene-based analysis tested for linkage between the phenotype and haplotypes of rare variants located within a particular gene (SEQLINKAGE and MERLIN). We assumed an autosomal dominant model with a disease allele frequency of 1% and a 90% penetrance for carriers and 10% phenocopy rate.

Both the variant-based and gene-based tests identified chromosomal loci that were highly suggestive or significant. Each linkage was unique to a family and many of the linkages were present along long linked haplotypes. The most interesting linked haplotypes were at 1p36.22-32.2, 12q14.2−21.1, 3p14.1-q13.13, and 4p14-q12. The signal at 5p was genome-wide significant (LOD = 3.3 in a single family). Two linkages replicate other published linkages for *MYP14* and *MYP3* at 1p36 and 12q21. All of the linked haplotypes contained rare variant(s) (minor allele frequency $<= 0.05$) that were part of the linkage.

These regions all contain good candidate genes, such as GUCA1C (3q13), which is known to be expressed in retinal photoreceptors. Targeted sequencing is planned on the linked haplotypes to elucidate the causal variant(s).

9 | Clinical and genetical study of dystrophinopathies in the Teaching Hospital Point G (Bamako, Mali)

Alassane B Maiga[1], Amadou Touré [2], Lassana Cissé[1], Seybou H Diallo[2], Salimata Diarra[1], Abdoulaye Yalcouyé[1], Mohamed Emile Dembélé[1], Kenneth Fischbeck[3], Cheick O Guinto[1], Guida Landouré[1,3]

[1]Service de Neurologie, CHU du Point « G », Bamako, Mali; [2]Service de Neurologie, CHU de Gabriel Touré, Bamako, Mali; [3]Neurogenetics Branch, NINDS, NIH, Bethesda, United States of America

Human dystrophinopathies are X-linked genetic disorders characterized by impairment in the function of dystrophin in a wide range of tissues. Although the clinical expression of dystrophinopathies and their evolutionary profile are

often evocative, their identification on molecular bases remains indispensable in our context for an accurate diagnosis and better management of such areas as genetic counseling and prenatal diagnosis. This is the reason why there is a need to carry out this clinical-genetic study on a larger sample To define the still poorly understood spectrum of this pathology in Mali in particular, and in the West African sub-region in general.

Patients with dystrophinopathies phenotype were seen and enrolled after giving their consent. In this study, physical exams were performed and blood samples were taken for genetic testing in seven families, totaling 11 patients (all male) with dystrophinopathies phenotype.

Seven families totaling 11 patients with dystrophinopathies phenotype (all male) were enrolled and of the 11 patients, 10 had Duchenne Muscular Dystrophy phenotype and one had Becker Muscular Dystrophy phenotype. The mean age of onset was 4.7 years old. The pattern of X-linked recessive inheritance was suspected in 63.6% of our patients who had similar case histories in their family history and the other cases were sporadic. Neurological examination has found that in all patients, proximal motor deficit of four limbs, calves hypertrophy and skeletal deformities were present and genetic testing also found Exon deletion in the *DMD* gene in two patients. Testing is still in progress for the remaining 9 patients.

Our study tends to confirm what is currently found in scientific literature regarding clinical and genetic patterns of dystrophinopathies and our study may be one of the rare studies in sub-Saharan Africa treating this subject. It is therefore important that further studies are conducted to better understand the spectrum of these poorly understood diseases.

10 | Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic disease

Robin N. Beaumont[1], Nicole M. Warrington[2], Momoko Horikoshi[3,4,5], Felix R. Day[6], The Early Growth Genetics (EGG) Consortium, Ken K. Ong[6,7], Mark I. McCarthy[4,5,8], John R.B. Perry[6], Rachel M. Freathy[1,9], David M. Evans[2,9,10]

[1]Institute of Biomedical and Clinical Science, University of Exeter Medical School, University of Exeter, Royal Devon and Exeter Hospital, Exeter, United Kingdom; [2]University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Australia; [3]RIKEN, Centre for Integrative Medical Sciences, Laboratory for Endocrinology, Metabolism and Kidney diseases, Yokohama, Japan; [4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [5]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom; [6]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom; [7]Department of Paediatrics, University of Cambridge, Cambridge, United Kingdom; [8]Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom; [9]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom; [10]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Individual variation in birth weight (BW) is associated with future cardio-metabolic health outcomes. These associations have generally been assumed to reflect the lifelong consequences of an adverse intrauterine environment. In earlier work, we demonstrated that much of the inverse correlation between BW and adult metabolic traits could instead be attributable to shared genetic effects. However, that work and other previous studies did not systematically distinguish the direct effects of an individual's own genotype on BW and subsequent disease risk from indirect effects of their mother's correlated genotype, mediated by the intrauterine environment. Here, we describe greatly expanded genome-wide association analyses of own BW ($n = 321{,}223$) and offspring BW ($n = 230{,}069$ mothers) which identified 305 association signals influencing BW (including >200 novel associations). We used structural equation modelling to decompose the contributions of direct fetal and indirect maternal genetic influences on BW, implicating fetal- and maternal-specific mechanisms and tissues. We then used Mendelian randomization to explore the causal relationships between factors influencing BW through fetal- or maternal routes and cardio-metabolic disease. Our results demonstrate that direct fetal genotype effects dominate the shared genetic contribution to lower BW and higher Type 2 Diabetes-risk. The relationship between lower BW and later blood pressure (BP) is instead driven by a combination of maternal and offspring genetic effects: indirect effects of maternal BP-raising genotypes act to reduce BW, but only direct effects of offspring BP-raising genotypes increase later offspring BP. This study establishes the contribution of genetic effects to observed correlations between BW and subsequent cardio-metabolic disease.

11 | Efficient fine-mapping to identify causal genetic variants and quantify their contribution to complex phenotypes

Christian Benner[1,2], Aki S. Havulinna[1,3], Veikko Salomaa[3], Samuli Ripatti[1,2,4], Matti Pirinen,[1,2,5]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; [2]Department of Public Health, University of Helsinki, Helsinki, Finland; [3]National Institute for Health and Welfare, Helsinki, Finland; [4]Broad Institute of MIT and Harvard, Cambridge,

United States of America; [5]Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Recent statistical approaches have shown that the set of all available genetic variants explains considerably more phenotypic variance of complex traits and diseases than the individual variants that are robustly associated with these phenotypes. However, an ultimate goal in genetics research is to narrow down the polygenic regional heritability into causal variants contributing to heritability. Because rapidly increasing sample sizes constantly improve fine-mapping of causal variants, it is useful to routinely estimate how much phenotypic variance the detected causal variants explain for each region. We have therefore extended the FINEMAP software to estimate the effect sizes and regional heritability under the probabilistic model that assumes a handful of causal variants per each region. Using the UK Biobank data to simulate genomic regions with only a few causal variants, we demonstrate that FINEMAP provides higher precision and enables more detailed decomposition of regional heritability into causal variants than the variance component model implemented in BOLT or the fixed-effect model implemented in HESS. Using data from 51 serum biomarkers and four lipid traits from the FINRISK study, we estimate that FINEMAP captures on average 24% more regional heritability than the variant with the lowest *P* value alone and 20% less than BOLT. Our simulations suggest how an upward bias of BOLT and a downward bias of FINEMAP could together explain the observed difference between the methods. We conclude that FINEMAP provides a computationally efficient framework to deduce a variant-level picture of the regional genetic architecture in the era of biobank scale data.

12 | Persistent organochlorine pollutants: genetic variations associated with p,p'-DDE and PCB153 blood levels among women in France

Takiy Berrandou[1], Emilie Cordina-Duverger[1], Claire Mulot[2], Thérèse Truong[1], Pascal Guénel[1]

[1]INSERM, Center for Research in Epidemiology and Population Health (CESP), Cancer & Environment Group, University Paris-Sud, University Paris-Saclay, Villejuif, France; [2] INSERM UMRS 1147, EPIGENETEC, University Paris Descartes, Paris, France

Persistent Organochlorine Pollutants including Dichlorodiphenyltrichloroethane (DDT) and polychlorinated biphenyls (PCBs) are ubiquitous in the environment. Many studies have attempted to link DDT or PCB levels in human tissues to the risk of breast and other cancers. However, little attention has been paid to the genetic determinants of DDT and/or PCBs levels.

Our aim was to identify genetic variants modifying blood levels of p,p′-DDE (the major breakdown product of DDT) and PCB153 measured in a sample of women in France.

Blood levels of circulating p,p'-DDE and PCB153 were measured in 1236 healthy women selected as control subjects in an epidemiological study on breast cancer. We genotyped 474 SNPs in 54 *CYP* and *GST* genes involved in the metabolism of DDT and PCBs. Blood concentrations of p,p′-DDE and PCB153 were explored in relation to each SNP separately and to the genetic variation in the genes using the Adaptive Rank Truncated Product method. This approach allows investigating the role of a gene seen as a set of SNPs or of a set of genes in p,p′-DDE and PCB153 levels, and to gain statistical power as compared to a SNP by SNP approach. *P* values were adjusted for multiple testing by False Discovery Rate (FDR) method.

Blood levels of DDE and PCB153 were significantly associated with several SNPs located in *CYP* genes. The top-SNPs were rs8192719 (FDR= 3.1×10^{-26}) for DDE and rs7255904 (FDR= 0.01) for PCB153, both located in *CYP2B6*. At the gene level, *CYP2B6* was associated with levels of DDE (FDR= 0.0002) and PCB153 (FDR= 0.01). Genetic variation in the whole gene-set including *CYPs* and *GSTs* was significantly associated with blood levels of p,p′-DDE ($P = 0.0002$) and PCB153 ($P = 0.0002$).

These findings show that polymorphisms in genes involved in the metabolism are important determinants of organochlorine compound levels measured in humans, and should be taken into account as possible modifiers of the association between PCB and p,p′-DDE levels and disease risk in epidemiological studies.

13 | Statistical framework for large-scale integration of pathway knowledge in GWAS

Shrayashi Biswas[1], Soumen Pal[1], Samsiddhi Bhattacharjee[1]

[1]National Institute of Biomedical Genomics, Kalyani, India

Traditional Genome-wide Association Studies (GWAS) are under-powered to detect all associated variants due to a huge multiple-testing burden, particularly variants with lower frequency or weaker effects. There is recent interest in conducting prioritized analysis of GWAS using SNP-level functional annotations. Another, very rich source of biological knowledge exists in the form of gene-level functional annotation (e.g. pathways). Pathway enrichment analysis, occasionally used after a GWAS, helps in

the interpretation; but does not per se, improve the power of detecting additional novel loci. It is imperative to use knowledge of pathways/networks a-priori to "inform" the genome-wide search. Here, we develop a statistical method to enable a prioritized genome-wide scan using knowledge from a very large number of gene-level annotations (e.g. pathways from a pathway database). Our method takes in *p* values of SNPs from a GWAS and re-weights them optimally with pathway information utilizing a flexible framework based on regularized logistic regression. We demonstrate using simulations and real GWAS data on Systematic Lupus Erythematosus (SLE), Coronary Artery Disease and so forth that such a "pathway-guided GWAS" can indeed gain substantial power over unbiased GWAS, specifically when the associated SNPs cluster into a few biological pathways. Unlike most prioritization tools using False Discover Rate (FDR), our method maintains global false-positive rates at the same stringency as in a usual GWAS ($p < 5e\text{-}08$). We have created an R package "GKnowMTest" (Genomic Knowledge-guided Multiple Test) for flexible incorporation of gene-level knowledge into GWAS. Further, we have extended our method to allow large number of SNP-level functional annotations retaining the computationally simple regression-based framework.

14 | *APOE*, Alzheimer's Disease, and Hispanic Populations

Elizabeth E. Blue[1], Andréa R.V.R Horimoto[2], Ellen M. Wijsman[1,2,3], Timothy A. Thornton[2]

[1]Division of Medical Genetics, University of Washington, Seattle, United States of America; [2]Department of Biostatistics, University of Washington, Seattle, United States of America; [3]Department of Genome Sciences, University of Washington, Seattle, United States of America

Although the relationship between *APOE* and both Alzheimer's disease (AD) risk and age-at-onset (AAO) is well-established in cohorts with European ancestry, there is debate regarding its effects in other ancestry groups. We compared individuals with European and Hispanic ancestry from the National Institute on Aging's Late-Onset Alzheimer's Disease and the National Cell Repository for Alzheimer's Disease (NIALOAD) study with the Columbia University Study of Caribbean Hispanics with Familial and Sporadic Late Onset AD (CU Hispanics) data using a combination of survival analyses and genome-wide association (GWAS) testing for both AD risk and AAO. Survival analyses show the difference between both *APOE* E2 and E4 heterozygotes versus non-carriers among the Europeans is approximately twice that observed among Hispanics. The effects of *APOE* on GWAS results differ dramatically between European and Hispanic data. Although all GWAS adjusted for population structure and relatedness, the correlation between p-values from models with and without *APOE* adjustment was much weaker among the European data versus the Hispanic data for both AD risk ($r^2 = 0.49$ vs. $r^2 = 0.97$ or $r^2 = 0.81$) and AAO ($r^2 = 0.58$ vs. $r^2 = 0.87$ or $r^2 = 0.83$). Significant association between the *APOE* region and both AD risk and AAO was detected in the European and the CU Hispanic data, although *p* values were much larger in the CU Hispanics. GWAS for AAO identified additional significant associations in the NIALOAD Europeans (11p11.2, p = 4.9e-08) and CU Hispanics (14q24.3, p = 6.39e-11) corresponding with genes implicated in AD risk or pathology.

15 | Bivariate genome-wide association study reveals new variants associated with eosinophil cationic protein and eosinophil-derived neurotoxin levels

Raphaël Vernet[1,2], Patricia Margaritte-Jeannin[1,2], Régis Matran[3], Farid Zerimech[4], Valérie Siroux[5], Marie-Hélène Dizier[1,2], Rachel Nadif[6,7], Florence Demenais[1,2], Emmanuelle Bouzigon[1,2]

[1]Inserm, UMR-946, Genetic Variation and Human Diseases Unit, Paris, France; [2]Université Paris Diderot, Paris, France; [3]Université Lille and CHU de Lille, Lille, France; [4]Pôle de Biologie Pathologie Génétique, Laboratoire de Biochimie et Biologie Moléculaire, CHU de Lille, Lille, France; [5]Inserm, CNRS, IAB, Université Grenoble Alpes, Grenoble, France; [6]INSERM U1168, VIMA (Aging and Chronic Diseases: Epidemiological and Public Health Approaches), Villejuif, France; [7]Université Versailles St-Quentin-en-Yvelines, UMRS 1168, Montigny-le-Bretonneux, France

The number of genetic factors identified for asthma remains limited. The study of biological phenotypes involved in the pathophysiological mechanisms of asthma may help in the understanding of the genetic architecture of the disease.

To characterize the genetic factors influencing two immunomodulatory enzymes released by eosinophils in the allergic response in asthma: eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN), we conducted univariate and bivariate genome-wide association (GWA) analyses of these phenotypes in 1,018 subjects from the French Epidemiological study on the Genetics and Environment of Asthma (EGEA) using 1000-Genomes imputed SNPs.

The univariate GWA analyses detected two genome-wide significant loci associated with ECP on chromosomes 14q11 (rs7141958, $P = 1.2 \times 10^{-10}$) and 18q21 (rs8097644, $P = 2.7 \times 10^{-8}$) and one significant locus at 14q11 associated with EDN (rs3790067, $P = 4.3 \times 10^{-11}$). The two lead SNPs in the 14q11 region were independent

signals ($r^2 = 0.03$). The bivariate analysis confirmed the associations with the 14q11 (rs3790067, $P = 1.7 \times 10^{-13}$) and 18q21 regions (rs8097644, $P = 6.9 \times 10^{-9}$) and detected a third region on 1p31 (rs12091953, $P = 3.7 \times 10^{-8}$). Fine-mapping of these regions, based on bivariate conditional analysis, showed that the 14q11 region contained two distinct signals: the lead SNP rs3790067 and another SNP, rs76185655, in moderate linkage disequilibrium (LD) with rs7141958 ($r^2 = 0.5$). Interrogation of expression quantitative trait loci (eQTL) databases showed that the latter SNPs (or proxies) were associated with expression of two genes in the blood, RNASE2 and RNASE3 ($P < 10^{-11}$), which are 73 kb apart on 14q11.

This study shows that bivariate analysis can increase the power to detect new loci and can help in refining the associated region.

16 | JM-SNP: joint modeling of multiple longitudinal and multiple survival traits to characterize the genetic architecture of complex traits

Myriam Brossard[1], Andrew D. Paterson[2,3], Osvaldo Espin-Garcia[1,3], Radu V. Craiu[4], Shelley B. Bull[1,3]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [2]Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [4]Department of Statistical Sciences, University of Toronto, Toronto, Canada.

Motivated by complex genetic architecture of Type 1 diabetes complications (T1DC) where multiple SNPs are associated with T1DC and/or intermediate time-dependent risk factors, we propose JM-SNP, a joint model of multiple longitudinal and multiple time-to-event traits to characterize SNPs having direct effects on T1DC and/or indirect effects mediated by longitudinal traits.

JM-SNP is formulated with two components, (a) a multivariate mixed model for longitudinal quantitative traits as a function of time and SNP effects and (b) a multivariate frailty model for joint analysis of survival traits, depending on SNP effects and individual trajectories from (a). The two components are linked through subject-specific shared random-effects. We develop parameter estimation under a 2-stage approach, where (a) and (b) are fitted successively. We classify a SNP association with a time-to-event trait as: indirect if the SNP has a non-null effect in component (a) but not in (b); direct if the SNP has a non-null effect in (2) but not in (a); direct/indirect if the SNP has non-null effects in both (a) and (b).

We assessed JM-SNP performance by simulation studies under a scenario of complex genetic architecture that mimics key features of T1DC from the Diabetes Control and Complications Trial. This involves two longitudinal traits (HbA1c, blood pressure), two time-to-event outcomes (retinopathy, nephropathy), one unmeasured longitudinal trait (unexplained correlation between survival traits) and five causal SNPs with direct and/or indirect associations. Overall, JM-SNP exhibits good performance (low bias, Type 1 error control, relative power) even when the analysis model is not fully specified.

17 | Should cases with a clinical diagnosis of Giant Cell Arteritis be included in genetic association studies? Analysis of the UK Giant Cell Arteritis consortium cohort

Charikleia Chatzigeorgiou[1], Sarah L. Mackie[1], UKGCA Consortium[1], Mark M. Iles[1], Javier Martin[2], Ann W. Morgan[1], Jennifer H. Barrett[1]

[1]School of Medicine, University of Leeds, and NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom; [2]Instituto de Parasitología y Biomedicina "López-Neyra," CSIC, PTS Granada, Granada, Spain

Giant cell arteritis (GCA) is the commonest form of vasculitis in individuals older than 50 years in Western countries. Genetic association studies are often restricted to cases with a positive temporal artery biopsy (TAB). Many GCA cases are diagnosed on the basis of clinical features alone and previous studies have excluded these cases, which may result in a loss of power. We explored the effects of including cases without a positive TAB in a genetic association study of 663 cases (356 with a positive TAB) from the UK Giant Cell Arteritis Consortium and 2619 controls from Welcome Trust Case Control Consortium. After the imputation with SNP2HLA, a total of 7924 genetic variants in the Human Leucocyte Antigen region were analysed, first using all cases and then only TAB-positive cases. Restricting to 1822 variants reaching P value < 0.1, P values from the two analyses were highly correlated (Spearman's $r^2 = 0.86$). Using the whole sample, P values were on average slightly more significant ($p < 0.0001$), but beta estimates were closer to zero (24% lower, $p < 0.0001$), suggesting an increase in power but reduction in effect size due to misclassification. On using the whole data set for fine-mapping, similar findings were observed to those of a recently published larger study based on TAB-positive cases: the association largely explained by the amino acid HLA-DRß1 His13 (odds ratio: 1.79, $p < 10^{-15}$) and an amino acid in position 56 of HLA-DQa1. In conclusion, there may be some advantages in relaxing the strict definition

of GCA cases based on TAB-positive status for discovery, but that this may lead to under-estimates of effect sizes.

18 | Phenome-wide investigation of Alzheimer's disease-related phenotypes

Hung-Hsin Chen[1], Lauren E. Petty[1], William S. Bush[2], Adam C. Naj[3], Jennifer E. Below[1]

[1]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, United States of America; [2]School of Medicine, Case Western Reserve University, Cleveland, United States of America; [3]University of Pennsylvania Perelman School of Medicine, Philadelphia, United States of America

Alzheimer's disease (AD) is a highly heritable but late-onset disease. Previous epidemiological studies have reported several diseases associated with AD, for example, diabetes, hyperlipidemia, and hypertension. However, the molecular genetic pathology these disorders share with AD is still unclear. Phenome-wide investigation of genetically regulated expression provide a chance to explore correlated traits through common regulation of gene expression. In this study, we aimed to identify AD-related diseases or symptoms based on a shared genetic architecture. We used genetically regulated expression to evaluate the association between gene and health outcome, utilizing associations from Vanderbilt's DNA biobank, BioVU, which has collected DNA samples from over 280,000 participants and linked with their electronic health records. PrediXcan was used to estimate the genetically regulated expression based on each individual's genotype data with genotype tissue expression (GTEx) trained model in a subset of 23,000 genotyped individuals. A total of 70 AD-related genes were chosen to identify associated health outcomes, and a significant association was defined by p value less than 0.05. To obtain a p value for each outcome, we compared the number of associated genes in each outcome with the empirical null distribution from 100,000 times permutation. Results show that four different health outcomes reached phenome-wide significance, including diabetes mellitus (count = 41, P value $<1\times10^{-5}$), cerebral degeneration (count = 36, P value $=1\times10^{-5}$), edema (count = 38, P value $=2\times10^{-5}$), and memory loss (count = 36, P value $=3\times10^{-5}$). In short, we investigated and identified several AD-related diseases via shared genetic architecture, highlighting the need for further validation of the range of phenotypic consequences of dysregulation of AD-associated genes.

19 | Prediction of treatment response from genome-wide SNP data in rheumatoid arthritis patients

Svetlana Cherlin[1], Heather J. Cordell[1], MATURA Consortium[2]

[1]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; [2]MAximising Therapeutic Utility in Rheumatoid Arthritis

Although a number of treatments are available for Rheumatoid Arthritis (RA), each of them shows a significant non-response rate in patients. Therefore, predicting a priori the likelihood of treatment response would be of great benefit. Here we conducted a comparison of a variety of statistical methods for predicting three measures of treatment response, between baseline and three or six months, using genome-wide SNP data from RA patients available from the MAximising Therapeutic Utility in Rheumatoid Arthritis (MATURA) consortium. Two different treatments (tumour necrosis factor inhibitors and methotrexate) and a variety of statistical methods covering a wide range of genetic architectures were evaluated. We used 10-fold cross validation to assess predictive performance, with nested 10-fold cross validation used to tune the model parameters when required. Overall, we found that SNPs add very little prediction information to that obtained using clinical characteristics only, such as baseline trait value. This observation can be explained by the lack of strong genetic effects and the relatively small sample sizes available; in analysis of simulated and real data with larger effects and/or larger sample sizes, prediction performance was much improved. Overall, methods that were consistent with the genetic architecture of the trait were able to achieve better predictive ability than methods that were not. For treatment response in RA, methods that assumed a complex underlying genetic architecture achieved slightly better prediction performance than methods that assumed a simplified genetic architecture.

20 | Modeling with semi-continuous predictors: power and bias in the presence of zero-inflated metabolites

Su H. Chu[1,2], Jessica Lasky-Su[1,2], Peter Kraft[3,4]

[1]Channing Division of Network Medicine, Brigham and Woman's Hospital, Boston, United States of America; [2]Department of Medicine, Harvard Medical School, Boston, United States of America; [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, United States of America; [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, United States of America

Zero-inflated metabolite levels are common when analyzing metabolomic data, whether due to true absence of the metabolite (as in the case of xenobiotics) or due to measurements falling below the limit of detection. While much statistical literature has explored modeling zero-inflated outcomes, less attention has been paid to the context in which the exposure of interest is semi-continuous. Furthermore, as genetics have been shown to associate with metabolite levels, best modeling practices for metabolites will have implications for integrative studies combining genetics and metabolomics.

We conducted simulation studies to compare power and bias of effect estimates for three methods: two typical approaches in case-control studies, naïve logistic regression (n-LR) and logistic regression with log-transformed metabolite levels (ln-LR); and a new and simple alternative employing logistic regression with interaction terms (2df-LR). Using a factorial design, we explored the effect of altering limit of detection thresholds, mean and standard deviation (SD) of distributions for observed metabolite levels, metabolite effect size, probability of xenobiotic ingestion, and metabolite sampling probabilities, with 1000 iterations each. Simulations were further conducted to investigate differences in power and bias estimates with respect to chronicity of the metabolite-disease association (i.e. acute vs chronic outcomes).

All approaches demonstrated appropriate Type-I error rates. In general, n-LR demonstrated good power performance across a wide range of simulations; however, 2df-LR performed best with higher sample proportions of xenobiotic ingestion in acute settings. Overall, most methods were biased toward the null, but best performance with respect to the variability and degree of bias varied by simulation.

21 | Invited abstract: Multiethnic GWAS and fine-mapping

David V. Conti[1]

[1]*Department of Preventive Medicine, University of Southern California, Los Angeles, United States of America*

Genome-wide association studies (GWAS) have identified numerous risk loci for many diseases and traits. While the majority of these discoveries have occurred with studies containing individuals of European ancestry, it is critical that future studies perform GWAS in non-European populations to provide insight into both ancestry-specific variation and common risk variation across populations. Additionally, follow up of regions identified through GWAS with multiethnic fine-mapping can improve characterization by leveraging different linkage disequilibrium (LD) structures across diverse populations. For a single population, we have previously developed an approach (JAM) that uses marginal summary statistics and estimated correlation structure to fit joint multi-SNP models. More recently, we have expanded this approach to explicitly account for the difference in covariance between SNPs across each population. Through simulations, we demonstrate that the approach is better for inference of the number of independent signals within a fine-mapping region and for identifying the specific causal variants. As the JAM approach uses linear regression as an underlying framework, it can be readily extended to more complex selection procedures and models, such as hierarchical regression models. These models can be used to incorporate prior information on variant characteristics, such as annotation, to better identify causal variants. Throughout this presentation, we demonstrate the use of these methods to genetic association studies for prostate and colon cancer and consisting of individuals with diverse ancestry. Overall, better characterization of risk loci will facilitate understanding the underlying biological mechanisms and application of genetic risk prediction models to all populations.

22 | Trans-ethnic meta-analysis of gestational diabetes reveals shared genetic background with type 2 diabetes

Natalia Pervjakova[1,2,3], James P. Cook[4], Andrew P. Morris[4], Teresa Ferreira[5,6], Reedik Mägi[1] – on behalf of the GenDip Consortium

[1]*Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia;* [2]*Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia;* [3]*Genomics of Common Disease, Division of Diabetes, Endocrinology and Metabolism, Department of Medicine, Imperial College, London, United Kingdom;* [4]*Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom;* [5]*Big Data Institute, Li Ka Shing Center for Health for Health Information and Discovery, Oxford University, Oxford, United Kingdom;* [6]*Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, United Kingdom*

Gestational diabetes mellitus (GDM), defined as glucose intolerance first recognized in pregnancy, has important implications for both mother and child. Offspring of mothers with GDM have an increased risk of birth complications associated with increased birth weight and adiposity, as well as for developing metabolic syndrome, type 2 diabetes (T2D) and cardiovascular disease in later life. Furthermore, in a 5–10 years perspective, it is estimated that around 50% of the women diagnosed with GDM will develop T2D. To date, the only genome-wide association

study (GWAS) for GDM, in 1,399 affected Korean women, revealed associations at two loci: *MNTR1B* and *CDKAL1*.

We conducted trans-ethnic meta-analysis of 21 GWAS in 5,374 cases and 346,506 controls of diverse ancestry (62% European, 16.5% East Asian, 3.2% Mexican-American and 18.3% Afro-Caribbean), each imputed up to reference panels from the 1000 Genomes Project or Haplotype Reference Consortium. The trans-ethnic meta-analysis, performed with MR-MEGA to allow for heterogeneity in allelic effects between ancestries, included 13,980,490 variants, after excluding those with minor allele count <5 and imputation quality < 0.4 in each GWAS.

We replicated both known GDM associations at genome-wide significance ($P < 5 \times 10^{-8}$): *MTNR1B* (rs10830963, $P = 2.3 \times 10^{-49}$) and *CDKAL1* (rs9348441, $P = 9.7 \times 10^{-15}$). We also identified three additional independent loci: *TCF7L2* (rs7903146, $P = 1 \times 10^{-14}$) and *CDKN2A/B* (rs10811660, $P = 1.9 \times 10^{-9}$) and *LOC105369513* (rs143421658, $P = 4.1 \times 10^{-8}$). Four of these loci (*MTNR1B*, *CDKAL1*, *TCF7L2*, and *CDKN2A/B*) have been robustly associated with type 2 diabetes, and lead variants are identical to those we have identified for GDM, supporting a shared underlying genetic architecture for both diseases.

23 | Incorporating transcriptome data to study genome-wide gene-environment interaction

Brandon J. Coombes[1], Beth Larrabee[1], Hugues Sicotte[1], Sue L. McElroy[2], Mark A. Frye[4], Robert Yolken[4], Joanna M. Biernacka[1,3]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, United States of America; [2]Lindnor Center of HOPE/University of Cincinnati, Cincinnati, United States of America; [3]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, United States of America; [4]Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, United States of America

We aim to study gene-environment (GE) interactions with Cytomegalovirus (CMV) exposure antibody titer levels that are associated with age-of-onset of bipolar disorder (BD). CMV IgG and IgM antibody levels were measured in 837 BD patients (664 early onset and 173 late onset). In a two-step genome-wide GE interaction scan, we found one significant interaction with a variant 150 kb upstream of a gene that encodes for glutamate metabotropic receptor (*GRM3*). *GRM3* is involved in brain function and has been found to be perturbed in individuals with schizophrenia. We next proposed several gene-level GE interaction tests which incorporate tissue-specific expression quantitative trait loci (eQTLs) information by either using the predicted gene expression for a tissue in a regression model with an interaction term, using the eQTL estimates for a tissue as weights in a variance component score test of interaction, or combining eQTL information across tissues. For each test, we studied the performance in our study of GE interaction affecting age of onset of BD. In a scan of 7866 genes with predicted expression in four candidate brain tissues, we found one gene, DNA polymerase epsilon (*POLE*), had significant interaction with CMV IgM. DNA damages caused by CMV infection combined with mutations in *POLE* that disrupt DNA repair may explain why subjects with low expression of *POLE* and high levels of CMV IgM antibody titer were more likely to develop BD at an earlier age. Combining eQTL information across tissues performed better than the other methods testing each tissue separately.

24 | Evaluation of a targeted custom capture bisulfite sequencing approach

Richard Munthali[1], Aida Eslami[1], Ming Wan[1], Andrew Sandford[1], Kathleen Oros Klein[2], Celia Greenwood[2,3], Denise Daley[1]

[1]Center for Heart and Lung Innovation, University of British Columbia, Vancouver, Canada; [2]Lady Davis Institute for Medical Research, Montreal, Canada; [3]Departments of Oncology, Epidemiology, Biostatistics & Occupational Health, and Human Genetics, McGill University, Montreal, Canada

Cytosine-guanine dinucleotides (CpG) methylation can be modified by environmental exposures, such as smoking and medication and differential methylation of CpG sites likely contributes to the etiology of many common complex diseases. As such there is increased interest in evaluating CpG methylation and conducting epigenome-wide association studies. CpG methylation can be assessed using a variety of methods including whole genome bisulfite sequencing which can be prohibitively expensive, pyrosequencing of small regions, or the relatively inexpensive, and commonly utilized chip based approaches such as Illumina's 27 K, 450 K and Epic arrays. While cost effective, there are significant limitations of the array based approaches including the limited number CpG sites interrogated and the number of CpG sites that demonstrate variable methylation.

An emerging alternative to array based approaches, targeted capture bisulfite sequencing can offer several advantages including greater number CpGs that can be assessed, greater percentage of CpG sites with variable methylation, joint sequencing of both SNPs and CpG sites facilitates evaluation of both haplotype and SNP effects. We recently piloted the Illumina TruSeq

Methyl Capture Epic bisulfite sequencing approach using 48 samples selected from asthma studies. We compared the library content (TruSeq) to Illumina's arrays (27 K, 450 K and Epic). We found that the majority of CpG sites on the arrays (27 K (100%), 450 K (98%) and Epic (97.5%)) overlapped with the targeted capture library. However, the proportion of CpG sites with variable methylation is considerably higher 50% TruSeq versus 9% 450 K. A full analysis of the pilot of 48 samples is currently underway but our initial findings suggest that bisulfite targeted capture sequencing offers a cost effective alternative to whole genome bisulfite sequencing.

25 | Integrative network analysis identifies relationships between metabolomics, genomics, and risk factors for Alzheimer's disease

Burcu F. Darst,[1,2], Qiongshi Lu,[1,3], Sterling C. Johnson,[1,4,5,6], Corinne D. Engelman[1,2,5,6]

[1]University of Wisconsin, Madison, United States of America; [2]Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, United States of America; [3]Department of Biostatistics & Medical Informatics, Madison, United States of America; [4]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, United States of America; [5]Geriatric Research Education and Clinical Center, Wm. S. Middleton Memorial VA Hospital, Madison, United States of America; [6]Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, United States of America

Integrative multi-omics approaches could be useful for the identification of underlying mechanisms that contribute to complex diseases, such as Alzheimer's disease (AD). We performed an integrated network analysis to investigate relationships between genomics, metabolomics, and AD risk factors using the Wisconsin Registry for Alzheimer's Prevention (WRAP). Analyses included 1,111 Caucasian participants with whole blood expression for 11,376 genes (imputed from genome-wide genotyping using PrediXcan), 1,097 fasting plasma metabolites, 364 fasting cerebral spinal fluid metabolites, and 19 AD risk factors. After adjusting each of the 12,856 variables for potential confounders, residuals were used to test all 82,631,940 possible pairwise Spearman correlations. Inter-omic correlations meeting a Bonferroni-adjusted $P < 6.1e\text{-}10$ were used to develop an undirected graphical network, which included 908 edges (correlations) and 532 nodes (variables). Although no genes were directly correlated with AD risk factors, many genes were indirectly linked to AD risk factors through plasma metabolites. One correlation between *CPS1* and glycine has been reported in numerous previous investigations, as have correlations between glycine, body mass index (BMI), waist-hip ratio (WHR), and homeostatic model assessment for insulin resistance (HOMA-IR), which were also identified in our network. Mediation analyses revealed that although there were no direct associations between *CPS1* and these three risk factors, there were strong mediation effects such that lower *CPS1* expression led to higher plasma glycine levels, which led to lower BMI, WHR, and HOMA-IR. Thus, *CPS1* may have protective effects for certain AD risk factors. These results demonstrate the potential utility of other hypotheses generated by this network.

26 | Multi-omics data integration under a general likelihood based framework with an emphasis on the missing values

Sarmistha Das[1], Indranil Mukhopadhyay[1]

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

Multi-omics data integration poses a major challenge due to the presence of missing values that varies across different omics platforms. But imputing the missing values might be misleading especially when missing percentage is large and/or the reason of missing is unclear. So, to deal with such situations, our idea is to integrate every available information on multiple omics data.

In this study, we develop a general likelihood based approach to integrate multi-omics data namely, genotype, gene expression, methylation, and phenotype. We consider qualitative phenotype, for example, disease status which could be affected or unaffected. We evaluate the effect of missing data percentage across various omics, for different genetic association tests. Our general likelihood framework allows us to test different null hypotheses based on biological knowledge like genetic association, mediation effect, effect of methylation on gene expression and so forth.

We assume complete genotype and phenotype information although gene expression and/or methylation values may be missing for a number individuals (overlapping or non-overlapping). Here we consider only genetic association test based on case-control data in presence of gene expression and methylation data. We evaluate and compare the power of the proposed likelihood test by varying missing value percentages for the omics data through simulations. Our method gives robust inferences although the percentage of missing values vary. Also, it shows that the power increases as sample size increase. We derive the asymptotic distribution of our test statistic to calculate *P* values fast.

Application of our method to a real data set also provides encouraging results.

27 | Association analysis for bivariate traits with family data using generalized estimating equations

Mauricio A. Mazo Lopera[1], Nubia E. Duarte[2], Mariza de Andrade[3]

[1]Department of Statistics, Universidad Nacional de Colombia, Mendellín, Colombia; [2]Department of Mathematics, Universidad Nacional de Colombia, Antioquia, Colombia; [3]Department of Health Sciences Research, Mayo Clinic, Rochester, United States of America

Genome-wide association study (GWAS) is becoming fundamental in the arduous task of deciphering the etiology of complex diseases. Most of the statistical models used to address the genes-disease association consider a single response variable. However, it is common for certain diseases to be related with several phenotypes that may be correlated with each other. In addition, GWAS typically sample unrelated individuals from a population and therefore shared familial risk factors are not investigated. In This study, we propose to apply a bivariate model that associates two phenotypes with a genetic region and we include the family risk factor assuming that we have repeated measures from family data. Using generalized estimation equations (GEE), we model two phenotypes, either discrete, continuous or a mixture of them, as a function of genetic variables and other important covariates. We incorporate the kinship relationships into the working matrix extended to a bivariate analysis with repeated measures. The estimation method and the different hypothesis tests are developed in This study. In addition, we evaluate the proposed methodology with simulation studies and an application to real data.

28 | Effect of genetic susceptibility to schizophrenia and type 2 diabetes mellitus on hyperglycaemia in patients with schizophrenia

Tesfa Dejenie Habtewold[1,2], Md. Atiqul Islam[3], Edith J. Liemburg[2,4], Richard Bruggeman[2], Behrooz Z. Alizadeh[1]

[1]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands; [2]University of Groningen, University Medical Center Groningen, University Center for Psychiatry, Rob Giel Research Center, Department of Psychiatry, Groningen, The Netherlands; [3]Shahjalal University of Science and Technology, Department of Statistics, Sylhet, Bangladesh; [4]University of Groningen, University Medical Center Groningen, Department of Neuroscience, Groningen, The Netherlands

Type 2 diabetes (T2D) is a common comorbidity in patients with schizophrenia (SCZ). The underlying pathophysiologic mechanisms are yet to be found, although it can be argued that genetic susceptibility and antipsychotic usage are involved in the incidence of T2D. We, therefore, aimed to investigate whether genetic susceptibility to T2D and SCZ and usage of antipsychotic medications are associated with T2D among patients with SCZ. Data of 820 patients with non-affective psychosis were extracted from the Genetic Risk and Outcome of Psychosis (GROUP) cohort in the Netherlands and Belgium. Multiple linear regression analysis was applied to identify factors associated with glycated haemoglobin (Hb_{A1C}), which is a proxy measure for T2D. The genome-wide significant genetic risk score of T2D (p value $= 0.002$) was significantly associated with Hb_{A1C} and explained 3.4% of the variance in Hb_{A1C}. The polygenic risk score of SCZ (PRS_{SCZ}) and high metabolic risk antipsychotics explained 0.7% and 0.1% of the variance of Hb_{A1C}, although the association was not significant (p value $= 0.31$ and p value $= 0.89$, respectively). In addition, gender, daily cigarette smoking and history of cardiovascular diseases were significantly associated with Hb_{A1C}. This study has the potential to shed light on the pathogenetic mechanisms underlying T2D and SCZ comorbidity and predicting T2D in patients with SCZ.

29 | Wavelet screaming: a novel approach to analyzing GWAS data

William Denault[1], Julius Juodakis[1,2], Bo Jacobsson[1,2], Astanand Jugessur[1,3,4], Håkon K. Gjessing [3,4]

[1]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway; [2] Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [3]Centre for Fertility and Health (CeFH), Norwegian Institute of Public Health, Oslo, Norway; [4]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

We present a new approach to performing genome wide association studies (GWAS) which takes into account the functional nature of the genome. Our method is based on a sliding window approach that sequentially screens the entire genome for associations. We consider SNPs as a genetic signal, and for every screened region, we transform the genetic signal into the wavelet space. The association is identified at the wavelet coefficient location using a Bayesian Hierarchical model.

Our sliding-window approach is fast. It reduces the number of tests to be performed, enhances the detection of association signals by improving the modeling, and also provides a natural fine mapping tool. We performed simulations to test our method using polygenic simulated phenotypes linked to real genetic data.

Our method has the same power as the standard GWAS methodology for monomorphic signal, but a higher power for the detection of polygenic signal. It performs better than the GWAS standard methodology for small polygenic effects. The wavelet transform allowed the detection of signals spread out at multiple independent SNPs which explained a small amount of the phenotypic variation (<0.1%). From our results, we consider that Wavelet Screaming is a suitable alternative methodology to perform GWAS. Moreover, this method is well suited for the meta-analysis of multiple cohort studies.

We applied our methodology to the HARVEST data set to well-studied phenotypes (e.g., gestational age). We investigated previous SNPs identified through large meta analysis that are not detected in HARVEST using GWAS methodology. Wavelet Screaming detected known loci as well as loci that were previously unidentified.

30 | Analytic strategies for polygenic risk score modeling of laboratory values from biobank data

Jessica Dennis[1,2], Guanhua Chen[3], Peter Straub[1,2], Donald Hucks[1,2], Jonathan Mosley[4], Douglas Ruderfer[1,2], Nancy J. Cox[1,2], Lea K. Davis[1,2]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, United States of America; [3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, United States of America; [4]Department of Medicine, Vanderbilt University Medical Center, Nashville, United States of America

Polygenic risk score (PRS) modeling of laboratory values from electronic health records could identify novel biomarkers. Optimal strategies for analyzing these data, however, are unclear. We aimed to determine how best-fit *P* value thresholds (P_T) for SNP inclusion in a PRS varied across (a) a two-stage strategy for P_T training and validation; and (b) strategies for control of medication and disease. We developed our approach in 25,461 patients (9,259 training, 16,202 validation) in the Vanderbilt University Medical Center biobank (BioVU) with at least one lipid measurement. Discovery genome-wide association studies (GWAS) summary statistics for PRS weighting and thresholding were obtained from the Global Lipids Genetics Consortium, and we defined P_T as the discovery GWAS *P* value threshold that maximized the proportion of variability explained (R^2) in BioVU. We found that: (a) the R^2 of the training P_T in the validation set was comparable to the R^2 of an independently-fit P_T for high- and low-density lipoprotein cholesterol (HDL and LDL), but not triglycerides (TG); and (b) restricting to pre-medication observations marginally lowered the P_T across all traits, in accordance with patient selection into the discovery GWAS. Excluding patients with coronary artery disease or type 2 diabetes had minimal effects on the P_T for HDL and LDL but increased the TG P_T > 16-fold. In conclusion, the consistency of the HDL and LDL P_T suggests that these discovery GWAS are approaching maximal power. A two-stage strategy may be useful for phenotypes (e.g., TG) that have pleiotropic SNPs, are sensitive to confounding, or that have underpowered discovery GWAS.

31 | Genome-wide meta-analysis of parent-of-origin effects of asthma, atopy and airway hyperresponsiveness in four cohorts

Aida Eslami[1], Loubna Akhabir[1], Judith M. Vonk[2], Allan B. Becker[3], Anita L. Kozyrskyj[4], Andrew J. Sandford[1], Gerard H. Koppelman[2], Catherine Laprise[5], Denise Daley[1]

[1]University of British Columbia, Vancouver, Canada; [2]Groningen Research Institute for Asthma and COPD (GRIAC), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [3]Department of Pediatrics and Child Health, Faculty of Medicine, University of Manitoba, Winnipeg, Canada; [4]Department of Pediatrics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Canada; [5]Université du Québec à Chicoutimi, Saguenay, Canada; [6]Centre intégré universitaire de santé et de services sociaux de Saguenay (CIUSSSS), Chicoutimi, Canada

The main genetic effects of the common SNPs identified by Genome-Wide Association Studies (GWAS) do not fully explain the heritability of asthma. Genomic imprinting (parent-of-origin effects) is a potential mechanism which may explain this missing heritability. We aimed to identify candidate genomic regions for imprinting in asthma, atopy and airway hyperresponsiveness (AHR).

We used GWAS data from four family-based studies (trios): Canadian Asthma Primary Prevention Study (CAPPS), Study of Asthma Genes and Environment (SAGE), Saguenay-Lac-Saint-Jean Québec Familial Collection (SLSJ), and Dutch Asthma GWAS (DAG). We used a likelihood-based variant of the Transmission Disequilibrium Test. Parent-of-origin effects were analyzed by including parental sex as a modifier in the analysis. Meta-analysis was conducted using the parent-of-origin effects results of SLSJ, DAG, and the joint analysis of CAPPS and SAGE (CAPPS/SAGE), weighted by the number of informative transmissions for each study.

Meta-analysis for asthma, using results of SLSJ (251 trios), DAG (316 trios), and CAPPS/SAGE (141 trios), resulted in five independent SNPs with significant parent-of-origin effects with $P \leq 1.49 \times 10^{-5}$ (suggestive threshold). Meta-analysis for atopy, using results of SLSJ

(229 trios), DAG (312 trios) and CAPPS/SAGE (217 trios) resulted in two independent SNPs. Meta-analysis for AHR using results of SLSJ (132 trios), DAG (260 trios) and CAPPS/SAGE (219 trios) resulted in seven independent SNPs. Of the significant results, 11 out of 14 of the SNPs were in or near long non-coding (lnc)RNA genes. Our results suggest a possible role for lncRNAs in parent-of-origin effects in asthma and allergic phenotypes.

32 | Optimality in two-phase sampling designs for post-GWAS studies

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, Canada; [2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [3]Department of Statistical Sciences, University of Toronto, Toronto, Canada

Given typically modest genetic effect sizes, large-scale population studies are generally necessary to achieve genome-wide significance. Post-GWAS analysis focuses on fine-mapping of targeted genetic regions, but as yet costs of next-generation sequencing remain prohibitively expensive for large studies. Two-phase sampling design and analysis is a cost-reduction technique that utilizes collected data during Phase 1 GWAS (e.g. outcome, auxiliary variable) to select an "informative" (biased) subsample in Phase 2 fine-mapping. The main goal is to make inference on a genetic variable that is costly and/or onerous to measure (e.g. sequencing data). In two-phase sampling designs, less attention has been paid to determination of optimal study designs compared to effect estimation and hypothesis testing. To date, most of the work on optimal designs has focused on case-control studies, where for example, a "balanced" design has been claimed to be "near optimal." Strategies and tools that help researchers to select the (optimal) Phase 2 subsample are not readily available beyond binary outcomes. Considering the effect size and haplotype distribution between a GWAS-auxiliary-variable and a sequence variant as inputs, we develop two approaches for optimal two-phase designs within the exponential family under a semiparametric maximum likelihood framework. One is based on analytic expressions for the variance-covariance matrix (VCM), and alternatively, another is based on a computational approach via a genetic algorithm for optimization of the Phase 2 sampling space. We evaluate measures of statistical efficiency based on features of the VCM that have been proven useful to define optimality criteria in design of studies.

33 | Sparse Canonical Correlation Analysis (sCCA) significantly improves power of cross-tissue transcriptome-wide association studies (TWAS)

Helian Feng[1,2], Bogdan Pasaniuc[3,4], Megan Major[3,4], Peter Kraft[1,2]

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

Transcriptome-wide Association Studies (TWAS) relies on accurate estimation of genetic effects on gene expression and suffers from low power at small expression quantitative trait loci (eQTL) sample sizes or when data from relevant tissues is not available. Cross-tissue TWAS tackles these issued by leveraging eQTL data across multiple tissues. Here, we use Sparse Canonical Correlation Analysis (sCCA) to integrate data across tissues (i.e. the weighted average of highly genetically correlated tissue expression levels) and show that sCCA-TWAS outperforms traditional TWAS. We use simulations starting from real Genotype Tissue Expression (GTEx) data to compare to single-tissue TWAS; cross-tissue TWAS adjusting for multiple testing with Bonferroni and Generalized Berk-Jones (GBJ); and summarizing cross-tissue expression patterns using Principal Component Analysis (PCA). All methods control the Type I error both when there is no genetic effect on expression or when expression is not associated with outcome. When expression drives trait, sCCA-TWAS had the greatest power to detect a gene associated with outcome, even when the causal expression trait was not directly measured: for example, when gene expression explains 2% of the variability in outcome and the GWAS sample size is 20,000, the average power difference between sCCA and single-tissue, multi-tissue, and PCA approaches was 0.23, 0.26, and 0.40. The large gain in power is likely due to sCCA cross-tissue features being more likely to be significantly heritable than single-tissue or PC expression features (2.78x, and 3.72x respectively). Our results suggest that aggregating eQTL data across multiple tissues using sCCA can improve the specificity and sensitivity of TWAS. Comparison of these approaches also applied to multiple public complex trait GWAS.

34 | Detection of genetic similarities using unsupervised random forest

Césaire J. K. Fouodo[1], Inke R. König[1]

[1]*Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany*

Genome-wide association studies (GWAS) have in the past been successful in the identification of associations with well-defined phenotypes as well as in the establishment of supervised classification models based on univariable analyses. To perform multivariable genome-wide analyses in GWAS data, classical statistical approaches like linear or logistic regression models are not suitable due to the high-dimensional nature of the data. As an alternative, random forest (RF) based approaches are known to be one of the more appropriate supervised machine learning methods for high dimensional multivariable data analysis. RF based approaches have been shown to be fast and to produce good predictive performances in high dimensional classification problems. Furthermore, the method can be extended to unsupervised learning analysis. Our aim is to apply an unsupervised random forest (URF) strategy to GWAS data to cluster individuals into genetically homogeneous subgroups. Incorporating a novel boosting approach, we further expect to obtain more stable clusters and improve the performance of the method provided each boosting iteration takes into account the result of the previous iterations. Clustering individuals according to their genetic similarities and without knowledge of a phenotype can allow us to discover novel subgroups of patients. The evaluation of our approach is demonstrated in an application to samples from the Pan-Asia SNP Consortium database with the objective to validate the derived clusters against the known ethnical background.

35 | Large-scale trans-ethnic genome-wide association study reveals novel loci, causal molecular mechanisms and effector genes for kidney function

Nora Franceschini[1], Thu Le[2], Artur Akbarov[3], Maciej Tomaszewski[3], Andrew P. Morris[4] on behalf of the COGENT-Kidney Consortium

[1]*Department of Epidemiology, University of North Carolina, Chapel Hill, United States of America;* [2]*Division of Nephrology, University of Virginia, Charlottesville, United States of America;* [3]*Division of Cardiovascular Sciences, University of Manchester, Manchester, United Kingdom;* [4]*Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom*

Chronic kidney disease (CKD) is a major public health burden and affects nearly 10% of the global population. We assembled published and de novo genome-wide association studies of estimated glomerular filtration rate (eGFR), a measure of kidney function used to define CKD, in up to 312,468 individuals of diverse ancestry. We identified 93 loci attaining genome-wide significant evidence of association with eGFR ($p < 5 \times 10^{-8}$), including 20 mapping outside those previously reported for kidney function, with the strongest novel signals mapping to/near *MYPN* (rs7475348, $p = 8.6 \times 10^{-19}$), *SHH* (rs6971211, $p = 6.5 \times 10^{-13}$), *XYLB* (rs36070911, $p = 2.3 \times 10^{-11}$) and *ORC4* (rs13026220, $p = 3.1 \times 10^{-11}$). Across loci, we identified 127 distinct association signals at locus-wide significance ($p < 10^{-5}$), including four in the regions mapping to *SLC22A2* and *UMOD-PDILT*. Allelic effects on eGFR of index variants were consistent across populations, with no evidence of heterogeneity due to ancestry (Bonferroni correction, $p_{HET} < 0.00039$). Integration with functional and regulatory annotation revealed that eGFR association signals were jointly enriched in coding sequence, kidney-specific histone modifications and binding sites for HDAC2 and EZH2. Class I histone deacetylases (including HDAC2) are required for embryonic kidney gene expression, growth, and differentiation, and EZH2 mediates the development of renal fibrosis by downregulating expression of Smad7 and PTEN, and activating profibrotic signalling pathways. Annotation-informed fine-mapping revealed 40 variants accounting for >50% of the probability (π) of driving eGFR association signals, including coding alleles *GCKR* p.Leu446Pro (rs1260326, $p = 2.0 \times 10^{-35}$, $\pi = 0.861$), *CPS1* p.Thr1406Asn (rs1047891, $p = 1.5 \times 10^{-29}$, $\pi = 0.980$), *CERS2* p.Glu115Ala (rs267738, $p = 1.7 \times 10^{-10}$, $\pi = 0.553$) and *CACNA1S* p.Arg1539Cys (rs3850625, $p = 2.5 \times 10^{-9}$, $\pi = 0.990$), and expression quantitative trait loci in kidney for *UMOD* and *GP2* (rs77924615, $p = 1.5 \times 10^{-54}$, $\pi = 0.861$), *FGF5* (rs12509595, $p = 4.7 \times 10^{-16}$, $\pi = 0.571$) and *TBX2* (rs887258, $p = 2.7 \times 10^{-13}$, $\pi = 0.622$). These results define novel causal molecular mechanisms underlying kidney function association signals, and highlight genes through which their effects are mediated, offering a potential route to clinical translation and CKD treatment development.

36 | Prediction of CpG methylation status from SNP genotype data

James J. Fryett[1], Andrew P. Morris[2], Heather J. Cordell[1]

[1]*Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom;* [2]*Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom*

In transcriptome-wide association studies, gene expression is predicted from genotype data, then tested for associated with a trait of interest. This approach could also be used to investigate the role that alternative biological measures play in complex traits. Here, we explore how well CpG (cytosine-guanine dinucleotides) methylation status can be predicted from SNP genotypes, with a view to conducting a methylome-wide association study. Using data from the Avon Longitudinal Study of Parents and Children (ALSPAC), we first explore the heritability of CpG methylation explained by nearby SNPs. For most CpGs, we find that the heritability is near zero, with the exception of a small subset of CpGs. We then investigate how well three methods - ridge regression, elastic net and LASSO (least absolute shrinkage and selection operator) - predict CpG methylation status using genotypes of local SNPs. For most CpG sites, methylation was poorly predicted, although for approximately 1% of CpG sites, it could be predicted well ($R^2 > 0.5$). On average, elastic net and LASSO outperform ridge regression, suggesting that the underlying architecture of methylation may be sparse. Estimates of prediction accuracy were marginally smaller than heritability estimates, indicating that prediction is as good as can be expected when predicting using only genotype data. We conclude that the subset of CpGs whose methylation can be predicted with reasonable accuracy represent a strong candidate set for inclusion in future methylome-wide association studies. This approach may help elucidate the role of CpG methylation in a wide range of complex traits.

37 | Differences in imputation accuracy from one gene to another and impact on rare variant association testing

Emmanuelle Génin[1], Thomas E. Ludwig[1], Ozvan Bocher[1] and FREX consortium

[1]*UMR 1078 Génétique, Génomique fonctionnelle et Biotechnologies, Inserm, Université de Brest, EFS, CHU Brest, Brest, France*

Imputation of untyped variants in genotyped data using sequenced reference panels is an interesting and cost-effective solution for association testing. Imputation has been extensively used in the context of common variant association tests where it has shown its utility. However, less studies of the interest and limits of imputation were done in the context of rare variant association tests using gene-based tests.

Here, we used the data from the French Exome (FREX) project (http://med-laennec.univ-brest.fr/FrExAC/) to assess how many and how well variants found in these 574 exomes are imputed when using only information on SNP-chip data on them (they were all genotyped on Illumina OmniExpress chip) and the imputation server with the Haplotype Reference Consortium (HRC) data (64,976 haplotypes at 39,235,157 SNPs). We looked at the results in a gene-centric view to highlight genes that are well imputed and genes that are poorly imputed. Indeed in most studies performed so far on imputation, results are given globally at the scale of the entire genome. This is useful for investigators who want to embark on genome-wide study but less useful for investigators who cannot afford genome-wide analysis and just want to know how genes from their favorite gene list will be imputed. For each gene, we provide information on the percentage of variants from the exome that are detected and on the average concordance of genotypes as well as the consequences in terms of expected power loss of using imputed rather than exome sequenced data.

38 | Investigation the expression levels of miR-155 and miR-365 in the plasma of relapsed and unconscious breast cancer patients compared with healthy subjects

Payam Ghasemi-Dehkordi[1], Morteza Hashemzadeh-Chaleshtori[1]

[1]*Cellular and Molecular Research Center, Basic Health Sciences Institute, Shahrekord University of Medical Sciences, Shahrekord, Iran*

Breast cancer is the most common cause of death in women worldwide and early diagnosis of a disease which is also regarded to the success of treatment and good prognosis. In recent years, the role of microRNAs has been defined in cancer therapy, tumorigenicity, anticancer drugs, and recurrence of cancers. miR-155 and miR-365 have a high potential role in the growth and aggression of breast cancer with high sensitivity and specificity and they can use as a biomarker for the detection of certain cancers. The purpose of the current study was to evaluate the effects and recurrence of miR-155 and miR-365 as diagnostic biomarkers in breast cancer patients. A total of 44, 24 and 20 blood specimens were collected from healthy women, women with primary and recurrent breast cancer, respectively. The plasma samples were isolated and the expression levels of miRNA were evaluated by qRT-PCR technique. All data were collected and analyzed using SPSS software version 20. The expression level of miR-155 and miR-365 in women with primary breast cancer were 2.82 and 2 times higher than their level in healthy women and these changes were significant for both microRNAs ($p < 0.001$). Plasma levels of miR-155 and miR-365 in women with

recurrent breast cancer compared with healthy women were increased by 5.8 and 2.2 times, respectively, and it was also significant for both microRNAs ($p < 0.001$). The miR-155 and miR-365 expression changes in plasma samples were not significant between these the two patient groups ($p = 0.562$ and $p = 1$, respectively). The findings of the current study were indicated that miR-155 and miR-365 are valuable biomarkers for the early detection and recurrence of breast cancer in the patients.

39 | Fast calling of copy number variations from cohort-wide high-depth whole genome sequencing

Grace Png[1,2], Daniel Suveges[1,3], Klaudia Walter[1], Kousik Kundu[1], Ioanna Ntalla[4], Emmanouil Tsafantakis[5], Maria Karaleftheri[6], George Dedoussis[7], Eleftheria Zeggini[1], Arthur Gilly[1]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom; [2]University of Cambridge, Cambridge, United Kingdom; [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom; [4]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [5]Anogia Medical Centre, Anogia, Greece; [6]Echinos Medical Centre, Echinos, Greece; [7]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens, Greece

Copy number variants (CNVs) are large deletions or duplications at least 50 to 200 base pairs long. They play an important role in multiple disorders, however, accurate calling of CNVs is still a challenge today. Several large-scale high-depth whole-genome sequencing efforts are underway, yet most current approaches to CNV detection use raw read alignments, which are computationally intensive to process. Here, we introduce UN-CNVC, a regression tree-based CNV detection algorithm that produces population-wide callsets at unprecedented speed, using depth at variant calls from whole-genome sequencing (WGS). As a proof-of-principle, we applied UN-CNVC to WGS (>18x) from 6,898 samples across four European cohorts, and describe a rich large variation landscape comprising 2059 CNVs. 38.4% of detected events were previously reported in the Database of Genomic Variants, a dbSNP equivalent for CNVs. 18% of high-quality deletions excise entire genes, and we recapitulate known phenotype-altering events such as those affecting the *GSTM1* and *RHD* genes. We then test for deletions associated with 58 medically-relevant traits in 3,079 individuals, and with 275 protein levels in 1,457 individuals to assess the potential clinical impact of the detected CNVs. We describe the linkage structure and copy number variation underlying an association between levels of the CCL3 protein and a complex structural variant affecting *CCL3L*, a paralog of the *CCL3* gene ($p = 3.6 \times 10^{-12}$). This study demonstrates that existing population-wide WGS callsets can be mined for CNVs with minimal computational overhead, delivering insight into a little-studied, yet potentially impactful class of genetic variant.

40 | Bayesian variable selection for Mendelian randomization

Apostolos Gkatzionis[1], Paul J. Newcombe[1]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

Mendelian randomization is the use of genetic information to assess the existence of a causal relationship between a risk factor and an outcome of interest. A Mendelian randomization analysis relies on a set of genetic variants (SNPs) that are strongly associated with the risk factor and only associated with the outcome through their effect on the risk factor.

Nowadays, SNPs associated with a risk factor can be obtained through large consortia genome-wide association studies (GWAS). The results reported by such studies, typically a list of univariate associations between SNPs and the risk factor can be used to perform variable selection among a set of candidate SNPs. The JAM algorithm (Joint Analysis of Marginal summary statistics, Newcombe, Conti and Richardson, 2016) is a Bayesian model search algorithm recently proposed for this task. JAM uses a stochastic stepwise addition-deletion procedure to perform SNP selection based on GWAS summarized data. It accounts for genetic correlations and can be parallelized to analyse large numbers of SNPs simultaneously.

We discuss the use of JAM for Mendelian randomization. In this context, we propose an extension that augments the JAM posterior with a loss function To penalize SNPs having a pleiotropic effect on the outcome. The performance of the new algorithm is illustrated with simulated and real data.

41 | How well can we classify coronary artery disease using all genetic data and choosing the best classification algorithm?

Damian Gola[1,2], Till Andlauer[3], Nazanin Mirza-Schreiber[3], Lingyao Zeng[4], Bertram Müller-Myhsok[3], Inke R. König[1,2,5]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]German Centre for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Lübeck, Germany; [3]Department of

Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany; [4]Deutsches Herzzentrum München, Technische Universität München, München, Germany; [5]Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL), Grosshansdorf, Germany

Coronary artery disease (CAD) is the leading global cause of mortality and has substantial heritability with a polygenic architecture. Recent approaches of risk prediction were restricted in that they did not utilize the complete genetic information available but rather focused on genetic loci found to be associated to CAD. Also, they were based on simple genomic risk scores (RS) not taking possible interdependencies between variants into account. We, therefore, benchmarked RS, logistic (penalized) regression, naïve Bayes (NB), random forests (RF), support vector machines (SVM) and gradient boosting (GB) on a data set of 7736 CAD cases and 6774 controls from Germany to identify the algorithms for most accurate classification of CAD patients. After training, the final models were validated on an independent data set of 527 CAD cases and 473 controls out of the same population and two further independent datasets from the UK (1874 CAD cases, 1874 controls) and Germany, France and UK (365 CAD cases, 401 controls).

We found RS using 50 633 genetic markers to be the most suitable algorithm for CAD classification, yielding an area under the receiver operating curve (AUC) of 0.92 (95% CI [0.90, 0.95]) in the validation data from the same population. NB and SVM performed better than RF and GB, with AUC of about 0.81 and 0.75, respectively. Applied to datasets from slightly different populations, however, all classification models had markedly reduced AUC of 0.5–0.6. We conclude that using all available genetic information can boost classification performance, although intensive validation of prediction models is crucial to assess their usability in populations different from those used to build the models.

42 | Defining trait core genes with networks

Britney E. Graham[1,2], Kevin Chesmore[3], Scott M. Williams[1,2]

[1]Systems Biology and Bioinformatics Program, Case Western Reserve University, Cleveland, United States of America; [2]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, United States of America; [3]Department of Molecular and Systems Biology Genetics, Geisel School of Medicine, Dartmouth College, Hanover, United States of America

The concept that many genes affect a single trait, that is the omnigenetic model, has recently gained momentum due to the observation from genome-wide association studies (GWAS) that numerous phenotypes each associate with hundreds of variants. Height is one such omnigenetic, complex trait. It is highly heritable with ~700 GWAS associated loci in the GWAS catalog. We propose a network-based analysis to determine which height associated loci are most important to the presumed genetic architecture of the trait, that is the "Core" genes as proposed by Boyle, LI and Pritchard, 2017. Using proteomic, pathway and gene interaction-based analyses of the height associated loci, we found that the best networks used confidence of relatedness between genes, edge weight, to determine nodes (genes) most important to the network (i.e., the Core genes). We investigated the pleiotropy of height genes and the relationship between phenotypes linked to those genes using a multiple phenotype-gene mapping method based on the GWAS catalog. Our resulting network revealed the number of Core genes to be approximately one third of the total, with several well-defined gene clusters, many concentrated around several oncogenes, many of which associate with more than one cancer. Phenotype-gene clusters include both height and body mass index (BMI) related traits as well as fetal size and preterm birth related traits. This last observation is supported by these Core genes being expressed with high confidence in placental and fetal tissues. This study shows that networks are a powerful tool for the mathematical identification of Core genes for complex disease.

43 | Deriving significance thresholds for genome-wide admixture mapping studies

Kelsey E. Grinde [1], Lisa A. Brown [1,2], Timothy A. Thornton [1], Sharon R. Browning [1]

[1]Department of Biostatistics, University of Washington, Seattle, United States of America; [2]Seattle Genetics, Bothell, United States of America

Admixture mapping is a powerful approach for identifying genetic variants associated with complex traits. Genome-wide admixture mapping studies have become more widely implemented in recent years, due in part to technological advances and growing international effort to increase the diversity of genetic studies. However, many open questions remain about appropriate implementation of admixture mapping studies, including how best to control for multiple testing, particularly in the presence of population structure. In this study, we extend an existing theoretical framework to characterize the correlation of local ancestry and admixture mapping test statistics in admixed populations with any number of ancestral populations and arbitrary population structure. Based on this framework, we propose an approach for deriving significance thresholds for genome-wide admixture mapping studies. We validate our approach via simulation studies and application to genotype data for

8,064 unrelated African American women and 3,425 unrelated Hispanic/Latina women from the Women's Health Initiative SNP Health Association Resource. Our approach yields significance thresholds of 2.1×10^{-5} and 4.6×10^{-6} for admixture mapping studies in the African American and Hispanic/Latina samples, respectively. Compared to other commonly used multiple testing correction procedures, our method is fast, easy to implement, and controls the family-wise error rate even in the presence of complex population structure. Importantly, we note that the appropriate significance threshold depends on both the number of ancestral populations and the population structure of the sample. Our findings hold many practical implications for researchers regarding best practices for admixture mapping studies in admixed populations with population structure.

44 | Modifiable risk factors and Parkinson's disease: systematic Mendelian randomization studies

Sandeep Grover[1], Fabiola Del Greco M.[2], Chistina Lill[3], Inke R König[1]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]Institute for Biomedicine, Eurac Research, Bolzano, Italy; [3]Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), Institutes of Neurogenetics & Cardiogenetics, University of Lübeck, Lübeck, Germany

Several risk factors have been tested for their causal role in Parkinson's disease (PD) using Mendelian Randomization (MR) approaches with protective associations reported for Body Mass Index (BMI) and serum iron levels. However, to date, a systematic approach in selection of risk factors and prioritisation of genetic instruments is lacking in literature. Furthermore, several genetic instruments have been updated in recent years in the latest genetic association reports. The aim of the current study is to explore the causal relationship between already reported and novel modifiable risk factors and PD in populations of European ancestry using a two-sample MR methodology. We follow a systematic approach to identify (a) potential risk factors in PD and (b) genetic variants validly associated with these risk factors based on the largest reported Genome-wide Association Studies (GWAS). Inverse variance weighted (IVW), MR Egger and weighted median methods are used to judge the causal effect of identified risk factors. Presence of pleiotropic variants is detected using the heterogeneity Q test and the I^2 index. Furthermore, sensitivity analyses are conducted on the basis of instruments comprising a subgroup of SNPs with known biological function for specific risk factors. In the absence of potential pleiotropic variants, the MR approach is believed to overcome the issues of confounding and reverse causality inherent in observational studies. This approach may help us to identify causal risk factors for PD and to better understand neurodegenerative pathways for the development of potential interventions.

45 | Quantification of genetic effects: the impact of model specification and misclassification

Felix Günther[1,2], Klaus Stark[1], Helmut Küchenhoff[2], Iris M. Heid[1]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany; [2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians-Universität Munich, Munich, Germany

Genome-wide association studies aim to identify genetic risk factors for complex diseases focusing on type-1 error control with little attention to potentially biased effect estimates. However, the adequate quantification of genetic effects is pivotal, for example, when calculating individual risk using genetic risk scores. For binary traits and logistic regression, several aspects are relevant to obtain unbiased effect estimates: inclusion of influential covariates, and consideration of missing information or misclassification. On the example of age-related macular degeneration (AMD), we illustrate the dependence of SNP effect estimation on modelling. We developed a maximum likelihood approach (MLA) that facilitates unbiased estimation of effects in logistic regression with outcome *Late AMD in worse eye*, even if single eye information was missing for some study participants and if there is misclassification in AMD diagnosis. Based on data of 1034 participants from the population-based German AugUR study, we estimated genetic effects of *CFH* and *ARMS/HTRA1* lead variants (rs1061170, rs10490924) using various models: (a) naïve logistic regression ignoring missing single eyes/misclassification with single SNPs as covariate; (b) adding age as the leading AMD risk factor as covariate, (c) MLA considering misclassification from missing "eyes"; (d) MLA with both SNPs and age as covariates, (e) model (4) with assumed sensitivity of 0.9 in single eye AMD-information. In these models (1)-(5), we estimated the ORs 1.91, 2.26, 2.35, 2.41, 2.46 for rs1061170 and 2.36, 2.37, 2.35, 2.48, 2.44 for rs10490924. We discuss assumptions of the approaches, reasons for differences in estimates and general implications for quantifying genetic effects.

46 | A weighted genetic risk score based on 279 signals of association with lung function predicts chronic obstructive pulmonary disease

Anna L. Guyatt[1], Nick Shrine[1], Phuwanat Sakornsakolpat[2,3], Victoria E. Jackson[1,4,5], Brian D. Hobbs[2,6], Michael H. Cho[2,6], Ian P. Hall[7], Martin D. Tobin[1,8], Louise V. Wain[1,8], on behalf of the SpiroMeta consortium

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, United States of America; [3]Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand; [4]Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia; [5]Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia; [6]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, United States of America; [7]Division of Respiratory Medicine and National Institute for Health Research – Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom; [8]National Institute for Health Research Biomedical Research Centre – Respiratory Theme, Glenfield Hospital, Leicester, United Kingdom

Lung function measures are used to diagnose and grade chronic obstructive pulmonary disease (COPD). We have shown that genetic determinants of lung function, as analysed in large population cohorts, are informative about COPD risk, and are a powerful alternative to case control studies.

An association was previously reported between COPD and an *unweighted* genetic risk score, comprising the 95 known variants associated with lung function at that time (OR per SD 1.36 [95%CI 1.30–1.43], $P\ value = 5.65 \times 10^{-36}$). In This study, we used newly discovered and previously reported signals to create a *weighted* score, aiming to create a better predictor of COPD.

Novel signals ($P\ value < 5 \times 10^{-9}$) were identified from genome-wide association analyses of spirometry in UK Biobank and the SpiroMeta consortium ($N = 404{,}165$). Previously reported signals ($P\ value < 1 \times 10^{-5}$) and novel signals were combined in a weighted genetic risk score, with weights obtained from studies not included in signal discovery. We tested the risk score in post bronchodilator defined COPD case control studies.

We combined 139 novel signals with 140 previously reported signals to construct a weighted score and tested for association with COPD in 5991 cases and 3378 controls. The mean (standard deviation [SD]) number of risk alleles per individual was 295 (10.4). The risk score was associated with COPD risk ($P\ value = 6.64 \times 10^{-63}$), with an OR of 1.55 (95%CI 1.47–1.63) per SD increase in the score.

Our findings show that increasing the number of signals in a weighted risk score increases predictive power for COPD, and provides new biological insight.

47 | A meta-analytic framework of a maximum score test for genetic association integrating a class of disease risk models for gene-environment interactions

Summer S. Han[1], Nilanjan Chatterjee[2]

[1]Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, United States of America; [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, United States of America

Current methods for detecting genetic associations lack full consideration of the background effects of environmental exposures. We previously developed a maximum score test for detecting genetic association, encompassing a broad range of risk models, including linear, logistic, and probit, for specifying joint effects of genetic and environmental exposures. However, this method is not applicable to a large consortium setting, where analysis is performed separately in each study site, hence only summary results are available. In this study, we develop a meta-analytic approach for the maximum score test that provides a combined genetic association result based on the summary outputs obtained across different studies. We extend the previously developed maximum score test that obtains the test statistics by maximizing over a range of index parameter thetas (i.e. over a class of score tests), each of which reflects the potential heterogeneity of the genetic effects by levels of environmental exposures under a particular disease risk model. To obtain a meta-analytic p value of the maximum score test, we collect the score and its variance for each theta as well as the correlation matrix for the scores across thetas. We then calculate the meta-analytic p value of the maximum score test using an extreme value theory related to estimating a tail probability of the maximum of stochastic processes, which requires the summations of the scores, variances, and the correlation matrices across different studies and varying thetas. Simulation studies demonstrate the robust power of the proposed method for detecting genetic associations under a wide range of scenarios. We apply the proposed method to genome-wide association studies data for lung cancer and Alzheimer's disease.

48 | Interactions between folate intake and genetic predictors of expression associated with colorectal cancer risk

Cameron B. Haas [1,2], Ulrike Peters[1,2], Colon Cancer Family Registry (CCFR), Colorectal Transdisciplinary Study (CORECT), and Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, United States of America; [2]Department of Epidemiology, University of Washington, Seattle, United States of America

Observational studies have shown increased folate consumption to be associated with lower risk of colorectal cancer (CRC), but the complex mechanism remains unclear. Previous genotype by environment interaction (G×E) studies have been limited by statistical power and sample size. Set-based SNP testing has the potential to increase statistical power to detect GxE interactions by aggregating functionally relevant SNPs.

We incorporated functional information from the transcriptome prediction tool, PrediXcan, into our novel set-based approach for testing G×E interactions in two independent datasets. We used variant weights from the PrediXcan models of colon tissue-specific gene expression as a priori variant information for a set-based GxE approach.

We used sex and study specific quantiles based on calculated total folate (mcg/day) harmonized across studies. Analyses were performed using the Mixed Effects Score Tests for interactions, adjusting for sex, age, study, total energy intake, and principle components. The discovery data set included 4,839 CRC and advanced adenoma cases and 5,884 controls of European ancestry. Replication was performed in 8,636 cases and 8,032 controls. Subsequently, results from discovery and replication were used in a pooled meta-analysis.

Although none of the top genes from the discovery phase were replicated, we detected several suggestive interactions in the discovery data set which remained above a false discovery rate of 0.2 in the meta-analysis (CREB1, AGA, SLC28A3, and GSTA1). Future work will replicate these procedures for dietary and supplemental folates in separate analyses, as the biological mechanisms could differ.

49 | Identifying hidden ancestries in publicly available summary data

Kendra Koach[1], Megan M. Sorenson[1], Tiffany Dinh[1], Yinfei Wu[1], Christopher R. Gignoux[2], Audrey E. Hendricks[1]

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

Recent summary level genetic data has enabled advancements in understanding genetics, health, and disease. Summary test statistics from Genome-wide Association Studies (GWAS) are particularly useful for secondary analyses and data dissemination. Publicly available databases of allele and genotype frequencies are rich resources for looking up potentially causal variants in studies of rare diseases and as controls in rare variant tests given new methods. Ancestries provided in the summary data are often broad (e.g. European, African, Asian) and do not include information on finer scale ancestries (e.g. British, Italian) or admixture (e.g. African Americans within African summary groups). Lack of precise ancestry information can cause confounding in secondary or association analyses and can contribute to variant misclassification in molecular diagnosis of rare diseases.

To address this, we have developed a modified Expectation-Maximization (EM) algorithm that estimates the proportion of finer scale ancestry within publicly available data (e.g. GWAS, genotype databases). Unlike traditional individual level ancestry estimation methods, our method only requires allele or genotype frequencies. We are able to accurately and precisely detect the underlying ancestry proportions across a wide variety of simulations. Specifically, we have sensitivity to detect ancestries that make up only 1% of the summary data. We further apply our method to NHLBI's Trans-Omics for Precision Medicine (TopMED) BRAVO interface allele frequencies and genome Aggregation Database (gnomAD) African and Latino groups identifying hidden ancestries not provided in the summary data. Our method has the ability to provide precise ancestry information for the continually growing genetic resources reducing inaccurate results due to unknown or hidden ancestry information.

50 | Imputation of complex biological data for Bayesian network analyses

Richard Howey[1], Heather J. Cordell[1]

[1]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom

Bayesian networks have been proposed as a way to identify possible causal relationships between measured variables based on their conditional dependencies and independencies, in particular, in complex scenarios with many variables. When there is missing data, the standard approach is to remove every individual with missing data

before performing any Bayesian network analysis. This can be wasteful and undesirable when there are many individuals with missing data, perhaps with only one variable missing, making imputation a natural choice. We present a new imputation method designed to increase the power to detect causal relationships whilst accounting for model uncertainty. This method uses a version of nearest neighbour imputation, whereby missing data from one individual is replaced with data from another individual, the nearest neighbour. An important feature of this approach is that it can be used with both discrete and continuous data. For each individual with missing data, subsets of variables that can be used to find the nearest neighbour are chosen by bootstrapping the complete data to estimate a Bayesian network. We show that, with the use of our imputation method, the power to detect the correct model in simulated data can be increased by as much as 50%. Such increases may be possible in real data when many individuals have missing data due to cost or practical reasons. Thus, the use of our imputation method has great potential to boost the power of Bayesian network analyses to identify possible causal relationships.

51 | Meta-analysis of ~1.3 M individuals identifies rare variants associated with blood pressure and implicates causal genes

Joanna M. M. Howson[1] on behalf of the Blood Pressure-International Consortium of Exome chip Studies (BP-ICE), CHD Exome+ consortium, CHARGE, Exome BP, GoT2D, T2Dgenes, deCODE and the Million Veteran Program

[1]BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

Elevated blood pressure (BP) is the primary modifiable risk factor for cardiovascular diseases. Worldwide, ~10.7 M deaths/year have been attributed to complications from raised BP. Therefore, improving the understanding of BP regulation in the population is of pivotal importance. Genome-wide association studies have identified common BP-associated SNPs but rarely pinpoints causal genes. In contrast, identification of rare coding BP-associated risk variants (MAF < 0.01) could help identify causal genes.

We tested ~240,000 variants on the exome array and ~8 M imputed variants in up to ~865,000 individuals for association with systolic BP, diastolic BP, pulse pressure, and hypertension. Results for variants with $P < 1 \times 10^{-7}$ were meta-analysed with up to ~448,000 additional individuals. Variants with $P < 5 \times 10^{-8}$ in the full meta-analysis of ~1.3 M individuals were considered significant. We evaluated a number of statistical approaches to identify independent associations in novel and known BP regions and performed gene-based tests, which aggregate rare variants in the exons of genes.

Over 30 novel rare variant-BP associations were identified including ten missense variants. Four genes (*DBH*, *COL21A1*, *NOX4* and *NPR1*) at known loci had associations consistent with multiple rare variant associations. Within known regions, there were ~40 new BP-rare variant associations, some of which were located in well-known BP genes (e.g. *AGT*) and others were located in novel genes not previously implicated in BP regulation. Integration of functional data, highlighted additional candidate genes. Our study demonstrates that rare BP variants can have larger effects than common BP variants and can provide insights into candidate causal genes and pathways.

52 | Mating asymmetry in an international orofacial cleft study

Julie Hudson[1], Mengche Tsai[2,3], Kelly M. Burkett[1], Marie-Hélène Roy-Gagnon[3]

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada; [2]Department of Pediatrics, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan; [3]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada

Mating asymmetry (MA) has been observed in human populations due to cryptic population forces such as assortative mating or sex specific migration. However, MA is not the only reason for differential mating type frequencies in case-parent trios, as maternal genetic effects or imprinting effects also alter mating frequencies. MA thus confounds maternal genetic effect tests in samples of case-parent trios where mating symmetry has to be assumed to test for maternal genetic effects. A combined analysis of MA and maternal/imprinting effects has not been conducted in studies without controls so the extent of the problem is unknown.

We used the Bourgey et al. 2011 mating asymmetry statistic (MaS) to estimate genome-wide levels of MA in orofacial cleft case-parent trios from an international consortium (Beaty et al. 2010). MaS measures deviation from expected symmetry in mate-pair frequency among the six informative mating types through the estimation of three asymmetry parameters. We observed different levels and patterns of differential mating type frequencies depending on ancestry: 29 SNPs had moderate to high MaS when all populations are considered together, and more high MaS SNPs exist in principal component

defined subgroups of Asian and European ancestries. This increase in MA within subgroups could be as a result of population-specific asymmetry or due to smaller sample sizes. Likelihood ratio tests were also conducted to determine the significance of high MaS SNPs. We also related MaS values to maternal genetic tests and imprinting tests to detect potential false positives. Our results support previous identified ancestry-specific asymmetry.

53 | A sex-specific genome-wide association study identifies novel loci associated with end-stage renal disease in chinese patients with type 2 diabetes

Guozhi Jiang[1,2], Claudia H.T. Tam[1,2], Andrea O.Y. Luk[1,2,3], Heung Man Lee[1,3], Cadmon King Poo Lim[1,3], Wing Yee So[1,2], Juliana Chung Ngor Chan[1,2,3,4], Ronald Ching Wan Ma[1,2,3,4] on behalf of the Hong Kong Diabetes Biobank and TRANSCEND Consortium

[1]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China; [2]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China; [3]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China; [4]Integrated Bioinformatics Laboratory for Cancer and Metabolic Diseases, The Chinese University of Hong Kong, Hong Kong, China

End-stage renal disease (ESRD) is a major cause of morbidity and mortality in patients with type 2 diabetes (T2D) worldwide. Both gender and genetic variants influence the risk of developing ESRD in patients with T2D. In this project, our aim is to identify genetic variants associated with ESRD in Chinese patients, including any sex-specific genetic variants. We performed a sex-stratified meta-analysis of genome-wide association study (GWAS) in 2047 ESRD cases (1037 men, 1010 women) and 7654 controls with duration of T2D >10 years and free of ESRD (3865 men, 3789 women) from the cohorts of the Hong Kong Diabetes Register (HKDR) and the Hong Kong Diabetes Biobank (HKDB). ESRD was defined as dialysis or transplant, or eGFR<15 ml/min/1.73 m^2. We imputed the genotypes using minimac 3, with 1000 genome project Phase III data as reference panel. Within gender, ~6.6 million high-quality common variants (minor allele frequency [MAF] $\geq$1%) were included in logistic regression, adjusting for age and principal components. In women, a total of 210 variants were associated with ESRD at threshold of $P<10^{-5}$ from 12 distinct genomic regions on chromosomes 3, 4, 6, 7, 11, 12, 13, 15 and 19. The strongest signal, which associated with ESRD in women (OR= 1.48[95% CI: 1.29–1.71], $P=5.8\times10^{-8}$) but not in men ($P=0.29$), falls in the exon region of 15q13.1. In men, in total 198 variants were identified with suggestive association with ESRD, with one of the strongest signal from a SNP located at 2q24.1 (OR = 1.59 [95% CI = 1.33–1.90], $P=2.8\times10^{-7}$ in men; $P=0.98$ in women). No overlap significant regions were identified between men and women, and evidence of heterogeneity of effects for top signals was presented between men and women ($P_{het}<0.05$). Our study has identified a number of novel regions associated with ESRD in T2D, and has highlighted the difference of genetic effects between men and women. Replications and extension of these findings are currently in progress.

54 | Estimating the effects of copy number variants on intelligence quotient using hierarchical Bayesian models

Lai Jiang[1], Guillaume Huguet[23], Catherine Schramm[123], Sebastien Jacquemont[23], Tomas Paus[4], Gunter Schumann[5], Patricia Conrod[235], Zdenka Pausova[6], Celia M.T. Greenwood[17]

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2]Center Hospitalier Universitaire Sainte-Justine Research Center, Montreal, Canada; [3]Department of Pediatrics, Université de Montréal, Montreal, Canada; [4]Departments of Psychology and Psychiatry, University of Toronto, Toronto, Canada; [5]Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom; [6] The Hospital for Sick Children, University of Toronto, Toronto, Canada; [7]Departments of Human Genetics and Oncology, McGill University, Montreal, Canada

In pediatric genetics clinics, many children are observed to carry genetic copy number variants (CNVs) that are either rare in the population, or that have never been previously observed. Predicting the effects of such rare CNVs on cognition is challenging. Recently, Huguet et al. (2018, JAMA Psych) showed, in two general population cohorts, that verbal and performance intelligence quotient scores (VIQ/PIQ) could be predicted by a linear model containing the sum of loss-of-function intolerance (pLI) scores for all genes deleted by the individual's CNVs. However, this model is unable to estimate individual CNV effects. Therefore, here we propose hierarchical Bayesian models to estimate the effects of individual CNVs on IQ in a hierarchical framework.

The effects of each CNV, β_j, were assumed to contribute additively to IQ across all CNVs carried by an individual. The effects β_j were then assumed, in a hierarchical manner, to depend on annotation scores applicable to CNV j. To improve convergence and performance, we explored winsorizing and de-correlating the annotation scores, as well as sigmoid transformations of effect sizes.

We find some annotation scores that are strongly associated with the CNVs' impact on IQ, particularly the Deletion Score and a score derived from expression quantitative trait loci (eQTL) knowledge. Furthermore, several regions in the genome are identified as possibly containing CNVs with particularly strong impact on IQ. However, the Bayesian models explain a lower proportion of the variance in IQ than the linear model. Validation of initial results is underway in another general population cohort.

55 | Understanding of DNA methylation in the biological basis of stress related cardiovascular disease

Rong Jiang[1], Lydia Kwee[2], Svati H. Shah[2], Abanish Singh[1], Michael A. Babyak[1], Beverly H. Brummett[1], Ilene C. Siegler[1], Redford B. Williams[1], Elizabeth R. Hauser[2]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, United States of America; [2]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, United States of America

DNA methylation potentially links environmental, genetic and metabolic risk factors for cardiovascular disease. Our previous studies have replicated an interaction between two SNP genetic variants (G) on chromosome 19q13.32 in the TOMM40-APOE region and psychosocial stress (S) associated with an adverse lipoprotein profile in stressed caregivers, but not in non-stressed controls. This study tests whether this interaction (G×S) may be due to differential DNA methylation at three levels: average methylation in the candidate region (AMCR), average global methylation (AGM) and genome-wide, probe-specific methylation (EWAS). Methylation data were assayed for two samples: CAREGIVER ($N = 304$) and CATHGEN ($N = 1038$) using HumanMethylation450 BeadChip and MethylationEPIC chip. AMCR was calculated across 103 kbp in NECTIN2-TOMM40-APOE-APOC region and AGM was computed over all sites. S was measured by caregiving stress or a constructed stress components. All association tests of main effects and G×S interaction were controlled for age, gender, race, cell counts, and batch effects. A weak G×S was observed with AMCR ($P = .025$–0.185) and AGM ($P = .035$–0.164). The effect of SNPs on methylation levels was observed in stressed individuals only; the pattern was consistent with the relationship between SNP and lipoprotein levels. The EWAS top hit for SNP effect (cg17928676, FDR $= 3.1 \times 10^{-12}$-9×10^{-43}) was in 19q13.32 and replicated in both samples. These findings indicate that the SNP may moderate the association of S through methylation, and have an impact on the lipoprotein phenotypes. This study links our genetic interaction findings with a functional consequence and suggests DNA methylation links stress and related disease susceptibility.

56 | ELECTRONIC MEDICAL RECORDS AND GENOMICS (eMERGE) – using emr-linked biorepositories to expand genomic medicine research

Sheethal Jose[1], Jyoti Dayal[1], Kenneth L. Wiley[1], Rongling Li[1], eMERGE Network

[1]Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, United States of America

The Electronic Medical Records and Genomics (eMERGE) Network, a consortium funded by the National Human Genome Research Institute (NHGRI) since 2007, leverages biorepositories linked to electronic medical records (EMRs) for large scale genomic research, develops sharable tools, and establishes best practices for implementing genomic medicine. During eMERGE Phases I and II, the Network demonstrated that EMR systems and biorepositories can serve as resources for genome-wide association studies (GWAS) of diseases and quantitative traits. eMERGE investigators validated that the phenome-wide association study (PheWAS) paradigm using real world EMR-linked genetic data could replicate 66% of known GWAS findings in the NHGRI Catalog. In eMERGE Phase III, the Network, consisting of 12 academic medical centers, is sequencing 109 clinically relevant genes and 1,552 SNPs in 25,000 individuals. Clinically actionable genetic variants are being returned to patients and their clinicians for use in their care and other uncertain variants are being analyzed for associations with clinical phenotypes. During the return of results process, the Network is collecting various outcomes data, such as changes in healthcare utilization, biomarker indicating benefit/harm, whether knowledge of a variant changes a patient's diagnosis and treatment, risk in family members, and so forth. All sites are also investigating the ethical, legal and social implications of incorporating genomic information into EMRs. The Network published that Institutional Review Boards (IRBs) are primarily concerned about incidental findings, informed consent, placement of results in EMR, and contacting family members. However, communicating with IRB staff during protocol review and including supplemental materials help to address their concerns.

57 | Machine learning in multi-omics data to assess longitudinal predictors of glycaemic trait levels

Marika Kaakinen[1,2], Laurie Prelot[1], Harmen Draisma[1], Mila D. Anasanti[1], Marjo-Riitta Jarvelin[3,4], Inga Prokopenko[1]

[1]Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, United Kingdom; [2]Centre for Pharmacology and Therapeutics, Department of Medicine, Imperial College London, London, United Kingdom; [3]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [4]Center for Life Course Health Research, University of Oulu, Oulu, Finland

Multi-omics data hold enormous potential for personalised medicine; however, analysis of high-dimensional data poses challenges. Type 2 diabetes (T2D) is a global health burden that will benefit from personalised risk prediction and tracking of disease progression. We aimed to identify longitudinal predictors of glycaemic traits relevant for T2D by applying machine learning (ML) approaches to multi-omics, including epigenetic and metabolomic data, from the Northern Finland Birth Cohort 1966 at 31 (T1) and 46 (T2) years. We predicted fasting glucose/insulin (FG/FI), glycated haemoglobin (HbA1c) and 2-hr glucose/insulin (2hGlu/2hIns) at T2 in 595 individuals using 1,010 variables from T1 and T2: body-mass-index (BMI), waist-hip-ratio, sex; 10 plasma measurements; 453 NMR-based metabolites; 542 methylation probes established for BMI/FG/FI/HbA1c/T2D/2hGlu/2hIns and measured with Illumina Infinium-HumanMethylation450 BeadChip (T1)/EPIC (T2) arrays. We used six ML approaches trained on 80% and tested on 20% of the data: random forest (RF), boosted trees (BT) and support vector machine (SVM) with the kernels of linear/linear with L2 regularization/polynomial/radial basis function. RF and BT showed consistent performance while SVMs struggled with higher-dimensional data. The predictions worked best for FG and FI. T2 branched-chain amino acids, HDL-cholesterol and body measurements already at T1 were amongst the most important predictors. Addition of methylation data did not improve the predictions; however, BMI-associated methylation probes T1_cg26361535 at *ZC3H3*/T1_cg00634542 at *SLC11A1* were within the top 5% of predictors of FI/FG variability. With ML we could narrow down hundreds of variables into a clinically relevant set of predictors and demonstrate the importance of longitudinal traits in prediction.

58 | Joint genetic factors of body mass index and ADHD components

Ville Karhunen[1], Petri Wiklund[2], CAPICE consortium, Marjo-Riitta Järvelin[1,2,3], Alina Rodriguez[1,4]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom; [2]Center for Life Course Health Research, University of Oulu, Oulu, Finland; [3]Biocenter Oulu, University of Oulu, Oulu, Finland; [4]School of Psychology, University of Lincoln, Lincoln, United Kingdom

There is evidence for genetic overlap between body mass index (BMI) and attention-deficit/hyperactivity disorder (ADHD), however the difference in the association between two components of ADHD – hyperactivity and inattention – remains unclear. We examined the effect of BMI polygenic risk score (BMI-PRS) on ADHD and its separate components in the Northern Finland Birth Cohort 1986 ($N = 2916$). ADHD-related questions were answered by teachers and parents at eight years and by adolescents and parents at 16 years. Based on age, respondent and ADHD component, we constructed 19 outcome variables. We generated the BMI-PRS based on external summary statistics. We used ordinal regression to examine the effect of BMI-PRS on the outcomes, adjusted for maternal education, pre-pregnancy BMI and offspring sex. We also examined BMI-PRS by sex interaction effects. We found evidence of association between BMI-PRS and ADHD-related phenotypes for 13 of the 19 outcomes examined (false discovery rate (FDR) adjusted P value < 0.05). The effects were the strongest at eight years, and similar for both hyperactivity and inattention dimensions. There was some evidence for effect-modification by sex, with boys having stronger effect sizes especially for inattention-related outcomes. These results suggest similar effects of BMI-PRS on both ADHD components.

59 | Geographic genetic ancestry associates with uterine fibroid traits in African Americans from the bioVU resource

Jacob M. Keaton[1,2,3], Jacklyn N. Hellwege[1,2,3], Ayush Giri[1,2,3,6], Michael Bray[2], Sarah H. Jones[4], Katherine E. Hartmann[5,6], Ky'Era Actkins[2,7], Lea K. Davis[2,8], Todd L. Edwards[1,2,3], Digna R. Velez Edwards[2,3,6]

[1]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, United States of America; [2]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, United States of America; [3]Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, United States of America; [4]Vanderbilt Epidemiology Center, Vanderbilt University, Nashville, United States of America; [5]Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, United States of America; [6]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, United States of America; [7]Department of Microbiology, Immunology, and Physiology, Meharry Medical College, Nashville, United States of America; [8]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, United States of America

The burden of uterine fibroids (UF) disproportionately impacts African American (AA) women. Evidence suggests AA women have earlier onset and higher cumulative risk. The contribution of genetic ancestry to UF risk is unclear. We investigated ancestry proportions for the 1000 Genomes Phase 3 populations ($n = 26$) for association with UF traits in AA subjects from a large electronic health record biorepository (BioVU, $n = 583$ cases, 797 controls). Global ancestry proportions were estimated using ADMIXTURE software. Dichotomous outcomes fibroids status (i.e. case/control) and fibroid count (i.e. 1 vs > 1) and continuous outcomes $\log_{10}$-transformed fibroid volume and largest dimension were modeled for association with ancestry proportions using logistic and linear regression, respectively, adjusting for age. Effect estimates are reported per 10% increase in genetically inferred ancestry proportion. Esan in Nigeria (ESN) ancestry was associated with UF risk (OR = 1.12, 95% CI = 1.05–1.20, $P = 0.001$), although Luhya in Webuye, Kenya (LWK; OR = 0.84, 95% CI = 0.76–0.94, $P = 0.002$), Mexican Ancestry from Los Angeles USA (MXL; OR = 0.35, 95% CI 0.13–0.98, $P = 0.045$), and Toscani in Italia (TSI; OR = 0.78, 95% CI = 0.65–0.94, $P = 0.009$) ancestries were protective for fibroids. Utah Residents (CEPH) with Northern and Western European Ancestry (CEU; OR = 0.82, 95% CI = 0.68–0.99, $P = 0.046$) ancestry was protective for multiple fibroids. MXL ($\beta = -1.04$, SE = 0.47, $P = 0.028$) ancestry was associated with decreased fibroid volume, although Gujarati Indian from Houston, Texas (GIH; $\beta = 1.79$, SE = 0.54, $P = 0.001$) ancestry was associated with increased volume. Gambia in Western Divisions in the Gambia (GWD; $\beta = 0.15$, SE = 0.06, $P = 0.012$) and GIH ($\beta = 0.44$, SE = 0.20, $P = 0.024$) ancestries were associated with increased largest UF dimension. These results suggest that fibroid risk may vary by geographic/genetic ancestry. Further investigation at the local ancestry and single variant levels may yield novel insights about disease architecture and genetic mechanisms underlying ethnic disparities in UF risk.

60 | The power of the allele-based N-test in linkage analysis

Sajjad Ahmad Khan[1]

[1]*Department of Statistics, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa, Pakistan*

There are many tests of inheritance based upon sibling information for diseases that have late onset. The N-test (Green et al. 1983) is one of these tests, which utilizes information from affected siblings. The N-test is the count in affected siblings of the most frequently occurring haplotype from the father plus the analogous count from the mother. When applied to haplotypes, the N-test excludes recombinant families from the analysis. In this study, we modified the N-test to be based on alleles instead of haplotypes. This modified allele-based N-test can include all families (recombinant as well as non-recombinant). We carried out a simulation study to compare the power of the allele-based N-test with the powers of the S_{all} and S_{pairs} non-parametric statistics as computed by Merlin. The powers of the allele-based N-test, S_{all} and S_{pairs} statistics are identical to each other for affected sibships of size 2 and 3. For affected sibships of larger sizes, the powers of the S_{all} and S_{pairs} statistics are larger than the power of allele-based N-test. These simulation-based results are consistent with earlier results based on analytical computations.

61 | Absolute risk of pancreatic cancer in the U.S. general population

Jihye Kim[1], Chen Yuan[2], Ana Babic[2], Ying Bao[3], Meir Stampfer[1], Edward L. Giovannucci[1], Howard D. Sesso[1], JoAnn E. Manson[1], Brian M. Wolpin[2], Peter Kraft[1]

[1]*Harvard T.H Chan School of Public Health, Boston, United States of America;* [2]*Dana-Farber Cancer Institute, Boston, United States of America;* [3]*Brigham and Women's Hospital and Harvard Medical School, Boston, United States of America*

Pancreatic cancer has low incidence in the general population, thus a targeted approach that screens only individuals at high risk may be most effective. To identify high-risk individuals for pancreatic cancer, we analyzed 491 incident cases and 1,131 age-matched controls in four large US cohorts; the cases were diagnosed after blood draw between 1984 and 2010. We obtained risk factor data from blood samples or questionnaires completed around the time of blood collection. Using conditional logistic regression, we built relative risk models incorporating lifestyle and clinical factors (body mass index, waist-to-hip ratio, physical activity, periodontal disease, and diabetes), a genetic risk factor (incorporating 22 SNPs identified to be associated with pancreatic cancer in genome-wide association studies), and emerging biomarkers. We accessed risk prediction by calculating the Area Under the receiver operating characteristic Curve (AUC) with a 4-fold cross-validation and estimated cumulative absolute (lifetime) risks by combining risk estimates with age- and gender-specific incidence rates from the Surveillance, Epidemiology, and End Results (SEER) and U.S. mortality rates. The AUC of our model was 0.62, which is over previously published risk models ($AUC = 0.61$). The model identified 2% of men and 3% of women who had 3 times greater risk than the average risk in the general population and had maximum 3% lifetime risk assuming that they do not have the disease until age 80. Our prediction models utilizing a broad range of risk factors

successfully identified a subset of individuals who could most benefit from screening for pancreatic cancer.

62 | Kernel-based tests for very rare variants

Stefan Konigorski[1], Christoph Lippert[2]

[1]Molecular Epidemiology Research Group, Max Delbrück Center (MDC) for Molecular Medicine in the Helmholtz Association, Berlin, Germany; [2]Statistical Genomics Research Group, Max Delbrück Center (MDC) for Molecular Medicine in the Helmholtz Association, Berlin, Germany

Kernel-based gene-level tests are a powerful tool for genetic association studies of rare variants. They allow to decrease the multiple testing burden compared to variant-level tests, provide robust inference when testing the association of very rare variants with binary as well as quantitative traits (Lee et al., 2012, DOI:10.1093/biostatistics/kxs014), and can incorporate the effects of variant sets in a gene in a flexible manner. Different tests have been proposed, but they still provide suboptimal performance for the analysis of sequencing data (Konigorski et al., 2017, DOI:10.1371/journal.pone.0178504), where the overwhelming majority of genetic variation is extremely rare (i.e., singletons and doubletons), and where single-point mutations are of interest. As a result, there is little consensus on how to incorporate very rare variants. In this study, we propose a new formulation of kernel-based tests that generalizes established and novel kernels and allows to incorporate prior information from various annotations, To increase the power to detect associations with very rare variants. The investigated annotations include genomic locations and distances as well as functional annotations, and they are evaluated in a cross-sectional analysis of the ADNI (Alzheimer's Disease Neuroimaging Initiative) data to identify loci associated with Alzheimer's disease. The results indicate that a mix of the genomic descriptives and functional annotations can yield the highest power increase, and demonstrate that novel kernels and the integration of domain-specific knowledge constitute a promising approach for the analysis of very rare variants.

63 | Genome-wide haplotype association studies: comparison of novel methods

Björn-Hergen Laabs[1], Inke R. König[1]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein Campus, Lübeck, Germany

Standard genome-wide association studies (GWAS) based on single SNPs are known not to exploit the entire genetic information, thus possibly missing relevant genetic associations. Important gaps might be filled analyzing multiple SNPs simultaneously by the analysis of genetic interactions or haplotypes. The latter are particularly promising in that a haplotype not only contains the information of the SNPs themselves but also of their interactions and unmeasured loci between them.

One of the disadvantages of haplotype-based GWAS over SNP-based GWAS is that they need preprocessing to be applicable. For example, it is necessary for the most haplotype-based approaches to determine haploblocks which build different haplotypes. Another challenge for haplotype-based GWAS in the past was the handling of rare haplotypes. Most of the older methods avoided the resulting statistical problems by summing up all rare haplotypes in one category. However, especially these rare haplotypes could carry information of rare functional variants. In the last few years, a number of new methods for haplotype-based GWAS have been proposed which are able to estimate the association with rare haplotypes. In this contribution, we compare some of the methods which are able to deal with rare haplotypes with regard to the power to detect functional haplotypes. Additionally, we use one of the standard approaches (haplo.glm) that pools rare haplotypes using a fixed threshold as reference. We apply these methods to a GWAS data set with given pre-specified haploblocks and compare the findings with the results of earlier studies using haplotype-based GWAS.

64 | Exploring the effect of parental height on a newborn's birth weight

Yunsung Lee[1], Håkon K. Gjessing[2,3], Astanand Jugessur[1,2,3], Per M. Magnus[3]

[1]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway; [2]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway; [3]The Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

Parental height and newborn's birth weight are associated. To understand the underlying etiology, we used parental polygenic risk scores for height as instrumental variables. Genotype data from 10,496 complete newborn parent trios were available from the "HARVEST" study, which is part of the Norwegian Mother and Child Cohort Study. We identified 697 SNPs in HARVEST that were genome wide significant ($p < 5 \times 10^{-8}$) for height, and these were used to generate a polygenic risk score (PRS). The birth weight (gram) of newborns was extracted from the Medical Birth Registry of Norway. Two stage least square regression was used, with and without adjusting for the newborn's own PRS. The effect of mother's height

(cm) on newborn's birth weight (gram) with and without adjustment for the newborn's own PRS for height were 12.1 (95% CI: 7.7, 16.5) and 20.5 (95% CI: 16.9, 24.1) respectively. The father's height also increased the newborn's birth weight (12.8; 95% CI: 9.3, 16.2), but this effect substantially decreased after adjusting for newborn's PRS (3.3; 95% CI: −0.8, 7.3). Maternal height increases newborn's birth weight independently of the newborn child's PRS. By contrast, the effect of paternal height was attenuated after adjusting for the newborn child's PRS. These results, with some assumptions, suggest that maternal height is causally related to fetal growth.

66 | Using external information to enhance the power of genome wide gene-environment interactions scans

Juan Pablo Lewinger[1], W. James Gauderman[1], David V. Conti[1], John L. Morrison[1], Andre Kim[1]

[1]University of Southern California, Department of Preventive Medicine, Keck School of Medicine Los Angeles, United States of America

The state-of-the art in methods for performing genome wide associations scans of gene environment interactions (G×E) are two step-approaches that perform an initial screening of SNPs based on marginal exposure-genotype associations and/or marginal genotype-trait associations, followed by a formal test of G×E interaction in a second step. Despite dramatic power gains of two-step approaches over conventional single step methods, very large sample sizes are still required to detect the modest G×E effect sizes expected for most traits. We investigate methods that use "external information" to improve power for detecting G×E interactions in a two-step testing framework. Such external information may consist, for example, of functional SNP annotations or prior gene expression studies in humans or model organisms. We desire a method that enhances power when the external information is predictive of G×E interactions while maintaining robustness when it is not. Through extensive numerical power calculations, we compare approaches to aggregate sources of external information into a single "functional" score and robustly use the score to enhance two-step genome wide GxE scans. For example, in a scenario where a two-step method has 60% power to detect a G×E interaction, our approach that utilizes the functional score increases the power to over 75% power, even though the score is only mildly predictive of a GxE (area under the receiving operating curve is 0.6.)

67 | Exome-chip association study of refractive error in U.S. caucasians

Deyana D. Lewis[1], Ishika Jain[1], Theresa Alexander[2], Anthony M. Musolf [1], Dwight Stambolian[3], Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America; [2]National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, United States of America; [3]Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States of America

Refractive error occurs when light fails to properly focus on the retina of the eye. In the case of myopia, the light is focused in front of the retina, causing blurry distance vision. Myopia is defined in terms of refractive error (RE), a quantitative trait measured in Diopters (D), such that individuals with RE of -0.5D or more negative generally require glasses to correct their vision. Individuals with high myopia (RE< -6D) are vulnerable to ocular complications later in life. Great efforts have been undertaken to identify and understand the mechanisms underlying the development and progression of myopia. Some environmental exposures contribute to risk and genome-wide association studies (GWAS) and linkage studies have identified loci underlying the distribution of refractive error and influencing the risk of myopia. However, few causal variants have been definitely identified and a large proportion of the variation in refractive error remains unexplained. Therefore, this study aims to determine whether rare variants may account for some of the variation of refractive error.

Here we present results of a single-variant association test for a quantitative trait, refractive error, in 1562 control individuals from the Age-Related Eye Disease Study (AREDS). After quality control, close to 100,000 variants were analyzed and EMMAX single-marker analysis identified 18 rare variants that were genome wide significant ($p < 5 \times 10^{-8}$). Interesting candidate variants were found in genes *NIPSNAP2*, *TSSK3* and *IDH1*. We are extending the analysis to gene-based tests and will present these results.

68 | QC software for analysis of sequence data in family-based studies

Qing Li[1], Joan E. Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America

In recent years, with the increased popularity of DNA sequence data, the research community has paid more

attention to family based studies of complex traits. Pedigree data pose new challenges in the quality control steps for sequence variants. Not only do we need to screen data based on various genotyping calling metrics, we also need to use the pedigree structure to detect any markers with inconsistent inheritance patterns. In This study, we present our quality control (QC) pipeline software, written in R, (a commonly used statistical software language), for sequence data analysis. It includes functions to extract various genotyping calling measurements, to detect Mendelian inconsistency, and to detect any excessive allele sharing due to remote inbreeding. We used our software to conduct quality control on whole exome sequence data on odd numbered chromosomes in extended pedigrees from the Genetic Analysis Workshop 18 (GAW18). We also simulated Mendelian errors and genotyping errors in the sequence data to evaluate our QC functions. Our simulations showed the functions can detect those errors with high accuracy and within a reasonable time frame. These softwares are freely available and will assist researchers in large family studies using whole exome and whole genome sequence data.

69 | Generalized linear discriminant analysis for high-dimensional genomic data

Sisi Li[1], Juan Pablo Lewinger[1]

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, United States of America

Because of its simplicity, linear discriminant analysis (LDA) performs well in classification problems when the sample size (n) is relatively small. However, standard LDA cannot be applied with high-dimensional genomic data (p>>n). We propose two different extensions of LDA for classification with high-dimensional data. The first extension is based on Scout, a family of regression and classification methods for high dimensional settings that uses a sparse estimate of the inverse covariance matrix of the features derived from the graphical least absolute shrinkage and selection operator (GLASSO). We show that introducing additional sparsity to estimate the difference in means between the classes results in improved predictive performance over Scout. The second extension of LDA, motivated by the connection between LDA and linear regression, is based on a 2-level hierarchical regression model regularized with ridge penalties. This version of LDA can also incorporate external information relevant to the classification of the classes, such as gene annotations or information from previous studies. The regularized hierarchical model can be efficiently fitted by applying standard ridge regression twice. Through simulation, we show that in a wide range of scenarios our proposed LDA approaches yield better prediction than existing methods extending LDA to the high-dimensional setting. We illustrate the methods with an application to the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) data.

70 | Genetic interaction analysis among oncogenesis-related genes revealed novel genes in lung cancer development

Yafang Li[1], Xiangjun Xiao[2], Olga Gorlova[1], Ivan Gorlov[1], Younghun Han[1], Rayjean Hung[3], Christopher I. Amos[2], OncoArray Consortium.

[1]Biomedical Data Science Department, Dartmouth Geisel School of Medicine, Lebanon, United States of America; [2]Baylor College of Medicine, Institute of Clinical and Translational Research, One Baylor Plaza, Houston, United States of America; [3]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

Cancer is driven by the accumulation of many genetic variants that may act epistatically during the development of the disease. To explore epistasis among oncogenesis-related genes in lung cancer development, we conducted a filtered pairwise genetic interaction analysis among 35,031 SNPs from 2027 oncogenesis-related genes. We first conducted an efficient epistasis analysis using "fast-epistasis" option in PLINK and then validated the signals using standard logistic regression interaction analysis in a discovery cohort comprising of 18,401 cases and 14,260 controls from OncoArray lung cancer consortium. Candidate SNP pairs were further submitted to replication study including 5,636 cases and 6,141 controls from two independent genome-wide association studies. SNP pairs with interaction p value $< 5 \times 10^{-10}$ in joint analysis were reported as significant signals. Genetic interactions in *RGL1:RAD51B* (OR = 0.44, p value $= 3.27 \times 10^{-11}$ in overall lung cancer and OR = 0.41, p value $= 9.71 \times 10^{-11}$ in non-small cell lung cancer), *SYNE1:RNF43* (OR = 0.73, p value $= 1.01 \times 10^{-12}$ in adenocarcinoma) and *FHIT:TSPAN8* (OR = 1.82, p value $= 7.62 \times 10^{-11}$ in squamous cell carcinoma) gene pairs were identified in our analysis. None of these genes have been identified from previous main effect association studies in lung cancer. The candidate gene pairs with interaction p value < 0.05 in replication study were further submitted to gene set enrichment analysis using the Ingenuity Pathway Analysis program. Potential pathways and gene networks underlying the identified genetic interactions were revealed in lung cancer development. The stratified epistasis analysis by histology subtypes also suggested differences in epistasis and

signaling pathways involved in lung cancer tumorigenesis. The reported study is one of the pilot studies to systematically explore the genetic interactions among oncogenesis-related genes in lung cancer development. The results display that genetic epistasis among cancer-related genes is a common mechanism involved in lung tumorigenesis and many latent genes contribute to lung cancer development through interacting with other modifier genes.

71 | A unified method for rare variant analysis of GxE interactions

Elise Lim[1], Han Chen[2,3], Ching-Ti Liu[1]

[1]Department of Biostatistics, Boston University, Boston, 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, 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, United States of America

Advanced technology in whole-genome sequencing has offered the opportunity to comprehensively investigate the genetic contribution, particularly rare variants, to complex traits. Many methods of rare variants analysis have been developed to jointly model the marginal effect but the method to detect gene-environment (GE) interactions is underdeveloped. Identifying the modification effects of environmental factors on genetic risk poses a considerable challenge. To tackle this challenge, we developed a unified method in detecting GE interactions of a set of rare variants using generalized linear mixed effect model. The proposed method can accommodate both binary and continuous traits in related or unrelated sample. Under this model, genetic variants, sample relatedness, and the GE interaction were modeled as random effects. We adopted a kernel-based method to leverage the joint information across the rare variants and implemented a variance component score test to reduce the computational burden. Our extensive simulation study shows that the proposed method maintains correct type I error and high power under various scenarios, such as differing the direction of main genotype effects and the number of variants in the model for both continuous and binary traits. We illustrated our method to test gene-based interaction with smoking on body mass index or overweight with Framingham Heart Study and replicated the *CHRNB4* gene reported in previous large consortium meta-analysis of SNP-smoking interaction. Our proposed set-based GE test is computationally efficient and is applicable to both binary and continuous phenotypes, while appropriately accounting for familial relationships.

72 | Gene-environment interaction with smoking on non muscle invasive bladder cancer size at the time of diagnosis

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

Urinary bladder cancer (UBC) is one of few cancers with an established gene-environment interaction, yet previous research has not distinguished between muscle invasive (MIBC) and non muscle invasive (NMIBC) groupings. However, not stratifying for these categories in genetic analyses can result in overlooking important biological mechanisms. Furthermore, little evidence is present for genetic associations with characteristics of NMIBC, such as tumour size, grade of the disease or patient's age. These are especially influential for the disease course and are more specific entities than broadly-defined outcomes of prognosis and/or recurrence.

First, we have carried out a genome-wide association study (GWAS) on 653 NMIBC cases for tumour (stage, grade, size) and patient (age) characteristics in the Bladder Cancer Prognosis Programme (BCPP) cohort. Following, we investigated if smoking status and/or smoking intensity interacts with the effect of discovered variants.

Out of 61 SNPs yielding genome-wide significance across outcomes of tumour size, stage, grade, and age in the GWAS of the BCPP cohort, 10 have reached P value < 0.05 for interaction with smoking, all with outcome of tumour size. Five of these SNPs were located in 6q14.1; two in 14q21.1; the rest were mapped to 1p31.3, 3p26.1, and 13q14.13.

In summary, our study suggests of interaction between genetic variance and smoking behaviour for increased NMIBC tumour size at the time of diagnosis in the BCPP cohort. These results may shed more light on one of the developing pathways of NMIBC.

73 | Tissue-wise sub-typing of complex trait based on genetics

Arunabha Majumdar[1], Na Cai[2,3], Claudia Giambartolomei[1], Huwenbo Shi[4], Jonathan Flint[5], Bogdan Pasaniuc[1,4]

[1]Department of Pathology & Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, United States of America;

[2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom; [3]European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom; [4]Bioinformatics Interdepartmental Program, University of California, Los Angeles, United States of America; [5]Brain Research Institute, University of California, Los Angeles, United States of America

Analyzing gene-expression and Genome-wide Association Studies (GWAS) data together can prioritize tissue/cell-type relevant to a complex trait which is often unknown. If multiple tissue/cell-type specific causal pathways underlie a phenotype, the phenotype can be classified into sub-types stratified by different causal mechanisms. For example, body mass index (BMI) can be regulated by genes expressed only in brain or adipose tissue, or both but with differential expression levels and an individual's BMI can be regulated more by genes specifically expressed in brain compared to adipose. We aim to learn about such sub-phenotype structure for a group of individuals based on their marginal phenotype and genotype data for sets of expression quantitative trait loci (eQTLs) each corresponding to a set of genes expressed in a tissue. We implement the expectation-maximization (EM) algorithm to estimate the posterior probability of an individual being assigned a sub-type (corresponding to a tissue). Our simulation study shows that, if such sub-phenotype structure exists, it can be retrieved meaningfully; and the accuracy of correctly inferring the subtype depends on the heritability of each subtype explained by the corresponding set of SNPs. We sub-typed BMI in the UK Biobank cohort. For cerebellum-brain and subcutaneous-adipose tissues, we obtained set of genes specifically expressed in each tissue and the corresponding LD-filtered eQTLs. In the analysis of BMI for a set of 150,000 people with their genotype data for tissue-specific eQTLs, based on 65% posterior probability cut-off, 11,000 individuals were assigned to cerebellum-brain and 9,000 individuals to subcutaneous-adipose. In summary, we present a novel approach to identify genetically defined sub-type of complex trait.

74 | Encouraging open science, replicability of analysis and collaborative cloud computing for whole genome sequence analysis of complex traits

Timothy Majarian[1], Alisa K. Manning[1,2]

[1]Metabolism Program, Broad Institute of MIT and Harvard, Cambridge, United States of America; [2]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, United States of America

The traditional model for data sharing and analysis breaks for whole genome sequence (WGS) association studies of complex traits and disease. By specifying a detailed analysis plan to be implemented at individual research sites, analysts must ensure their data is in the correct format, that software is compatible with their local high-performance computing infrastructure, that there is adequate space for data, and that there exists a mechanism for transferring results to a central repository for meta-analysis. Methods for WGS analyses are under active development, often leveraging distributed computing architectures not available on premises. Furthermore, WGS data sets and the resulting summary statistic files needed for meta-analysis are massive, rendering the traditional model of individual data downloads ineffective.

Collaborative cloud computing on the FireCloud platform brings researchers together within a shared workspace with transparent cost management, open-source analysis pipelines, batch analysis at scale, distributed computing on modern Spark architecture, and reproducible Jupyter notebooks. We have made analysis notebooks and pipelines for WGSAS of complex traits available through our open-source repository, including data conversion tools, common and rare variant association analysis with hail, EPACTS, GENESIS and SAIGE, functional fine-mapping with PAINTOR, and genomic enrichment analysis with LD Score Regression. We demonstrate the use of these notebooks and pipelines with 1000 Genomes data and simulated traits in a public FireCloud workspace. We have leveraged these tools in our analysis of NHLBI's TOPMed WGS data and are extending them for forthcoming massive data sets like the *All of Us* Research Program.

75 | Genetic simulation resources at your service: registration and certification of genetic simulation software

Bo Peng[1], Man Chong Leong[1], Huann-Sheng Chen[2], Melissa Rotunno[2], Katy R. Brignole[2], John Clarke[3], Leah E. Mechanic[2]

[1]Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, United States of America; [2]Division of Cancer Control and Population Sciences, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, United States of America; [3]Cornerstone Systems, Lynden, United States of America

Genetic Simulation Resources (GSR) (https://popmodels.cancercontrol.cancer.gov/gsr/) is an online catalogue of genetic simulators. The catalogue contains detailed information on features of the simulation programs and of the generated data and typical applications for more than 140 simulators. Search and comparison tools are provided to allow users to

easily identify the most suitable simulation tools for their specific research topics.

To further assist users of the catalogue and encourage better practice in the development, dissemination, and maintenance of genetic simulators, the GSR team rolled out a GSR Certification program to proactively appraise simulators based on criteria in the categories of accessibility, documentation, application, and support. (https://popmodels.cancercontrol.cancer.gov/gsr/certification/). The certificates are evaluated by the GSR committee with assistance from package authors. Certified packages are displayed more prominently on GSR and packages certified in all categories are invited to display a GSR certified logo on their websites.

The GSR Certification program provides a service to the community by evaluating genetic simulation software, with the ultimate goal of promoting the development and application of genetic simulation software. We encourage authors of genetic simulators to register and certify their packages and welcome investigators who are interested in genetic simulations to join our committee.

76 | Combined association of a polygenic risk score with 313 genetic variants and established environmental risk factors in relation to breast cancer risk

Pooja Middha[1,2], Nasim Mavaddat[3], Roger L. Milne[4,5], Jacques Simard[6], Marjanka K. Schmidt[7,8], Peter Kraft[9,10], Paul D.P. Pharoah[3,11], Douglas F. Easton[3,11], Montserrat Garcia-Closas[12], Jenny Chang-Claude[1,13] on behalf of Breast Cancer Association Consortium

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Faculty of Medicine, University Heidelberg, Heidelberg, Germany; [3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [4]Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Australia; [5]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia; [6]Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Canada; [7]Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [8]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands; [9]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, United States of America; [10]Program in Molecular and Genetic Epidemiology, Harvard T. H. Chan School of Public Health, Boston, United States of America; [11]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom; [12]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, United States of America; [13]Genetic Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Genome-wide association studies have identified multiple common breast cancer susceptibility variants. These, along with the environmental risk factors, will improve risk prediction of breast cancer to inform stratified prevention and screening. We evaluated the combined associations of a newly-derived polygenic risk score (PRS) including 313 single nucleotide polymorphisms (SNPs) with 13 established breast cancer environmental risk factors in women of European ancestry.

Data from the Breast Cancer Association Consortium (BCAC) were used for analyses and comprised two independent sets of 28,176 cases and 32,209 controls genotyped with the iCOGS array and 44,109 cases and 48,145 controls genotyped with OncoArray. Joint associations and multiplicative interactions between a 313-SNP PRS and each of 13 breast cancer risk factors were evaluated using logistic regression and a likelihood ratio test. Results from iCOGS and OncoArray were meta-analyzed using fixed effects model. We also performed global and tail-based goodness of fit tests in logistic regression models. The outcome was breast cancer, overall and by estrogen receptor (ER) status.

The global and tail-based goodness of fit tests showed no significant departure from the multiplicative risk model. The strongest evidence of multiplicative interaction with the 313-SNP PRS was found for current use of combined estrogen and progesterone therapy (interaction OR = 1.12, $P = 0.03$) and body mass index in premenopausal women (interaction OR = 0.95, $P = 0.009$) for overall breast cancer, and current smoking (interaction OR = 1.08, $P = 0.03$) for ER-positive disease.

Overall these analyses indicate that the 313-SNP PRS and the environmental risk factors are likely to have multiplicative effects on breast cancer risk.

77 | Analysis of the *CDKN2A* gene in FAMMM syndrome families reveals early age of onset for additional syndromic cancers

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

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America; [2]Hereditary Cancer Center, Creighton University, Omaha, United States of America; [3]The Rose Salter Medical Research Foundation, Newport Coast, United States of America; [4]Department of Hematology/Oncology, Creighton University, Omaha,

United States of America; [5]Department of Obstetrics and Gynecology, Creighton University, Omaha, United States of America

Familial atypical multiple mole melanoma (FAMMM) syndrome is a familial cancer syndrome that results from mutations in several genes including the *CDKN2A* gene. The syndrome is currently comprised of dysplastic nevi as well as melanoma, breast, pancreatic and lung cancer which are referred to as "concordant" cancers. As families with known *CDKN2A* mutations have been studied longitudinally, clinicians observed an abundance of other cancers or "discordant cancers." However, it was unknown whether these cancers were related to the syndrome. We sought to determine whether these discordant cancers also occur at higher frequencies and at earlier age of onset in carriers than non-carriers.

We studied 10 FAMMM syndrome families ($N = 1085$ individuals) in which a causal mutation in the *CDKN2A* gene was identified. We performed survival analysis as well as a mixed effects Cox Regression with age at follow-up or cancer event as our time variable and presence or absence of a concordant or discordant cancer as our censoring variable. The survival curves showed a significant age effect with carriers having a younger age at cancer onset for concordant (as expected) as well as discordant cancers than that of non-carriers. The cox regression models were also highly significant (P = 1.24E-27 and P = 5.00E-13 for the concordant and discordant cancers, respectively). These analyses support the hypothesis that carriers of mutations in *CDKN2A* in FAMMM syndrome have increased risk for early onset of several additional cancer types, suggesting that early screening for these cancers would be beneficial to carriers.

78 | Imputed expression of the Mendelian disease gene *SLC39A4* uncovers individuals at risk for zinc deficiency in biobank populations

Tyne W. Miller-Fleming[1], Xue Zhong[1], Jessica E.H. Brown[1], Eric Gamazon[1], Lisa Bastarache[2], Joshua C. Denny[2], Nancy J. Cox[1]

[1]Vanderbilt University Medical Center, Department of Medicine, Division of Genetic Medicine, and Vanderbilt Genetics Institute, Nashville, United States of America; [2]Vanderbilt University Medical Center, Department of Biomedical Informatics and the Center for Precision Medicine, Nashville, United States of America

Mendelian diseases arise from the loss of single genes and provide an opportunity to evaluate the efficiency of genetic tools, such as PrediXcan and phenome-wide association studies (PheWAS), to impute transcript levels and identify associations between gene expression and the phenome. Loss of the zinc transporter gene, *SLC39A4* results in the Mendelian disorder, Acrodermatitis enteropathica (AE). Based on the current understanding of AE, we can test the prediction that diminished *SLC39A4* expression is associated with hallmark symptoms of the disease, including blistered skin, diarrhea, anemia, and increased susceptibility to infections.

PrediXcan was applied to 23,000 Caucasian individuals in the Vanderbilt biobank, BioVU, which contains genetic data linked to electronic health records. PheWAS was performed to identify phenotypes associated with altered expression of *SLC39A4*. We identified significant associations between predicted *SLC39A4* expression levels and phenotypes consistent with AE (ulcerative colitis $p = 3 \times 10^{-5}$, anemia $p = 4 \times 10^{-3}$, *Staphylococcus aureus* infection $p = 3 \times 10^{-3}$). Interestingly, polymorphisms in *SLC39A4* occur more frequently in individuals of African descent. Using data from the Genotype-Tissue Expression project (GTEx), we identified a coding polymorphism found disproportionately in African American individuals (rs75920625, $AF_{EUR} = 0.005$, $AF_{AFR} = 0.14$) that significantly downregulates *SLC39A4* expression across all tissues ($p < .0001$).

These findings suggest that PrediXcan can model Mendelian disease gene expression and uncovers individuals presenting with Mendelian-like phenotypes without possessing the known monogenic disease. *SLC39A4* variants differ across populations and can significantly affect gene expression. Future work will examine whether predicted *SLC39A4* levels and susceptibility to AE phenotypes vary across populations. Because AE manifestations are alleviated upon zinc supplementation, nutraceutical intervention may be beneficial for individuals suffering from AE-like symptoms due to aberrant *SLC39A4* expression levels.

79 | Exploring genotype-environment (G × E) interactions phenome-wide

Rachel Moore[1,2], Inês Barroso[1], Oliver Stegle[2]

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

Phenome-wide association studies (PHEWAS) are a useful tool to study the effect of loci on multiple phenotypes, establishing links between biological processes or diseases not previously known to be related. However, such pleiotropy for genotype-environment (G × E) interactions has been poorly explored. Such

analyses may reveal greater aetiological overlap between traits than revealed through PHEWAS, or identify loci that are enriched for interaction effects. Consequently, these phenome-wide interaction studies may improve our understanding of biological processes involved in trait or disease risk.

To explore phenome-wide G×E effects, we used UK Biobank data ($N_{max} = 251{,}443$), focussing on a set of 57,328 variants (1,104 loci, 1MB window, LD $r^2 < 0.1$) significantly associated ($P < 5 \times 10^{-8}$) with basal metabolic rate (BMR), and a set of 64 lifestyle factors. Interaction effects were evaluated using StructLMM (biorxiv doi: 270611), a computationally efficient method for detection and characterisation of loci interacting with multiple environments.

We identify eight significant interaction loci for BMR (estimated heritability 0.29). In contrast, for other traits with lower heritability (< 0.27), including BMI and hip circumference, we detect >=30 interaction loci (conditional FDR 5%) and note that 32% of the interaction loci are shared across at least three phenotypes. Estimating the proportion of phenotypic variance explained by interaction effects across traits, at these 1,104 loci confirms differences in their interaction architecture. Namely, interaction effects explain 0.20-fold the amount of phenotypic variation explained by direct genetic effects for BMR, and 1.35-fold for diastolic blood pressure. Together, these results provide first insights of the G×E landscape across multiple phenotypes.

80 | G×E Scan: Software for Genome wide Discovery of G×E Interactions

John Morrison[1], Jim Gauderman[1]

[1]*Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, United States of America*

The discovery of gene-environment (G × E) interactions may help unravel the etiology of complex human traits, such as cancer, cardiovascular disease, and multiple sclerosis. We have developed an R Package, G × EScan for discovery of G × E interactions in a genome wide association study. The program implements several approaches, including the commonly used 1-degree-of-freedom (1-*df*) test of G × E, the 1-df test of marginal G association, and the 2-*df* joint test of G and G × E. It also implements several "2-step" approaches, which can be much more powerful for detecting G × E interactions. The program can analyze a disease or continuous trait, and the environmental factor can be categorical or continuous. Adjustment covariates can be included in the model. Measured or imputed SNPs can be analyzed, and we provide a utility to convert large genotype files into binary format for improved efficiency. The program utilizes RCpp, an R package that allows the incorporation of C++ code into an R program, leading to faster run times. For a given environmental factor E, a single call to the program can be used to produce 1-*df*, 2-*df*, and 2-step genome wide G×E results. Output includes text files of parameter estimates, test statistics, and *p* values for all analysis methods and all SNPs. Also produced is a single Excel file with Manhattan plots, QQ plots, and tables of top SNPs for each of the analysis methods. An additional utility program is available to extract SNPs within a specific genomic region for possible follow-up analysis. The program is freely available and can be downloaded from http://biostats.usc.edu/software.

81 | Using Bayes model averaging for admixture mapping

Lilit C. Moss[1], Xin Sheng[1], Christopher A. Haiman[1], David V. Conti[1], On Behalf of the African Ancestry Prostate Cancer Consortium and the Ellipse Game-On Consortium

[1]*Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, United States of America*

Admixture mapping is typically performed in recently admixed populations to detect genetic risk loci for diseases that have differential risk by ancestry. Admixture mapping relies on the deviation of local ancestry, alleles inherited from one ancestral population at a particular locus, from global ancestry, the proportion of the entire genome inherited from one ancestral population. The approaches most commonly used are case-only or case-control models which respectively compare ancestry deviations in cases only or test the effect of the deviation between cases and controls on the disease outcome. Although the case-only approach has potentially more power than the case-control approach, the case-only approach is prone to spurious associations when deviations of ancestry are not solely due to noise within controls. In this study, we Bayes model averaging from estimates obtained from case-only and case-control models to yield a novel statistical test for admixture mapping. The novel approach offers more power than a case-control method while remaining robust in scenarios where the case-only method is most susceptible to false positives. We use simulations to demonstrate that our approach detects admixture signals with increased power and robustness over case-control and case-only methods and illustrate the approach in the African Ancestry Prostate Cancer (AAPC) and the Ellipse Game-On Consortia.

82 | A meta-analysis of more than 237,380 men of diverse ancestries identifies 40 new risk loci for prostate cancer

Lilit C. Moss[1], Xin Sheng[1], Ed Saunders[2], Mark Brook[2], Tokhir Dadaev[2], Rosalind A. Eeles[2], Zsofia Kote-Jarai[2], David V. Conti[1], Christopher A. Haiman[1], On Behalf of the PRACTICAL Consortium

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, United States of America; [2]The Institute of Cancer Research, London, United Kingdom

Genome-wide association studies (GWAS) have identified over 150 loci associated with prostate cancer risk to date, primarily in populations of European ancestry. To identify additional loci associated with prostate cancer risk, we conducted a meta-analysis of ~30 million SNPs in 110,406 cases and 126,974 controls with individuals of African, European, Asian, and Latino ancestry. After excluding previously reported risk loci and regions of 800 kb length surrounding these markers, our meta-analysis identified over 40 new loci associated with overall prostate cancer risk at P value $< 5\times10^{-8}$. With these newly discovered risk loci and the previously reported loci, we examine the polygenic risk score, heritability and proportion of familial relative risk within and between populations. These findings demonstrate the value of combining populations with individuals of diverse ancestry for discovery of new loci, and provide new regions for fine mapping and for improving risk prediction. Additionally, these results will further guide future research into understanding the biological mechanisms that lead to prostate cancer.

83 | A novel robust statistical method for isoform quantification from RNA-seq data

Pronoy K. Mondal[1], Raghunath Chatterjee[1], Indranil Mukhopadhyay[1]

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

Technological advances trigger the generation of massively parallel genome-wide transcriptome data, known as RNA-sequencing data. A major problem to analyse such data is the correct quantification of transcripts. However, comprehending the distribution of reads, ambiguity in mapping reads to proper isoforms and so forth lead to problems in modeling and estimation of transcript abundance.

We develop a statistical method for estimation of isoform level abundance using maximum likelihood approach under general conditions of the nature and distribution of reads. Our likelihood function is multinomial type with indicators as latent variables. We adopt expectation-maximasation (EM) algorithm to obtain exact estimates and avoid approximations or plug-in estimates in maximizing the likelihood function unlike existing methods.

We have studied our method extensively using simulated and real datasets. We did simulations under various models assuming different distribution of reads. Our method shows promising result and outperforms other methods significantly, especially when (a) the number of alternately spliced isoforms is large, and (b) some isoforms are extremely low abundant. Our method is also robust to the probability distribution of reads, more accurate and applicable even with a mixture of paired- and single-end reads, scalable with respect to memory allocation, and computationally very fast. It shows high correlation with qRT-PCR estimates when applied to a real data set. Confidence intervals calculated using our method are narrower than Cufflinks estimates. Based on its performance on simulated and real datasets, we believe that it will be extremely useful and feasible approach in practical implementation with real data.

84 | Pharmacogenetic effects in population-based metabolic profiles

Martina Müller-Nurasyid[1,2,3], Katharina Schramm[1,2,3], Margit Heier[4], Maik Pietzner[5,6], Kathrin Budde[5,6], Jerzy Adamski[7], Christian Gieger[4,8], Karsten Suhre[9], Gabi Kastenmüller[10], Konstantin Strauch[1,11]

[1]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [2]Department of Internal Medicine I (Cardiology), Hospital of the Ludwig-Maximilians-University (LMU) Munich, Germany; [3]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany; [4]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [5]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany; [6]DZHK (German Centre for Cardiovascular Research), partner site Greifswald, Greifswald, Germany; [7]Institute of Experimental Genetics, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [8]Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [9]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Doha, Qatar; [10]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [11]Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Germany

Genome-wide association studies (GWAS) of metabolic phenotypes have revealed new insight into the genetic basis of human metabolism. Some of the identified loci

were known to be associated with toxicity or adverse reactions to medication, suggesting the exploration of a new class of gene-environment interactions. The aim of this project was to discover novel loci that show a genotype-dependent reaction on drugs.

Using 2906 samples from the population-based KORA F4 study from Southern Germany, we performed systematic GWAS allowing for effect modification through medication intake for metabolites quantified with the AbsoluteIDQ p150 assay. Therefore, we introduced an interaction term drug*SNP in our linear regression models. Metabolite levels were log2-transformed and adjusted for age, sex and batch effects for the analysis. We defined medication by active agents of barcode scanned drugs taken within the last seven days Before blood draw. We analyzed single active agents as well as combinations of two active agents with a minimum of 100 users.

We identified 43 loci that are associated with medications. Five of these loci could be replicated in an independent German study (SHIP-TREND). All these loci are associated with metabolite hexose and mainly show association with combination of Hydrochlorothiazide and Ramipril.

Our study suggests potential mechanisms leading to known adverse reactions to active agents. More work is needed to elucidate the downstream effects of the identified interactions in the associated pathways. If our results can be corroborated, it might be conceivable to provide optimized recommendations for individualized drug prescription.

85 | Whole exome sequencing in lung cancer families identifies significantly linked loci on multiple chromosomes

Anthony M. Musolf[1], Claudio W. Pikielny[2], Yafang Li[2], Richard K. Wilson[3], Christopher I. Amos[4], Ramaswamy Govindan[5], 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, United States of America; [2]Geisel School of Medicine, Dartmouth College, Lebanon, United States of America; [3]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, United States of America; [4]Baylor College of Medicine, Houston, United States of America; [5]Division of Oncology, Washington University School of Medicine, St. Louis, United States of America

Lung cancer (LC) kills more people in the United States than any other cancer: 25% of all cancer deaths nationwide. LC risk is known to be increased by a variety of environmental factors, especially smoking, however, the disease has a significant genetic component as well. We performed whole exome sequencing (WES) *on* 204 individuals from 25 extended families. These families have a strong history of LC and are highly aggregated for the phenotype. Individuals from each family were chosen for sequencing to maximize information in a genetic linkage analysis. The WES was recalled with PICARD/GATK and standard quality controls were performed, leaving approximately 500,000 SNVs and indels for analysis.

We performed two-point parametric linkage analysis on LC and the WES data including phenotype data for both sequenced and unsequenced family members (493 total) using an autosomal dominant model with disease allele frequency of 1%, 10% penetrance for carriers and a 1% phenocopy rate. We performed two-point linkage analysis on the individual markers (SNVs and indels) using an Elston-Stewart algorithm.

It is likely that LC risk is caused by a rare, highly penetrant variant. To increase power on rare variants, we created multiallelic pseudomarkers. The pseudomarkers were constructed from the haplotypes of the rare variants (MAF < 0.01) located in each gene. Two-point linkage was performed on the pseudomarkers.

Initial analyses found multiple significant gene-based linkages on multiple chromosomes, including known cancer genes OBSCN, RYR2, and NBPF. We are currently performing additional analyses and adding more sequenced families to increase power.

86 | Telomere length and vascular phenotypes in a population-based cohort of children and mid-life adults

John Nguyen[1,2], Melissa Wake[1,2]

[1]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia; [2]Murdoch Children's Research Institute, Melbourne, Victoria, Australia

Telomere length may be a potential contributor to the lifecourse development of cardiovascular disease risk. Current evidence is lacking in children and in epidemiological settings. We investigated the associations of telomere length with vascular structure and function in children aged 11–12 years and mid-life adults.

Telomere length (T/S ratio) was calculated by quantitative real-time polymerase chain reaction from blood-derived genomic DNA. Vascular structure was assessed by carotid intima-media thickness (IMT), and vascular function by carotid-femoral pulse wave velocity (PWV) and carotid elasticity. Data were available for 1206 children and 1343 adults. Adults had lower T/S ratio (0.81 vs 1.09; $P < 0.001$), higher carotid IMT (663 vs 580 μm; $P < 0.001$), faster carotid-femoral PWV (7.0 vs

4.4 m/s; $P < 0.001$), and lower carotid elasticity (2.4 vs 4.8%/10 mmHg; $P < 0.001$), compared to children. Adjusted linear regression models showed no association between T/S ratio and carotid IMT, carotid-femoral PWV or carotid elasticity in children. Similarly, there was no association between T/S ratio and carotid IMT and carotid-femoral PWV in adults, but a marginal difference with carotid elasticity (0.14%/10 mmHg; 95% CI 0.05–0.2; $P < 0.05$).

In a cohort of healthy children and mid-life adults, there is no evidence of an association between telomere length and vascular structure or function, and weak evidence of an association between longer telomere length and increased carotid elasticity in adults. In conclusion, telomere length is not associated cross-sectionally with clinically meaningful differences in cardiovascular parameters. Longitudinal studies and studies in high risk populations are warranted.

87 | Invited abstract: What African genomes tell us about the origins of breast cancer

Olufunmilayo I. Olopade[1]

[1]*Center for Clinical Cancer Genetics and Global Health, University of Chicago, Chicago, United States of America*

Analysis of cancer genomes has provided fundamental insights into the process of malignant transformation, and cancer genomes have rapidly become an integral part of the practice of clinical oncology, with implications for diagnosis, prognosis, treatment, and prevention. Inherited and sporadic cancers often share common mutational events. Pathogenic *BRCA1* and *BRCA2* mutations are the strongest predictors of breast and ovarian cancer risk and can now also be categorized as the strongest predictors of aggressive prostate cancer risk. Within two decades of identifying *BRCA1* and *BRCA2* as major breast and ovarian cancer susceptibility genes, there are already several FDA approved targeted therapies to treat BRCA associated cancers. To date, risk reducing interventions have been driven by outdated perspectives and approaches with no defined pathways for proactively assisting patients at risk of aggressive breast cancers. The situation is worsened by the heterogeneity and diversity of the social and cultural context in which individuals and families with inherited cancer gene mutations are identified. Work from our group and others have defined the genomic landscape of common cancers such as breast, colon and prostate cancers. Using high throughput whole genome strategies, including genome-wide association studies, whole exome sequencing, and whole genome sequencing, we are deeply exploring the most foundational instigators of the most aggressive forms of cancers across the African Diaspora. In addition, mutation burden appears to predict response to immunotherapy that is rapidly changing options for treating all cancers. To accelerate progress and promote health equity, we have embarked on innovative interventions that couple genomic analysis for risk prediction with innovative interventions to reduce the high mortality from aggressive young onset cancers in low resource settings in the US and Nigeria. I will present our recent findings and future directions for genetic epidemiology research in underserved and understudied African ancestry populations.

88 | Significance testing for allelic heterogeneity

Yangqing Deng[1], **Wei Pan**[1]

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

It is useful to detect allelic heterogeneity (AH), that is the presence of multiple causal SNPs in a locus, which for example may guide developing new methods for fine mapping and may determine how to interpret an appearing epistasis. In contrast to Mendelian traits, the existence and extent of AH for complex traits had been largely unknown until the recent publication of Hormozdiari et al. (2017, *AJHG 100* : 789–802.). Hormozdiari et al. (2017) proposed a Bayesian method, called causal variants identification in associated regions (CAVIAR), and uncovered widespread AH in complex traits. However, there are several limitations with CAVIAR. First, it assumes a maximum number of causal SNPs in a locus, typically up to 6, to save computing time; this assumption, as to be shown, may influence the outcome. Second, its computational time can be too demanding to be feasible since it examines all possible combinations of causal SNPs (under the assumed upper bound). Finally, it outputs a posterior probability of AH, which may be difficult to calibrate with a commonly used nominal significance level. Here we introduce an intersection-union test (IUT) based on a joint/conditional regression model with all the SNPs in a locus to infer AH. We also propose two sequential IUT-based testing procedures to estimate the number of causal SNPs. Our proposed methods are applicable to not only individual-level genotypic and phenotypic data, but also Genome-wide Association Studies (GWAS) summary statistics. We provide numerical examples

based on both simulated and real data, including a large-scale schizophrenia (SCZ) and a high-density lipoprotein (HDL) GWAS summary datasets, to demonstrate the effectiveness of the new methods. In particular, for both the SCZ and HDL data, our proposed IUT not only was faster but also detected more AH loci than CAVIAR. Our proposed methods are expected to be useful in further uncovering the extent of AH in complex traits.

89 | SNP-derived transcriptomics and hierarchical clustering to identify inversely regulated genetic expression patterns between Alzheimer's & cancer

Gita A. Pathak[1], Nicole R. Phillips[1]

[1]Department of Microbiology, Immunology, and Genetics, Graduate School of Biomedical Sciences, UNT Health Science Center, Fort Worth, United States of America

Alzheimer's and cancer are two aging-associated diseases, and several epidemiological findings have reported an inverse correlation between the two diseases. Even though aging is a risk factor for Alzheimer's and cancer, the two diseases have been known to have discrete molecular mechanisms. The goal of this study is to identify genetic variants that may be influencing functional changes on Alzheimer's disease and cancer biology. The genotype data was acquired via authorized access, and after quality control filtering, the data was analyzed for 677, 578 and 3857 individuals for Alzheimer's, breast and prostate cancer respectively. All the subjects were >50 years of age. The gene expression profiles were predicted for five brain, breast and prostate tissues from cis-SNPs within 1 kb of the gene position using genotype tissue expressions (GTEx's) reference transcriptome data set. Gene expression-phenotype association was analyzed using generalized linear model and adjusted for population stratification using first two eigenvectors as covariates. Common significant genes ($p < .05$) across the three association analyses, and with significant differences between Alzheimer's and cancer are being reported here, followed by tissue-based functional enrichment analysis using gene ontology and protein interaction network. The genotype-derived expression profiles are also analyzed using hierarchical clustering against common controls after merging all the datasets for imputation to remove batch effect for the identification of inversely associated genetic expression signatures in Alzheimer's and cancer. Understanding the influence of SNPs on disease biology of Alzheimer's and cancer can be used for development of therapeutic strategies derived from their protective pathogenicity over each other.

90 | Incorporation of heterogeneity through a mixture model to boost power of association tests

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

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

For most, if not all common human diseases and complex traits, individuals are etiologically heterogeneous. Genome-wide Association Studies (GWAS) aim to discover common genetic variants that are associated with the complex traits, typically without considering heterogeneity. Heterogeneity, as well as imprecise phenotyping, significantly reduces the power to find genetic variants associated with human diseases and complex traits. Disease subtyping can explain some of the heterogeneity through well-developed unsupervised clustering techniques such as latent class analysis; however, existing disease subtyping techniques are either based on clinical features or on genomics data, but none incorporate heterogeneity into the association framework by considering the clinical features and genomics data together. Here, we use a finite mixture model with logistic regression to incorporate heterogeneity into the association testing framework for a case-control study. In the proposed model the disease outcome is modeled as a mixture of two Bernoulli distributions. One of the component distributions refers to the subgroup of the population for which the genetic variant is not associated with the disease outcome, (where is the coefficient of the genetic variant in the logistic regression model) and another component distribution corresponds to the subgroup for which the genetic variant is associated with the disease outcome. The mixing parameter corresponds to the proportion of the population for which the genetic variant is associated with the disease outcome. The proposed association test is expected to have more power than traditional methods of association testing. A simulation study of a trait with differing levels of prevalence, SNP minor allele frequency, and odds ratio will be performed to compare power between the models with and without incorporating heterogeneity.

91 | Bioshrink: a R shiny application for Bayesian analysis of genetic association studies that incorporates biological information

Miguel Pereira[1,2], John R. Thompson[3], Peter G. Burney[1], Cosetta Minelli[1]

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

We present Bioshrink, a user-friendly R Shiny-based application that implements a Bayesian hierarchical shrinkage model which aims to improve the analysis of genetic association data by incorporating external biological information in a joint SNP analysis. Users of Bioshrink require only a basic knowledge of Bayesian statistics and shrinkage regression to understand how the analysis is performed.

External biological information provided by the user as a prior information score is included through differential shrinkage. The Bayesian model implemented assumes the SNP effects follow a normal distribution centered at zero, with SNP-specific variance controlled by a shrinkage parameter. Bioshrink uses biological information to inform the amount of shrinkage; SNPs without biological support are more strongly shrunk towards zero, thus favoring the detection of SNPs with biological support.

Bioshrink takes the following inputs: the data set with phenotype and genotype data, a biological knowledge score for each SNP, number of expected true signals, and shrinkage parameter range (lower and upper bounds). The user can perform the analysis using different shrinkage ranges and can choose the best range according to the variance ratio, a measure that we developed to identify the best shrinkage range in this scenario. The application runs the Bayesian analysis using an efficient estimation algorithm and outputs a matrix with the SNPs ordered by Bayesian p value. Results can be looked-up interactively or downloaded as a space-separated file.

Bioshrink is freely available online and can be accessed at: http://www.bioshrink.org.

92 | Population stratification in the Estonian biobank and its confounding with complex traits

Natalia Pervjakova[1,2,3], Kristi Läll[1], Merli Mändul[1], Andrew P. Morris[1,4], Reedik Mägi[1], Krista Fischer[1]

[1]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia; [2]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; [3]Genomics of Common Disease, Division of Diabetes, Endocrinology and Metabolism, Department of Medicine, Imperial College London, United Kingdom; [4]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom

Principal components (PCs) based on a multidimensional scaling (MDS) analysis of the genetic relatedness matrix have often been used for adjusting genome-wide association analyses for population stratification. Biobanks collect growing amounts of genetic information, increasing the potential impact of fine-scale stratification and cryptic relatedness.

We analysed 49,363 participants from the Estonian Genome Center University of Tartu (EGCUT) cohort genotyped using various Illumina arrays. We calculated the first 250 PCs using 61,044 independent variants. We used analysis of variance (ANOVA) for assessing how the PCs are associated with the county of birth, linear regression analysis for prevalent coronary heart disease (CAD) and type 2 diabetes (T2D), and polygenic risk scores (GRS) for the same diseases.

Most of the first 250 PCs were significantly associated with the county of birth, showing that even though adjustment for the PCs can help reduce the impact of potential confounding due to population stratification, very large numbers of them should be considered.

Three first PCs, as well as county of birth, were significantly associated with prevalent CAD status, but not with T2D. However, the GRS for T2D was significantly associated with the first 11 PCs and the GRS for CAD with first and third PC.

Our results indicate that although PCs may help to adjust for population stratification in some cases, it is not clear whether a small number of them can fully capture the confounding due to geographic variation. We suggest that studies covering a large proportion of a population should consider using a mixed modelling framework with a random effect for the genetic relatedness matrix.

93 | Phenotypic consequences of lipid trait gene dysregulation

Lauren E. Petty[1], Hung-Hsin Chen[1], Global Hispanic Lipids Consortium, Jennifer E. Below[1]

[1]*Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, United States of America*

Genome-wide association studies (GWAS) have successfully identified more than 300 loci associated with serum lipid levels, and hypothesis-driven research into pleiotropic effects of these loci has demonstrated shared genetic architecture of lipid traits with metabolic traits, including glucose metabolism, and cardiovascular dysfunction, including hypertension. There has been little work, however, searching phenome-wide. Building off of novel results from our recent meta-GWAS and S-PrediXcan study of serum lipids (HDL cholesterol, LDL cholesterol, total cholesterol, and triglycerides) in the largest assemblage of Hispanic/Latinos to date ($N \approx 26k$), we sought to identify the impact of known and novel lipids genes across the phenome. PredixVU is a resource of genetically regulated gene expression association test results using PrediXcan GTEx-derived models with extensive electronic health record phenotypes from BioVU, including >23k currently genotyped samples. We queried the set of identified genes for each lipid trait in PredixVU and compared phenotype code counts to random permutation gene set counts to discover phenotypes enriched for association with these genes. We identified disorders of the female reproductive system, kidney and excretory phenotypes, and several blood count phenotypes as additional phenotypic consequences of dysregulation of lipid-associated genes. We also see additional evidence of pleiotropic effects of these genes with hypertensive phenotypes, however, we do not observe metabolic phenotypes among the most enriched. These results suggest that hypothesis-driven approaches may be missing important sequelae of lipid trait genes; further study of our identified phenotypic effects will allow for better understanding of their shared mechanism.

94 | Multi-phenotype genome-wide association study of protein levels in individuals with pulmonary arterial hypertension

Edita Pileckyte[1], Marika Kaakinen[1], Christopher J. Rhodes[1], NIHR Bioresource for Rare Diseases, Martin Wilkins[1], Inga Prokopenko[1]

[1]*Imperial College London, London, United Kingdom*

Genome-wide association studies (GWAS) identify numerous genetic risk factors associated with complex diseases but large proportion of heritability remains unexplained. We aimed to identify genetic risk variants underlying susceptibility to pulmonary arterial hypertension (PAH), a rare but severe complex disease. We performed single-phenotype (SP) and multi-phenotype (MP) GWAS of SOMAscan platform 1,124 blood plasma protein levels in 128 European PAH patients to identify one-variant-one-phenotype and one-variant-multiple-phenotypes associations, respectively. We defined proteins related to cardiovascular, metabolic, inflammatory, pulmonary, immune processes, and stress response, based on SOMAscan classification. Five proteins ERBB1, Hemoglobin, HMG-1, Notch 1, and TNF-α – formed a cluster related to all above functions. Before analysis, each protein level was natural logarithm transformed and adjusted for age, sex, and four principal components to control for population stratification. The resulting residuals were further inverse-normal transformed to assure normality. In SP-GWAS, we discovered and replicated 27 protein quantitative loci (pQTLs), corrected for multiple testing (P value $< 4.45 \times 10^{-11}$), among which four were novel trans-pQTLs. This included an association at *ELK2AP,* located on the border of the *immunoglobulin heavy locus (IGH),* for death receptor 3 (DR3) protein, a major regulator of inflammatory response. We performed MP-GWAS using SCOPA software that implements "reverse regression" approach. The five-protein cluster analysis demonstrated a pQTL at *TPK1* gene, encoding thiamine pyrophosphokinase. Coding *TPK1* mutations lead to thiamine metabolism dysfunction syndrome that manifests as dizziness and dystonia, symptoms common to PAH patients. These large-scale proteomics GWAS in PAH patients identified PAH-specific pQTLs not reported previously in studies of healthy individuals.

95 | Epigenome-wide association study of change in body mass index from young- to middle adulthood in 626 northern Finland birth cohort 1966 participants

Harmen Draisma[1], Marika Kaakinen[1,2], Laurie Prelot[1], Mila D. Anasanti[1], Zhanna Balkhiyarova[1], Matthias Wielscher[3], Sylvain Sebert[1,4], Marjo-Riitta Jarvelin[3,4], Inga Prokopenko[1]

[1]*Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, United Kingdom;* [2]*Centre for Pharmacology and Therapeutics, Department of Medicine, Imperial College London, London, United Kingdom;* [3]*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom;* [4]*Center for Life Course Health Research, University of Oulu, Oulu, Finland*

The genomewide association studies (GWAS) have identified thousands of genetic loci for various phenotypes crosssectionally. However, GWAS haven't been successful in identifying *longitudinal* genetic effects. A number of associations of interindividual variation in body mass index (BMI) with variation in the degree of DNA methylation have been reported. However, the relationship between *longitudinal* changes in BMI and their effects on differential methylation is underexplored. We aimed to develop novel methodology for the detection of longitudinal effects on DNA methylation and tested it on the example of BMI change over time. For 626 individuals from the Northern Finland Birth Cohort 1966, we calculated the average change in BMI per year using measurements at ages 31 (T1) and 46 (T2). We used BMI change residuals corrected for sex to test for association with the degree of methylation measured in blood at T2 for 832,569 markers from Illumina (San Diego, CA, USA) MethylationEPIC BeadChip using methylSCOPA software, developed by us. We quality-controlled the methylation data, regressed out the effects of measured (potential) confounders, and normalized the methylation signal intensity data. We detected epigenome-wide significant associations ($P < 1 \times 10^{-7}$) with BMI change at cg14476101, cg11376147, and cg27243685 representing *PGHDH*, *SLC43A1*, and *ABCG1* loci, established for concurrent BMI. For the first time, we report associations between DNA methylation levels and BMI change in Europeans. We proposed a novel method to detect associations between change in phenotype over time and DNA methylation and provide the first account of its use on example of longitudinal changes in BMI.

96 | Epigenetic age acceleration is associated with target organ damage in African Americans

Jeremy Raisky[1], Wei Zhao[1], Scott M. Ratliff[1], Stephen T. Turner[2] Thomas H. Mosley[3], Jennifer A. Smith[1,4]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, United States of America; [2]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, United States of America; [3]Division of Geriatrics and Gerontology, University of Mississippi Medical Center, Jackson, United States of America; [4]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, United States of America

Hypertension leads to damage in multiple target organ systems, including the brain, kidney, heart, and peripheral arteries. Additional biomarkers for specific types of target organ damage are needed to enhance prevention and treatment efforts and better characterize disease mechanisms. In this study, we evaluated whether epigenetic age acceleration, a measure of biologic aging based on DNA methylation, is associated with six types of organ damage in a primarily hypertensive cohort. Subjects included 333 African Americans from the Genetic Epidemiology Network of Arteriopathy (GENOA) study. DNA methylation was measured from peripheral blood leukocytes collected at baseline (1996–2000) with the Illumina HumanMethylation450 BeadChip, and was used to estimate intrinsic (IEAA) and extrinsic (EEAA) epigenetic age acceleration. Measures of target organ damage were assessed in a follow-up visit (2000–2004). Linear regression was used to examine the association between age acceleration with each measure of organ damage, and to test for interaction between age acceleration and blood pressure. IEAA was significantly associated with urine albumin to creatinine ratio (UACR) and ankle-brachial index (ABI) ($P < 0.01$). A ten-year increase in IEAA was associated with an 11.1 mg/g increase in UACR and a 0.35 decrease in ABI. IEAA and systolic blood pressure (SBP) had a significant interaction on ABI ($P = 0.03$), and the effect of increasing age acceleration on ABI was more pronounced in those with higher SBP. In conclusion, measures of increased cellular aging such as IEAA may be valuable subclinical biomarkers for multiple types of target organ damage, and additional studies are needed to elucidate the functional mechanisms underlying these relationships.

97 | Invited abstract: Genetic contribution to obesity in African populations: the H3Africa AWI-gen study

Michèle Ramsay[1,2], Ananyo Choudhury[1], Scott Hazelhurst[1,3] as members of AWI-Gen and the H3Africa Consortium

[1]Sydney Brenner Institute for Molecular Bioscience, Johannesburg, South Africa; [2]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, Johannesburg, South Africa; [3]School of Electrical Engineering, University of the Witwatersrand, Johannesburg, South Africa

Across Africa, the health and epidemiological transition is marked by an increase in obesity and related cardiometabolic diseases (CMDs). To identify genetic and environmental risk factors for susceptibility to common complex traits, we developed a population cross-sectional study of over 10,500 participants from six communities in four African countries (Burkina Faso, Ghana, Kenya and South Africa). Data include demography, health history, anthropometry, behavior and blood and urine biomarkers. Participants were genotyped on the newly designed Human Heredity and Health in Africa Consortium SNP genotyping array. The genotype data was enriched with imputation up to ~20 million

SNPs using the Sanger Institute African imputation reference panel. Regional and sex-specific differences show that women are more likely to be obese (body mass index (BMI) $\geq$ 30) than men in South (42–66% vs 3–17%); East (32% vs 5%); and West Africa (1–4% vs 1–2%). The covariates associated with obesity in different regions also show clear variation, including life style, behaviour and relevant biomarkers, posing analytic challenges. Furthermore, principal component analysis revealed significant population sub-structure between and within geographic regions. Increased genetic diversity and generally lower linkage disequilibrium (LD) in African populations present an advantage for fine mapping and the identification of causal variants. While revealing universal and African-specific associations with CMD-related traits these studies could enhance our understanding of the biology of obesity and related traits, including diabetes and hypertension, among Africans.

98 | German Neonatal Network – resource for genetic analyses in very low birth weight infants

Tanja K. Rausch[1,2], Inke R. König[1], Wolfgang Göpel[2]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]Department of Pediatrics, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Founded in 2009, the German Neonatal Network examines the development of very low birth weight infants until the age of five in Germany with 63 study centers.

During their stay in hospital after birth, the assessment includes the infants' basic data, medications, surgeries, and therapies. After discharge from the hospital, parents report on their infants' development via an annual questionnaire. A follow-up team of study nurses and a medical doctor from Lübeck visits all study sites to examine the infants at the age of five years. Besides recording basic data, they conduct several tests to examine eyes, ears, lung, movement, and development. Umbilical cord tissue frozen after birth is used to genotype the DNA of the infants. Affymetrix Axiom™ Genome-Wide CEU 1 Array Plate 2.0 and Illumina Infinium® Global Screening Array-24 v1.0 were used for chip genotyping.

Our continuously growing database already contains clinical data of 17,900 very low birth weight infants, genetic data of 7,400, and five-year follow-up data of 1,850 infants. In the main data set 49.1% of 15,462 very low birth weight infants are female, the mean birth weight is 1,059.82 $\pm$ 306.76 g and the mean gestational age is 28.7 $\pm$ 2.76 weeks.

Ultimately, genetic information will be available from all infants who were examined at the age of five years. A resource is already available providing answers to a wide variety of research questions.

99 | The EXCEED study: a resource for genomics of multimorbidity, with consent to recall by phenotype

Nicola F. Reeve[1], Catherine John[1], Alexander T. Williams[1], Susan E. Wallace[1], Ron Hsu[1], Robert C. Free[2], Edward J. Hollox[3], David J. Shepherd[1], Louise V. Wain[1], Martin D. Tobin[1], on behalf of the EXCEED investigators

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre – Respiratory, Glenfield Hospital Groby Road, Leicester, United Kingdom; [3]Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

Genetic studies of susceptibility to individual complex diseases have yielded many insights, but determinants of disease progression and multimorbidity are poorly understood. Multimorbidity is common in socioeconomically deprived and older age groups, and increasing in low- and middle-income countries. The universal use of doctor-coded primary care records since the 1990 s in the UK National Health Service provides an opportunity for genetic studies of these under-studied phenotypes.

We, therefore, established the Extended Cohort for E-health, Environment, and DNA (EXCEED) study which has so far recruited 9,600 participants aged 40–69 years, primarily from primary care services in Leicestershire, UK All participants provide baseline lifestyle data, a saliva sample for DNA extraction, and consent to link to their primary healthcare record, and to recall-by-phenotype and -genotype. Primary care record linkage has been completed, with 4.2 million data points relating to historical symptom, diagnostic, prescription and laboratory codes. In combination these allow longitudinal studies of disease development and progression, and observational pharmacogenetics studies. Genome-wide genotyping has commenced with the UK Biobank array, and similar primary care data linkage is underway in UK Biobank.

In EXCEED, we found that the prevalence of multimorbidity ($\geq$ 2 chronic diseases) was 25.6%, with 8.9% having 3 or more chronic diseases. This highlights the utility of EXCEED and UK Biobank for genetic studies of multimorbidity, given the similarities in ascertainment strategy and age between EXCEED and UK Biobank.

100 | Development of reporting guidelines for pharmacogenetic studies to facilitate evidence synthesis

Marty Richardson[1], Jamie Kirkham[1], Kerry Dwan[2], Derek Sloan[3], Geraint Davies[4], Andrea Jorgensen[1]

[1]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom; [2]Cochrane Editorial Unit, London, United Kingdom; [3]School of Medicine, University of St Andrews, St Andrews, United Kingdom; [4]Department of Clinical Infection, Microbiology and Immunology, University of Liverpool, Liverpool, United Kingdom

Outcomes in pharmacogenetic studies are often explained by several genetic variants each having a small effect on outcome. Consequently, large sample sizes are typically required to detect statistically significant associations between a pharmacogenetic marker and treatment response. Meta-analysis allows aggregation of data from several studies to increase sample size, and consequently power to detect significant genetic effects. However, differences often exist between pharmacogenetic studies in terms of the genetic variants investigated, how genetic subgroups are defined, outcome definitions, and the underlying assumptions made, for example about mode of inheritance, within the analyses. Since combining studies within a meta-analysis relies on them investigating the same underlying effect, these differences can significantly reduce the number of contributing studies. This problem is compounded by poor reporting of key data in study reports. The aim of our project is to develop reporting guidelines for authors of pharmacogenetic studies To facilitate the conduct of high-quality systematic reviews and meta-analyses. To produce this set of guidelines, we established a preliminary checklist of reporting items by i) including items from existing relevant guidelines, and ii) supplementing this list with any additional items thought to be important, identified through discussion and personal experience in conducting meta-analyses of pharmacogenetic studies. We have identified additional criteria specific to pharmacogenetic studies that are not specified in existing guidelines, and that are often not adhered to in pharmacogenetic study reports. We are currently planning a Delphi survey to gain consensus opinion on reporting items for our final reporting guideline.

101 | Next-generation sequencing aligned to high-resolution Nuclear Magnetic Resonance (NMR) measurements reveal role of rare variation in circulating metabolic biomarkers

Fernando Riveros-Mckay Aguilera[1], Clare Oliver-Williams[2,4], Savita Karthikeyan[2], Klaudia Walter[1], Kousik Kundu[1,5], John Danesh[1,2,3], Eleanor Wheeler[1], Eleftheria Zeggini[1], Adam Butterworth[2,3], Inês Barroso[1,6]

[1]Wellcome Sanger Institute, Cambridge, United Kingdom; [2]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [3]The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge, United Kingdom; [4]Homerton College, Hills Road, Cambridge, United Kingdom; [5]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge United Kingdom; [6]University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom

Circulating metabolite levels can be used as biomarkers, or risk factors, for cardiovascular disease (CVD). To establish the role of rare coding variation in circulating metabolic biomarker levels, we measured 226 serum lipoproteins, lipids and other amino acids and proteins in INTERVAL participants by Nuclear Magnetic Resonance (NMR). Gene-based association analyses included whole-exome sequence (WES) data from 3,741 participants (discovery) and 3,401 whole-genome sequences (WGS) from independent participants (validation). To increase power at the validation stage, we used correlated NMR traits as covariates in the analysis. Compared to unadjusted tests, this approach increased power in known gene-trait associations (e.g. 11 additional gene-trait associations for *PCSK9*), and identified *ACSL1*, *MYCN*, *FBXO36* and *B4GALNT3* as novel gene-trait associations ($P < 2.5 \times 10^{-6}$). Gene-set analyses identified novel association of loss-of-function (LoF) variants in the regulation of the pyruvate dehydrogenase (PDH) complex pathway, with lipoproteins in three broad categories intermediate-density lipoproteins (IDL), low-density lipoproteins (LDL) and circulating cholesterol ($P_{METASKAT} < 2.41 \times 10^{-6}$). In addition, genes near established HDL-associated loci were enriched for missense and LoF variants ($P_{Bonferroni} < 0.005$) in 18 HDL related traits, suggesting this gene set is enriched for effector transcripts. Exploring phenotypic tails of 49 lipoprotein and lipid measurements, we found enrichment ($P_{permutation} < 0.00037$) of likely deleterious rare variation in lipoprotein disorder and metabolism gene sets in the lower tails of four measurements which are CVD risk factors (e.g. S-VLDL-C), demonstrating that rare "protective" variation is prevalent in the phenotypic tails of a healthy population. These findings demonstrate the value of rare variant analyses to reveal novel loci, and pathways, linked to these important metabolic biomarkers.

102 | Rare variant tests for association in affected sib pairs

Razvan G. Romanescu [1], Gesseca Gos [1], Irene L. Andrulis [1,2], Shelley B. Bull [1,3]

[1]Lunenfeld-Tanenbaum Research Institute, Toronto, Canada; [2]Department of Molecular Genetics, University of Toronto, Toronto, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Recent progress in sequencing technologies has made it possible to investigate the role of rare variants (RVs) in disease etiology. Tests based on affected sib pairs (ASPs) can be more powerful than for case-control designs, as RVs tend to be enriched in families. The key idea is that RVs will be found preferentially on haplotypes shared identical by descent (IBD) versus haplotypes not shared IBD. Motivated by a whole exome sequencing (WES) study of sisters with early onset breast cancer (proband age-at-onset <45 years), we construct two burden-type test statistics which measure the departure of the joint distribution of RV allele count for the pair and their IBD state (0–2) under ascertainment compared to expectation under simple Mendelian transmission. The sisters, ascertained from the Ontario Breast Cancer Family Registry, have been screened negative for known susceptibility mutations (including BRCA1 & 2), thereby increasing the chances of finding rare familial mutations. We evaluate our methods on simulated sister pair data under early age-at-onset ascertainment, where the genetic model reflects a scientific hypothesis of locus heterogeneity, that is that different families tend to harbor different susceptibility variants. We demonstrate good type I error properties for the tests and show that one of them (at least) is more powerful than existing methods for testing association in ASPs. We then extend our methodology to test RVs in a pathway, as opposed to a single gene or region, which is expected to improve power and biological interpretability, especially under extreme locus heterogeneity across regions.

103 | Genome-wide regional genetic association of quantitative traits adapted to Linkage Disequilibrium (LD) under genomic partitioning

Delnaz Roshandel[1], Myriam Brossard[2], Sun-Ah Kim[3], Andrew D. Paterson[1,4], Yun J. Yoo[3], Shelley B. Bull [2,4]

[1]Hospital for Sick Children Research Institute, Toronto, Canada; [2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [3]Seoul National University, Seoul, Korea; [4]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Motivated by characterizations of genomic architecture where multiple-variant analysis can uncover novel associations missed by single-variant analysis, we apply computationally-efficient regression-based quantitative trait testing for region-based genomic discovery. To specify regional units, we apply a novel haplotype block detection algorithm (BigLD, Kim 2018) to cluster correlated variants and partition the genome into a large number of non-overlapping, quasi-independent linkage disequilibrium **(**LD) block regions amenable to parallel processing.

We demonstrate performance of this approach by a proof-of-principle application in GWAS of quantitative lipid traits (HDL-, LDL-cholesterol) measured at baseline in 1,340 participants of the Diabetes Control and Complications Trial in type 1 diabetes. Genotyping was by Illumina Human Core Exome Array with imputation to 1000 Genomes. Following standard quality control procedures, analysis included a total of 6.6 M genotyped and imputed variants (MAF > 5%) on autosomes. Application of BigLD yielded 91 K blocks with two or more variants (average 70 per block) and 57.5 K singleton SNPs. Within each LD block, we applied multiple linear regression including variants, sex, age, and sex by age interaction; and compared three tests for global association. All methods, including the generalized Wald statistic, reduced *df* multiple-linear-combination statistic (MLC, Yoo 2017), and regression of variant principal components (Gauderman 2007), recapitulated regions harbouring well-established associations (*CETP, LDLR, APOE*).

Although BigLD is agnostic to gene boundaries, it captures gene regions reasonably well. It also systematically captures intergenic regions, facilitating comprehensive testing of regional association using global test statistics such as MLC that is feasible for genome wide studies of imputation-dense data.

104 | Accounting for cryptic relatedness across families between subjects with no genotype data

Mohamad Saad[1], Ehsan Ullah[1], 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, United States of America

Complex traits continue to provide challenges for identifying disease-risk genes. Whole genome sequencing has become inexpensive and fast. This provides opportunities for investigation of 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.

Association testing is often used to identify variants of interest. When related subjects are included in the sample, a kinship matrix must be used to account for relatedness between subjects. Relationships may be known from the pedigree structures or can be inferred using observed genotypes. For subjects without genotype data, however, no approach is able to infer relatedness between subjects in different pedigrees.

Pedigree-based imputation increases the sample size and thus association power, and also provides genotype probabilities in subjects without genotype data. When phenotypes for such subjects exist, including them in the analysis adds information for detecting association between genotypes and phenotype. However, knowing the relationship between such subjects and the remaining subjects, especially across pedigrees, is crucial to control the type 1 error.

We propose a solution for inferring cryptic relatedness between subjects with completely missing genotype data. We use GIGI to estimate the probabilities of missing genotypes, and incorporate them in an Expectation-Maximization approach to estimate the kinship coefficients. Through simulation, our approach succeeds in inferring many types of relationships with relatively high confidence. It yields average kinship estimates of 0.19, 0.1, and 0.05 for underlying kinships of 0.25, 0.125, and 0.0625, respectively, for subjects without genotype data.

105 | ComPaSS-GWAS reduces the type I error rate of a quantitative trait GWAS when the normality assumption of regression residuals is violated

Jeremy A. Sabourin[1], Alexander F. Wilson[1]

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

When a quantitative phenotype is not adequately transformed in a regression based genome-wide association study (GWAS), the violation of the normality assumption of the residuals may result in inflation of the type I error rate. Previous work has shown that the amount of inflation is affected by at least two factors. The first is the degree of non-normality, where more non-normality results in more inflation; the second is the minor allele frequency (MAF) of the SNP, where the lower the MAF, the higher the inflation. ComPaSS-GWAS is a new method based on Complementary Pairs Stability Selection that uses a regression based GWAS aggregated over many random sample splits for internal corroboration. This approach has been shown to ameliorate the loss of power seen in a traditional two-stage GWAS approach with a single split. In this study, the type I error rate for ComPaSS-GWAS and a traditional GWAS are investigated for different degrees of non-normality for different MAFs. Simulations were performed based on the GAW19 data (1764 unrelated individuals, 73,488 SNPs). One thousand replicates of non-genetic traits following normal and gamma distributions were simulated and used to evaluate type I error rate. For non-normally distributed null traits, the traditional GWAS had minor inflation in SNPs with MAF $< 5\%$ and substantial inflation for SNPs with MAF $< 1\%$ (consistent with previous findings) while ComPaSS-GWAS had only minor inflation for SNPs with MAF $< 1\%$. These results suggest that ComPaSS-GWAS is more robust to the violation of normality.

106 | Chromosome X association analysis of Hemoglobin A1c (HbA1c) in African Americans using TOPMed Whole Genome Sequence (WGS) data

Chloé Sarnowski[1], Aaron Leong[2], Daniel DiCorpo[1], Laura Raffield[3], Xiuqing Guo[4], Paul S. de Vries[5] for the TOPMed Diabetes Working Group

[1]Department of Biostatistics, Boston University School of Public Health, Boston, United States of America; [2]Division of General Internal Medicine, Massachusetts General Hospital, Boston, United States of America; [3]Department of Genetics, University of North Carolina, Chapel Hill, United States of America; [4]Los Angeles Biomedical Research Institute at Harbor-University of California Los Angeles Medical Center, Los Angeles, United States of America; [5]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, United States of America.

Using whole genome sequence (WGS) association analysis of HbA1c (a test used to diagnose type 2 diabetes (T2D) and estimate glycemia) in 3,224 African-Americans from the Trans-Omics for Precision Medicine (TOPMed) program, we identified a chromosome X association in the *G6PD* locus (most associated SNP, rs1050828 (p.Val98Met), minor allele frequency (MAF) = 0.12, $\beta = -0.41$, $P = 4.4 \times 10^{-183}$). We sought to identify additional distinct or sex-specific associations in this region through conditional and sex-stratified analyses.

Restricting the analyses to the ± 500 kb window flanking *G6PD*, we performed conditional analysis on rs1050828 using linear mixed-effect models adjusted for

age at HbA1c measurement, study, sex, with an empirical kinship matrix to account for relatedness. Associations with $P < 1.7 \times 10^{-5}$ (0.05/2,886 variants with a minor allele count > 20) were considered distinct from rs1050828. We then performed analyses in males ($N = 1{,}312$) separately from females ($N = 1{,}912$) and meta-analyzed the sex-stratified results. Analyses were conducted on the Analysis Commons.

In addition to rs1050828, we identified a distinct ($r^2 = 0.0006$, $D' = 1$) rare signal (rs76723693 (p.Leu353-Pro), MAF = 0.005, $\beta = -0.44$, $P = 5.4 \times 10^{-9}$) which was more frequent in females ($MAF_{\_Females} = 0.006$ vs. $MAF_{\_Males} = 0.003$) and had a larger effect in males ($\beta_{\_Males} = -0.49$, $P = 2.4 \times 10^{-5}$, $\beta_{\_Females} = -0.38$, $P = 1.2 \times 10^{-4}$). Both rs1050828 and rs76723693 are missense and putative pathogenic (ClinVar). Heterogeneity between sex-stratified results was detected ($P_{het} \leq 0.10$) for 230 variants among 322 associated with HbA1c at $P \leq 5x10^{-8}$ (rs1050828, $\beta_{\_Males} = -0.43$, $P = 2.6 \times 10^{-148}$, $\beta_{\_Females} = -0.36$, $P = 1.0 \times 10^{-69}$, $P_{het} = 2.9 \times 10^{-18}$).

Two previously reported *G6PD* coding variants (rs1050828 and rs76723693) are independently associated with lower HbA1c values and their associations differ by sex. More people than just rs1050828 carriers may be underdiagnosed for T2D.

107 | Familial recurrence risk with varying amount of family history

Daniel J. Schaid[1], Shannon K. McDonnell[1], Stephen N. Thibodeau[1]

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

The familial recurrence risk is the probability that a person will have disease, given a specified number of affected relatives. Although much work has been done on ways to correct for ascertainment when there is at least one affected family member, there is little guidance on how to correct for ascertainment when focusing on families with more than one affected. The purpose of this report is to describe how to estimate familial recurrence risk when there are at least k affected family members. For complete ascertainment, each family can be viewed as sampled from a truncated binomial distribution and a moment estimator [Rider, 1955] can be used to estimate the recurrence risk. For family history surveys obtained through a proband, we assume that the family history information provided by the proband is complete and not dependent on the number of affected relatives, beyond the proband. Then, to compute recurrence risk for family history of $k > 1$, we exclude the proband (i.e., the single-ascertainment step), and use a truncated binomial to condition on having at least (k-1) affected relatives among the remaining family members. Robust variances of the recurrence risk are derived using estimating equations.

Application of our methods to a survey of prostate cancer shows that the recurrence risk increases from 15% when at least 1 family member is affected, to an asymptote of approximately 50% when at least 4 family members are affected.

108 | Genome-wide association study in 404,165 individuals identifies 139 novel signals of association with lung function

Nick Shrine[1], Anna L. Guyatt[1], Victoria E. Jackson[1], A. Mesut Erzurumluoglu[1], on behalf of the SpiroMeta consortium and the Lung eQTL study, Andrew P. Morris[3], Ian P. Hall[4], Martin D. Tobin[1,2], Louise V. Wain[1,2]

[1]*Department of Health Sciences, University of Leicester, Leicester, United Kingdom;* [2]*National Institute for Health Research Biomedical Research Centre – Respiratory theme, Glenfield Hospital, Leicester, United Kingdom;* [3]*Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom;* [4]*Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom*

Chronic obstructive pulmonary disease (COPD) is the third leading cause of death globally. Spirometrically-determined lung function is used in the diagnosis and grading of COPD. Population-based cohort studies of lung function are a powerful alternative study design to case-control studies of COPD.

Genome-wide associations of forced expiratory volume in 1 s (FEV_1), forced vital capacity (FVC), FEV_1/FVC and peak expiratory flow (PEF) were undertaken in 321,047 European individuals from UK Biobank and individuals from the SpiroMeta consortium (FEV_1, FVC, FEV_1/FVC $n = 83{,}118$; PEF $n = 24{,}218$). To maximise statistical power, whilst maintaining appropriate significance thresholds, we used: (a) a two-stage approach where variants with $P < 5 \times 10^{-9}$ in UK Biobank and $P < 10^{-3}$ in SpiroMeta were reported and (b) a one-stage approach reporting variants with $P < 5 \times 10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta and with $P < 10^{3}$ in each cohort. Novel and previously reported signals were interrogated for expression quantitative trait loci in whole lung and blood resources.

We report 139 signals novel signals of association with FEV_1, FEV_1/FVC, FVC or PEF, bringing the number of signals of association with lung function to 279. Fine-mapping of all new and previously reported signals, together with gene and protein expression analyses with improved tissue specificity and stringency, has implicated new genes and pathways highlighting the importance of cilia development, transforming growth factor beta (TGFB) signaling, and elastic fibres in the aetiology of

airflow obstruction. Many of the genes and pathways contain drug targets including those in development for indications other than COPD.

109 | Bayesian time-to-event analysis of high blood pressure

Daniel Shriner[1], Amy R. Bentley[1], Adebowale Adeyemo[1], Charles N. Rotimi[1]

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

Systolic blood pressure increases throughout the entire lifespan, whereas diastolic blood pressure increases until the sixth decade of life and then either plateaus or decreases. Given a late age-of-onset, we investigated time-to-event models of hypertension. The primary data comprised African Americans from Washington, D.C. and replication data comprised nationally representative samples of African Americans, European Americans, and Mexican Americans. First, we performed Bayesian time-to-event analysis, using a correlated gamma prior process for the baseline hazard. Instantaneous hazards increased until 59–60 years and then decreased. The inflection point in the hazard function lagged behind the inflection point in the curve of diastolic blood pressure, but never reached zero. Including a permanent stayer fraction did not improve model fit. These results indicate that true controls do not exist. Median lifetimes were 41.5 years in African Americans, 56.5 years in European Americans, and 55.5 years in Mexican Americans. Lifetime risks at age 85 years were 91.7% in African Americans, 78.7% in European Americans and 79.8% in Mexican Americans. Second, despite the absence of true controls, logistic regression with age and age^2 yielded equally good fit compared to the time-to-event model. Third, we used forward-backward regression in a non-parametric proportional hazards framework to assess 43 potential covariates. Six covariates were selected in the final model: weight, uric acid, chloride, potassium, low-density lipoprotein, and insulin; the conventional covariates sex, race, and body mass index were not selected. In a logistic model, the variance explained by age, age^2, these six covariates, and heritability was ~73%.

110 | Harmonizing psychosocial stress and CVD-risk variables for developing robust estimates of G×E analyses

Abanish Singh[1,2,3], Michael A. Babyak[1,2], Beverly H. Brummett[1,2], William E. Kraus[3,4], Ilene C. Siegler[1,2], Elizabeth R. Hauser[3,5], Redford B. Williams[1,2]

[1]Behavioral Medicine Research Center, Duke University School of Medicine, Durham, United States of America; [2]Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, United States of America; [3]Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, United States of America; [4]Department of Medicine, Duke University School of Medicine, Durham, United States of America; [5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, United States of America

Cardiovascular disease (CVD) is a strongly heritable life-course and lifestyle disease. Although advances in CVD epidemiology in recent past have enhanced our understanding of CVD pathogenesis, thus far, CVD research has not translated into a full genetic architecture of its risk prediction. Among many challenges are interactions of genes with environment (e.g., psychosocial stress), race, and/or sex and developing robust estimates of these interactions. Improved power with larger sample size contributed by the accumulation of epidemiological data could be helpful, but integration of these datasets is difficult due the absence of standardized phenotypic and environmental measures. In This study, we present the detailed illustration of a number of decisions made in our undertaking to harmonize a dozen datasets of varying size and demography of the samples. We harmonized candidate SNPs and CVD-risk variables related to demography, adiposity, hypertension, lipodystrophy, hypertriglyceridemia, hyperglycemia, depressive symptom, and chronic psychosocial stress. Using our algorithm based on proxy indicators of stress (Singh et al., 2015), we constructed synthetic psychosocial stress in nine out of 12 studies, where a formal self-rated stress measure was not available. The mega-analytic partial correlation between stress and depression while controlling for the effect of study variable in combined data set was significant ($Rho = 0.27$, $P < 0.0001$). The uniformity in the distributions of harmonized CVD-risk variables supported the tenability of the harmonization process. Our work demonstrated that it is possible to harmonize the inconsistencies and operationalize the existing data To gather large pool of samples for developing robust estimates of G×E interactions.

111 | Longitudinal analysis of DNA methylation reveals novel smoking-related loci in African Americans

Jiaxuan Liu[1], Wei Zhao[1], Xiang Zhou[2], Jennifer A. Smith[1,3]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, United States of America; [2]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor,

United States of America; [3]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, United States of America

Previous studies have implicated DNA methylation changes as a potential mechanism for the effects of smoking on physiological function and subsequent disease risk. Given the dynamic nature of epigenome, longitudinal studies are indispensable for investigating smoking-induced methylation changes over time. Using blood samples collected approximately five years apart in 380 African Americans (mean age 60.7 years) from the Genetic Epidemiology Network of Arteriopathy (GENOA) study, we measured DNA methylation levels using Illumina HumanMethylation BeadChips. Following quality control and removal of confounding effects, we quantified the methylation level of each DNA methylation (CpG) site in terms of its relative difference with respect to mean methylation level across individuals within each phase. We then evaluated the association between baseline smoking status and rate of methylation change using generalized estimating equation models for 6958 CpG sites. Smoking status was associated with methylation change for 22 CpG sites (false discovery rate $q < 0.1$), with the majority (91%) becoming less methylated over time. Methylation change was greater in ever-smokers than never-smokers, and differences in rates of change ranged from 1.96% to 29.0% per decade faster in smokers. Although biological pathway analyses were not significant, several CpGs were within genes associated with cardiovascular disease, cancers, and aging (*ICAM4*, *GUCY1A2*, and *FOXA1*). Significant enrichment was observed for CpG islands, enhancers, and DNAse hypersensitivity sites ($P < 0.05$). In conclusion, we identified novel epigenetic signatures for cigarette smoking that may have been missed in cross-sectional analyses, providing insight into the epigenetic effect of smoking and highlighting the importance of longitudinal analysis in understanding the dynamic human epigenome.

112 | Investigation of post-colonial demographic structure within the United States and implications for association analyses

Elena Sorokin[1], Kristin A. Rand[1], Julie M. Granka[1], Jake Byrnes[1], Keith Noto[1], Eurie Hong[1], Kenneth Chahine[2], Catherine A. Ball[1]

[1]AncestryDNA, San Francisco, United States of America; [2]AncestryDNA, Lehi, United States of America

Recent work has highlighted the need to evaluate the impact of fine-scale population structure on association analyses. Despite thousands of high-powered genome-wide association studies (GWAS), only about one-fifth of all studies include individuals of non-European descent. It is understood that GWAS conducted in European populations are not always generalizable to non-European populations and that continental population structure is an important factor to consider. However, it remains unclear whether fine-scale population structure, particularly more recent within-continental structure, contributes to bias in the interpretation of GWAS results. In over one million AncestryDNA customers who have consented to research, we leverage prior work determining fine-scale population structure within the United States to explore patterns of regional allele frequency variation for known genetic susceptibility loci. To account for effects of continental population structure, we examine regional sub-population frequencies of variant subsets that have been identified in each continental population. Our results show that, even when considering very recent time scales, small frequency differences may impact GWAS conclusions and interpretations. This study highlights the fundamental need to understand the generalizability of GWAS results through the lens of both continental, and within-continental, population structure, particularly in large-scale analyses.

113 | Detecting and correcting for bias in Mendelian randomization analyses using gene-by-environment interactions

Wes Spiller[1], David Slichter[2], Jack Bowden[1], George Davey Smith[1]

[1]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [2]Economics Department, Binghamton University, Binghamton, United States of America

Mendelian randomization (MR) has developed into an established method for strengthening causal inference and estimating causal effects, largely due to the proliferation of genome-wide association studies. The utility of genetic instruments remains controversial, as pleiotropic effects can introduce bias into causal estimates. This has prompted the emergence of many sensitivity analyses, though most are limited to cases in which many instruments are available. To address this limitation, we introduce MR using Gene-by-Environment interactions (MRGxE) as a framework capable of identifying and correcting for pleiotropic bias. This builds upon previous work leveraging gene-environment interactions to detect and correct for pleiotropic bias, whilst having the novel feature of allowing the validity of individual instruments to be assessed. Under the

MRG × E approach an instrument-covariate interaction, which induces variation in the association between a genetic instrument and exposure, allows for pleiotropic effects to be detected and controlled for within a linear regression framework. Further, the approach is the similarity in interpretation to conventional summary Mendelian randomization approaches. As an illustrative example, we investigate the effect of adiposity upon systolic blood pressure using data from the UK Biobank and The Genetic Investigation of ANthropometric Traits (GIANT) consortium. We use a single instrument (a weighted allelic score), with MRGxE producing findings in agreement with MR Egger regression in a two-sample summary MR setting. Finally, we assess the performance of MRG × E with respect to identifying and correcting for horizontal pleiotropy in a simulation setting, highlighting the utility of the approach even when the MRG × E assumptions are violated.

114 | Longitudinal strategies for identifying genetic associations with epigenetic changes over time

James R. Staley[1], Josine L. Min[1], Matthew Suderman[1], Tom R. Gaunt[1], Kate Tilling[1]

[1]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Epigenetic markers, such as DNA methylation, vary over time and modelling their trajectories with genetic variation could improve our understanding of biological mechanisms. However, a systematic assessment of the genetic regulation of epigenetic changes over time (e.g. for DNA methylation at 450,000 cytosine-guanine dinucleotides (CpGs)) using linear mixed models is not currently possible. To investigate these longitudinal trajectories, we have therefore developed a two-stage strategy where the initial genome-wide association study (GWAS) scan is performed using a variation of a standard linear model and the longitudinal SNP-CpG associations with P value $< 5 \times 10^{-8}$ are then assessed using a linear mixed model. Three approaches for the first stage were tested: a linear model ignoring the within-individual correlation of the repeated methylation measures, and two approaches where for each CpG the slope for methylation change is estimated for each individual and is used as the outcome to test against SNPs in a linear model. The first of these approaches estimates the subject-specific methylation changes using linear mixed models (two-step linear mixed model) and the second fits linear models to each individual separately (slope-as-outcome). These methods were applied to repeatedly measured blood DNA methylation profiles from the Accessible Resource for Integrated Epigenomics Studies (ARIES) to identify SNP associations with epigenetic changes over time. More than 100,000 longitudinal SNP-CpG associations were identified (P value $< 1 \times 10^{-13}$) using the two-stage strategies, with the slope-as-outcome approach the most efficient in terms of identifying associations to test in the second stage. These associations will be followed-up to investigate their robustness and function.

115 | Genetic architecture of the human plasma metabolome

Isobel D. Stewart[1], Praveen Surendran[2] on behalf of the mGAP (metabolome Genetic Architecture Programme) Investigators

[1]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom; [2]Cardiovascular Epidemiology Unit (CEU), Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom

Plasma metabolites are circulating small molecules with important predictive, diagnostic and causal roles for human disease. Advances in high-throughput untargeted assessment of the plasma metabolome now enable the characterisation of its genetic architecture on an unprecedented scale.

We performed meta-analyses of genome-wide association studies (GWAS) of plasma metabolites measured using Metabolon's (Durham, NC) untargeted DiscoveryHD4™ platform in up to 14,296 participants of the EPIC- Norfolk and INTERVAL studies. Variants were genotyped using the UK Biobank Axiom™ array, imputed using UK10K/1000 G/HRC reference panels and analysed using BOLT-LMM, SNPTEST and METAL. A total of 913 metabolites were detected in both studies, including lipids, amino acids, xenobiotics, co-factors and vitamins and currently unidentified metabolites. We defined 964 independent loci associated with any metabolite at conventional genome-wide significance ($p < 5 \times 10^{-8}$) and validated genetic associations with 646 metabolites involving 330 loci ($p < 5.48 \times 10^{-11}$, 5×10^{-8} adjusted for 913 tests) in a meta-analysis that included an independent set of 5,698 EPIC-Norfolk participants. Validated signals represented 1,847 locus-metabolite associations. Loci ranged from those associated with a single metabolite ($n = 124$) to the most pleiotropic region, at 11q12.3 containing the FADS1/FADS2 cluster, associated with >100 metabolites. Conditional analyses revealed many independently associated variants, giving a total of 2,599 variant-metabolite associations, representing the most comprehensive characterisation of genetic influences on human metabolism to date. Integration of transcriptomics using MetaXcan indicated levels of 529 metabolites associated with genetically predicted

expression of 1,735 genes in at least one Genotype Tissue Expression (GTEx) tissue. Results will be accessible via an online webserver to enable use by the scientific community.

116 | 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. Current results have identified over 5000 variants that are being filtered and prioritized for further analysis and cross comparison to additional datasets, as well as employing integrative approaches. Additional 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.

117 | TRIO_RVEMVS: a fast Bayesian variable selection method for trios that identifies individual rare variants

Duo Yu[1], Abigail C. Sedory[1], Kusha Mohammadi[1], Matthew D. Koslovsky[2], Michael D. Swartz[1]

[1]Department of Biostatistics and Data Science, The University of Texas School of Public Health at Houston, Houston, United States of America; [2]Department of Statistics, Rice University, Houston, United States of America

We have developed TRIO_RVEMVS, a novel method for jointly modeling common and rare variants using trio families that identifies individual rare variants (RV) driving the association of a genetic region with a disease, in addition to assessing the burden of RVs in genetic regions. TRIO_RVEMVS combines a conditional logistic regression model for family trios with a spike and slab variable selection prior and estimates the posterior modes for the probability of inclusion using the expectation-maximization (EM) algorithm, treating variable inclusion as missing. Using Cosi2 to simulate realistic trio families, we compare our method to PEDGENE using the weighted average correct association probability (WACAP, the sum of true positive and true negative rates, normalized by the total true associated and unassociated variants). At the gene level, TRIO_RVEMVS had a WACAP of 73.32% when analyzing common and rare variants jointly, outperforming PEDGENE (WACAP = 66.17%) while tying with PEDGENE when analyzing rare variants only (TRIO_RVEMVS WACAP = 61.07% vs. PEDGENE WACAP = 61.25%). Runtime for TRIO_RVEMVS (12 min) is competitive with PEDGENE (10 min) on a 30 kb region ($n = 1500$ trios). For identifying individual RVs with TRIO_RVEMVS, the average true positive rate is 2.6% with an average false positive rate of 0.1%. Thus, TRIO_RVEMVS exhibits comparable results with one of the top methods for identifying regions with important rare variants while simultaneously identifying influential variants; providing an important next step in identifying genes associated with birth defects. We anticipate applying TRIO_RVEMVS to trio data available through the Gabriella Miller Kids First Pediatric Research Program.

118 | Extending SNP-based heritability analysis: how many variants show strong effect in a GWAS

Fumihiko Takeuchi[1], Norihiro Kato[1]

[1]Research Institute, National Center for Global Health and Medicine, Tokyo, Japan

In SNP-based heritability analyses, the allele substitution effects of causal variants are modeled to follow a zero-mean normal distribution, whose variance represents the heritability. One topic unstudied so far is the proportion of SNPs showing strong associations in a genome-wide association study (GWAS), which is large for lipids and small for body mass index (BMI) and height. In terms of probability distribution, the proportion corresponds to the tail-heaviness, which characterizes a distribution independently of the variance. Without understanding that proportion, we cannot calculate the power of a GWAS, which is the expected number of loci to be discovered.

To address those topics, we modeled the allele substitution effects of single SNPs by the *t*-distribution, which can adjust tail-heaviness by the degrees-of-freedom parameter. We extended the Popcorn program, which analyzes GWAS summary statistics and estimates heritability in a population and genetic correlation between two populations of different ancestries. Using the observed parameters, we simulated possible distribution of SNP effect-size. GWAS datasets for blood pressure, lipids, type 2 diabetes, BMI and height in East Asian or European-ancestry populations were analyzed.

The estimated heritability ranged 0.07–0.31, and the genetic correlation ranged 0.77–0.99. The degrees of freedom were ~3 for lipids, ~5 for BMI, and ~4 for other traits, indeed indicating heavy tail for lipids. Using simulation, we computed the power of GWAS under various sample sizes.

The extension of SNP-based heritability analysis by adopting *t*-distribution enabled the modeling of the spectrum of associations in a GWAS and power calculation. The source code is available from https://github.com/fumi-github/Popcorn-t

119 | African, Native American, East Asian and European genetic ancestries and fetal growth in diverse populations

Fasil Tekola-Ayele[1], Deepika Shrestha[1], Tsegaselassie Workalemahu[1], Katherine L. Grantz[1], Cuilin Zhang[1], Jagteshwar Grewal[1], Stefanie N. Hinkle[1], Jing Wu[2]

[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, United States of America; [2]Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, United States of America

Fetal growth is an important predictor of health in children and adults. Fetal size varies by ethnicity and geography, with the smallest fetal size observed among Africans and Asians. We investigated the influence of maternal genetic ancestry on fetal growth in self-identified African Americans ($n = 591$), Asians ($n = 216$), Hispanics ($n = 535$), and Whites ($n = 603$) recruited through the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Fetal Growth Studies. Genome-wide proportions of African, Native American, East Asian and European genetic ancestries in the study samples were estimated. Associations of genetic ancestry with fetal weight at first trimester (gestational Week 14), mid-gestation (Week 20), second trimester (Week 27), and third trimester (Week 40) of pregnancy, and with risks for small- and large-for-gestational age at birth were tested. Fetal weight was positively correlated with European ancestry proportion in Whites and African Americans throughout pregnancy and with Native American ancestry proportion in Hispanics at the third trimester of pregnancy ($P < 0.05$). Fetal weight was inversely correlated with African ancestry proportion in African Americans and Hispanics at mid-gestation, second and third trimesters of pregnancy ($P < 0.05$). Among African Americans, the highest quartile of European ancestry proportion was associated with 66% lower odds for small-for-gestational age compared to those in the lowest quartile (95% CI 0.15, 0.76, $P = 0.008$) whereas, the highest quartile compared to lowest of African ancestry proportion had 2.4 times higher odds for small-for-gestational age (95% CI 1.00, 5.52, $P = 0.049$). In all, the findings indicate that genetic ancestry may contribute to variations in fetal growth throughout pregnancy and in small birth size.

120 | Methods for left-censored biomarker data: a simulation study in the two-sample case

Dominik Thiele[1,2], Inke R.König[1,2]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL), Germany

Classifying patients into subgroups in precision medicine strongly relies on the availability of biomarker data like gene expression profiles. Although there is a huge amount of candidate data, finding suitable profiles still is challenging due to the lack of reproducibility and statistical power. In addition, data is frequently left-censored, and it is yet unclear how best to handle data where a non-negligible proportion has values under a given detection limit. In this study, we target this issue by investigating the analysis of left

censored data. While imputation methods are commonly used in practice, it was shown that these perform badly. Current literature suggests that the best method relies highly on the research question and setting. However, recent studies considered only specific small sample settings without comparing methods comprehensively. To fill this gap, we performed a comprehensive simulation study considering univariate distributions and varying sample sizes to systematically compare different suggested methods. The distributions and sample sizes were chosen based on observed data from the German Center for Lung Research (DZL) All Age Asthma Cohort (ALLIANCE). Our results will help to guide researchers to select the most efficient methods for a specific setting.

121 | Latent structure within the UK Biobank sample: better the devil you know

Nicholas J. Timpson[1], Neil Davies[1], Dan Lawson[1], Simon Haworth[1], for the MRC IEU DryLab group[1]

[1]Medical Research Council Integrative Epidemiology Unit, Population Health Science, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Large studies use genotype data to discover genetic contributions to complex traits and infer relationships between those traits. Recent developments in resources, applications and understanding warrant a re-exploration of latent structure in datasets. Before 2015, very large samples were only achieved by aggregation of smaller studies whose structural properties and geographical footprints were neither detectable within single studies nor coordinated across the collection of studies. Now analysis can be undertaken in very large individual collections with the capacity to capture a single geographical footprint, such as UK Biobank. Co-incident geographical variation in genotypes and health traits can bias these analyses. Using data from the UK Biobank study, we found that single genetic variants and genetic scores composed of multiple variants are associated with birth location within UK Biobank and that geographic structure in genotype data could not be accounted for using routine adjustment for study center and principal components. Major health outcomes appear geographically structured and that coincident structure can yield biased associations. The ability of very large studies to detect effects indistinguishable from artefactual biases or ancestral differences demands reworked approaches to exploit, or at least account for, structure.

122 | A copula-based approach for modeling cancer risks in hereditary breast cancer syndrome families

Fodé Tounkara[1], M'Hammed Lajmi Lakhal Chaieb[2], Hae Jung[3], Yun-Hee Choi[3], Laurent Briollais[1,4]

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada; [2]Laval University, Quebec, Canada; [3]Western University, London, Canada; [4]Division of Biostatistics, Dalla School of Public Health, University of Toronto, Toronto, Canada

The effect of breast cancer (BRCA) gene mutations on breast cancer (BC) risk among Hereditary Breast Cancer Syndrome (HBOC) families has been evaluated by known BC risk models such as the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA). We propose to extend these models to incorporate the effects of mammography screening and prophylactic surgery as time-varying covariates while accounting for family selection bias and residual family correlations. We introduce a Cox and Oaks (CO) time-varying covariates (TVC) Cox model to handle the screening effect and prophylactic oophorectomy on BC risk. A Copula model is specified to account for familial residual dependence and the bias induced by the sampling ascertainment is corrected via a prospective likelihood approach. The CO time-varying covariate model is compared to a Permanent Exposure (PE) model using the AIC criterion. We illustrate the interest of our method through an analysis of BRCA mutation carrier families recruited through the Breast Cancer Family Registries (BCFRs). Numerical results show that prophylactic oophorectomy significantly decreases the hazard of BC among mutation carriers while having three mammographic screening increases the change to detect breast among carriers. Based on AIC score, the CO TVC model is preferable for the time effect of screening where this effect decreases over time, while the PE model fit the effect of prophylactic oophorectomy over time better. Finally, our simulation studies demonstrated the good empirical properties of our proposed method.

124 | GIGI2: a fast approach for parallel genotype imputation in large pedigrees

Ehsan Ullah[1], Khalid Kunji[1], Ellen M. Wijsman[2], Mohamad Saad[1]

[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, United States of America

Imputation of untyped SNPs has become important in Genome-wide Association Studies (GWAS). Imputation of rare variants, which are enriched in pedigrees, is more suitable in family-based designs. The costs of performing relatively large family-based GWAS can be significantly reduced by fully sequencing only a fraction of the pedigree and performing imputation on the remaining subjects. The program GIGI can efficiently perform imputation in large pedigrees but can be time consuming. Here, we implement GIGI's imputation approach in a new program, GIGI2, which performs imputation with computational time reduced by at least 25x on one thread and 120x on eight threads using multi-threaded imputation. The memory usage of GIGI2 is reduced by at least 30x. This reduction is achieved by implementing better memory layout and a better algorithm for solving the Identity by Descent graphs. We also make GIGI2 available as a webserver based on the same framework as the Michigan Imputation Server. GIGI2 is freely available online at https://cse-git.qcri.org/eullah/GIGI2 and the webserver is at https://imputation.qcri.org

125 | Investigation of genome-wide gene-by-sex interactions on time-to-asthma onset

Raphaël Veil[1], Raquel Granell[2], Catherine Laprise[3], Alan James[4], Maxim B. Freidin[5], Elza K. Khusnutdinova[6], Erika von Mutius[7], Nicole M. Probst-Hensch[8], Bénédicte Leynaert[9], Emmanuelle Bouzigon[1]

[1]Inserm, UMR-946, Genetic Variation and Human Diseases Unit; Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France; [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom; [3]Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Canada; [4]Busselton Population Medical Research Institute, Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands, Western Australia; School of Population Health, University of Western Australia, Crawley, Western Australia; [5]Research Institute of Medical Genetics, Tomsk NRMC, Tomsk, Russia; [6]Institute of Biochemistry and Genetics - Subdivision of the Ufa Federal Research Center of the Russian Academy of Science; Bashkir State University, Department of Genetics and Fundamental Medicine, Ufa, Russia; [7]Dr von Hauner Children's Hospital, Ludwig Maximilian University; Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research, Munich, Germany; [8]Swiss Tropical and Public Health Institute, Basel, Switzerland; University of Basel, Basel, Switzerland; [9]Inserm, Unit 700, Team of Epidemiology; Université Paris-Diderot Paris 7, Paris, France

Asthma is a complex disease with sex-specific differences in prevalence, clinical and biological features. Asthma is more prevalent in males during childhood, while it becomes more frequent in females in adolescence and adulthood. The mechanisms behind these sex-specific differences are not well understood and may involve hormonal changes together with differential genetic predisposition. To identify genetic variants interacting with sex that influence time-to-asthma onset (TAO), we conducted a large-scale meta-analysis of nine gene-environment-wide interaction studies (GEWIS) of TAO (totaling 7,104 men and 6,970 females of European ancestry) by using survival techniques applied to pediatric and adult asthmatic and non-asthmatic subjects. We detected three independent loci showing SNP×Sex interaction at the 10^{-5} level. The most significant association with TAO was female-specific in an intergenic region at 5q32 ($P_{female} = 9.1 \times 10^{-8}$ *vs.* $P_{male} = 0.56$ for rs6872558). The other two associations were male-specific: within *SORCS2* intron 2 at 4q16 ($P_{male} = 1.3 \times 10^{-7}$ *vs.* $P_{female} = 0.15$ for rs10005462) and within *DGKB* intron 1 at 7p21 ($P_{male} = 3.9 \times 10^{-7}$ *vs.* $P_{female} = 0.23$ for rs2189717). None of these loci had been previously associated with asthma phenotypes. Functional annotations indicated co-localization of these genetic variants with epigenetic marks and DNA regulatory elements in fibroblasts, lung or blood. By testing gene-by-sex interactions, we identified novel loci influencing asthma risk in a sex-specific manner. Candidate genes in these loci are involved in inflammatory process and immune cell regulation. Further replication of these findings are ongoing.

126 | Invited abstract: Leveraging big GWAS data to address question about selection, pleiotropy, assortative mating and epidemiology

Peter M. Visscher[1], Naomi R. Wray[1], Jian Yang[1] and the Program in Complex Trait Genomics

[1]Institute for Molecular Biology, University of Queensland, St Lucia, Australia

The availability of genome-wide genetic information and deep phenotyping on hundreds of thousands of human samples is driving research across multiple areas, including disease genetics, epidemiology, neuroscience and the social sciences. We will give examples of how genome-wide association study data on more than a million subjects can lead to insights into the genetics and biology of complex traits; how imprints of assortative mating and natural selection can be detected in the human genome; and, in general, how old questions in genetics and (genetic) epidemiology can be addressed with new (big) data.

127 | Histone H3 levels and modifications in association with gestational particulate matter exposure: the ENVIRONAGE cohort study

Karen Vrijens[1], Ann-Julie Trippas[1], Bram Janssen[1], Wouter Lefebvre[2], Charlotte Vanpoucke[3] Michelle Plusquin[1], Tim S. Nawrot[1,4]

[1]Center for Environmental Sciences, Hasselt University, Diepenbeek, Belgium; [2]Flemish Institute for Technological Research (VITO), Mol, Belgium; [3]Belgian Interregional Environment Agency (IRCELINE), Brussels, Belgium; [4]Department of Public Health, Environment & Health Unit, Leuven University (KU Leuven), Leuven, Belgium

Particulate air pollution is an important environmental health issue with adverse health effects, starting as early as in fetal life (during pregnancy). Epigenetic modifications have been suggested to mediate those effects and may increase disease predisposition in later life. To this extent, histone H3 modifications can influence gene expression by altering the chromatin state. Here, for the first time, the potential prenatal effects of exposure to particulate matter with a diameter less than 2.5 μm ($PM_{2.5}$) exposure on global histone H3 levels and H3K4 and H3K36 tri-methylation in cord blood are explored.

In 630 mother-newborn pairs from the ongoing birth cohort ENVIRONAGE, the levels of global histone H3, tri-methylated H3K4, and H3K36 protein were measured in cord blood by means of ELISA. Linear regression models were used to associate the relative H3K4me3, H3K36me3 and total histone H3 levels in cord blood with different $PM_{2.5}$ exposure windows during pregnancy. H3K4me3 and H3K36me3 levels were normalized against total histone H3 protein.

An inverse association was observed between H3K36me3 levels in cord blood and gestational $PM_{2.5}$ exposure, exposure during the last trimester of pregnancy was most significantly associated with H3K36me3 levels. Similar observations were shown with exposure to black carbon and NO_2 during pregnancy.

Our results suggest histone H3 modifications might play a role in the response to $PM_{2.5}$ exposure during pregnancy.

128 | Family-specific genetic associations with metabolic syndrome in linkage regions

Jia Y. Wan[1], Emileigh L. Willems[2], Trina Norden-Krichmar[1], Stephanie A. Santorico[2,3,4], Karen L. Edwards[1]

[1]Department of Epidemiology, School of Medicine, University of California, Irvine, United States of America; [2]Department of Mathematical and Statistical Sciences, University of Colorado, Denver, United States of America; [3]Human Medical Genetics and Genomics Program, University of Colorado, Denver, United States of America; [4]Department of Biostatistics & Informatics, University of Colorado, Denver United States of America

Specific quantitative traits that characterize the metabolic syndrome (MetS) include body weight, waist circumference, systolic and diastolic blood pressure, triglycerides (TG), high-density lipoproteins (HDL), fasting glucose, and fasting insulin. Using the GENetics of NonInsulin-dependent Diabetes mellitus (GENNID) Study as a resource of multiplex families and genetic data, we previously identified four candidate linkage regions containing putative quantitative trait nucleotides (QTNs) that influence MetS quantitative traits. Using the NimbleGen SeqCap EZ Target Enrichment protocol, these regions were then resequenced in a subset of six European American families (78 subjects) that showed high evidence of linkage. To find associated QTNs in regions of linkage, MERLIN software was used to perform family-based association testing while assuming linkage in a two-point analysis. For each candidate region, selected families were analyzed together as well as individually. The simpleM method was used to correct for testing multiple variants. Family-specific variants and variants shared across linked families were found to be associated with particular MetS quantitative traits. A number of these variants were located in genes with plausible biological functions for MetS traits. With recently developed technology and accessibility of DNA testing, personalized medicine can benefit from family-specific analyses, which allow an individual to identify possible at-risk variants and genes within his or her own family. However, family-specific results should be verified with other established population association results and genetic functional annotation databases.

129 | Likelihood-ratio based approach to select X-chromosome inactivation model

Jian Wang[1], Rajesh Talluri[3], Sanjay Shete[1,2]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, United States of America; [2]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, United States of America; [3]Department of Data Science, The University of Mississippi Medical Center, Jackson, United States of America

Analyzing X-chromosomal genetic variants is challenging because of the complexity of the X-chromosome inactivation (XCI) process for female X-chromosome loci. To address such complexity, we previously developed a unified statistical test to assess the association between X-chromosomal SNPs and complex diseases of interest, accounting for different biological possibilities of XCI:

random, skewed and escaping XCI. In the original study, we focused on the SNP–disease association test but did not provide knowledge regarding the underlying XCI models. Such knowledge is useful for understanding the contribution of SNPs to the disease, as well as the inheritance patterns of diseases. To identify the XCI model given a SNP, one can use the highest likelihood ratio (LLR; max-LLR approach). However, that approach does not formally compare the LLRs corresponding to different XCI processes to assess whether the models are distinguishable. Therefore, we proposed a LLR comparison procedure (comp-LLR approach), inspired by the Cox test, to formally compare the LLRs of different XCI models to select the most likely XCI model(s) that describe the underlying XCI process. We conducted simulation studies to investigate the performance of the max-LLR and comp-LLR approaches and applied both approaches to a head and neck cancer genetic study to investigate the underlying XCI processes for the X-chromosomal genetic variants.

130 | Approaches for curating phenotypes for pharmacogenomic genome-wide association studies of smoking cessation drug in the United Kingdom biobank

Qingning Wang[1], Chiara Batini[1], Catherine John[1], Robert C. Free[2], David Shepherd[1], Ron Hsu[1], Louise V. Wain[1], Nicola Reeve[1], Martin D. Tobin[1]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]NIHR Leicester Biomedical Research Centre - Respiratory, Glenfield Hospital Groby Road, Leicester, United Kingdom

The genetic architecture underlying treatment outcomes of varenicline, a smoking cessation aid, is not well understood in pharmacogenomic genome-wide association studies (GWAS), and clinical trials are underpowered. With a plan to conduct a GWAS for long term effectiveness of varenicline in real world data, we propose definitions of exposure and response in pharmacogenetic studies of varenicline response utilising electronic health care records.

Given the imminent availability of linked primary care records in UK Biobank, we utilised 255 varenicline recipients' data from the Extended Cohort for Ehealth, Environment, and DNA study (EXCEED). With these data, we curated definitions of exposure to varenicline treatment by constraining the sequence of prescription codes and the time gap between them. Subsequently, we investigated quit rates at different time periods after varenicline prescription to determine a definition of long term treatment outcome.

We propose to define the varenicline treatment exposure by at least 2 varenicline prescription codes with time gaps ≤ 30 days. Of all exposure definitions, this definition resembles a real life scenario and retains 221 (86.6%) individuals. Because 93.6% exposed individuals have smoking status recorded up to 1 year and most relapse occurs within a year, we propose to define the phenotype as successful cessation at 1 year follow up after the 1st treatment exposure of the proposed definition.

Our curated definitions of the phenotype has informed power calculations for our planned GWAS in UK Biobank and can inspire definitions of utility to pharmacogenomic studies of other smoking cessation aids in real world data.

131 | Efficient gene-environment interaction tests for large-scale sequencing studies

Xinyu Wang[1], Han Chen[2,3]

[1]Department of Biostatistics and Data Science, the University of Texas Health Science Center at Houston, Houston, 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, 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, United States of America

Complex human diseases and related quantitative traits are the interplay of many risk factors, including genetic and environmental components. Gene-Environment interaction (GEI) tests have been applied to identify genetic variations that modify the effect of environmental exposures. For rare genetic variants from whole genome sequencing studies, GEI tests for variant sets have been developed, such as those in the rareGE package in the generalized linear mixed model (GLMM) framework. However, these tests become impractical in large-scale sequencing studies due to the intensive computational efforts required, as fitting separate GLMMs for each variant set across the whole genome is a nontrivial task. Here we propose a computationally efficient method, Mixed-model Association tests for GEne-Environment interactions (MAGEE), for detecting the interactions between rare genetic variants and an environmental risk factor on binary and quantitative phenotypes in large-scale sequencing studies with complex related samples. We first fit a global null GLMM accounting for various levels of relatedness, but without any genetic main effects, which only needs to be fitted once in a whole genome GEI study. Statistical tests for variant sets are then performed as a variance component test by accounting for the genetic main effects using matrix projections to reduce the computational complexity from cubic to quadratic with the sample size. Simulation results show that the new method yields similar p values

rareGE, while being two orders of magnitude times faster in a sample of 10,000 individuals. We also illustrate our method using a real data example from a whole genome sequencing study.

132 | Impute multiple phenotypes using ridge regression approach

Xuexia Wang[1], Qiuying Sha[2], Shuanglin Zhang[2]

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

Recently, joint analysis of correlated traits is more and more popular in genome-wide association studies. Not only can it increase power on detecting disease associated genetic variants, but also address the issue of pleiotropy which may shed light on understanding how biochemical pathways relate to complex diseases. However, the power for joint analysis of multiple traits depends on the number of individuals whose phenotypes are collected. Some phenotypes are hard to collect due to high cost and loss of follow-up. When phenotypes are difficult to collect, the sample size might be insufficient to achieve the desired statistical power, which may lead to the failure of a study. In this paper, we propose to impute missing phenotype values using a ridge regression approach. Extensive simulation studies show that our method outperforms existing state-of-the-art methods in imputation accuracy and can boost positive findings in genetic association study. In addition, we apply our proposed method to the whole-genome chronic obstructive pulmonary disease (COPD) Gene Study. Compared with the existing methods, the proposed method also identified some novel COPD associated genes, which can improve our understanding of the etiology of COPD.

133 | Hierarchical regularized regression for incorporating external information in high-dimensional prediction models

Garrett M. Weaver[1], Juan Pablo Lewinger[1]

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, United States of America

Growing repositories of data that describe the structure and function of the genome may contain annotations that are relevant to the effects of genomic features on clinical outcomes. We propose a novel extension of regularized regression that enables the inclusion of such annotations and has the potential to improve the prediction of health-related outcomes in high-dimensional regression models that utilize genomic data. A sparsity-inducing penalty on the external information allows our model to identify relevant annotations for the prediction task at hand. Through simulation, we show that when the external data is informative, our model has improved predictive ability compared to standard approaches that do not include the external information. We also show that the additional penalty on the external data ensures there is little to no reduction in prediction performance when the external data is non-informative. Our model is applied with Gene Ontology annotations and other external information to identify gene expression signatures that predict clinical recurrence in prostate cancer patients and survival in breast cancer patients. Our method is available as an R package (https://github.com/USCbiostats/hierr) that utilizes coordinate descent to efficiently fit our model with the ability to apply some of the most commonly used penalties to both genomic features and external annotations.

134 | Imputed gene associations identify replicable *trans*-acting and target gene pairs enriched in transcription factor pathways

Heather E. Wheeler [1,2,3], Sally Ploch [1], Alvaro N. Barbeira [4], Hae Kyung Im [4]

[1]Department of Biology, Loyola University Chicago, Chicago, United States of America; [2]Department of Computer Science, Loyola University Chicago, Chicago, United States of America; [3]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood, United States of America; [4]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, United States of America

Genetic variation often affects complex traits through regulation of gene expression. Much progress has been made in uncovering *cis*-acting expression quantitative trait loci (cis-eQTLs), but *trans*-acting eQTLs have been more difficult to identify and replicate. Here, we apply PrediXcan to map *trans*-acting genes, rather than SNPs. We compared results between the discovery Framingham Heart Study (FHS, $n = 4838$) and the replication Depression Genes and Networks (DGN, $n = 922$) whole blood cohorts. We used *cis*-acting models built in the Genotype Tissue Expression (GTEx) Project whole blood cohort to predict gene expression in FHS and DGN. Examining the correlations between predicted and observed gene expression of gene pairs on different chromosomes, 124 pairs were significantly correlated in FHS, with an expected true positive rate (π_1) of 0.70 in DGN. To determine if combining predicted expression across tissue models could improve our ability to detect *trans*-acting/target gene pairs, we used the multivariable regression

approach called MulTiXcan, which accounts for correlation among predicted expression levels across 44 GTEx tissues. When we applied MulTiXcan to the FHS data, 3657 *trans*-acting/target gene pairs were discovered and 693 replicated in DGN, a dramatic increase compared to the single tissue model, indicating shared *cis*-eQTLs allow us to leverage multiple tissues to detect more *trans*-acting effects. Pathway analysis of replicated MulTiXcan gene pairs revealed the *trans*-acting genes are enriched in transcription and nucleic acid binding pathways and target genes are enriched in transcription factor binding sites and highly conserved motifs, indicating that our method identifies genes of expected function.

135 | Correcting for confounding from batch effects and genotype imputation with whole genome sequence data: application to the ADSP family sample

Ellen M. Wijsman[1,2], Tyler R. Day[1]

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, United States of America; [2]Department of Biostatistics, University of Washington, Seattle, United States of America

Disease gene identification often employs whole genome sequence (WGS) data. Evaluation of WGS variants may be carried out with association methods used in genome-wide association studies (GWAS), including adjustment for related subjects, if necessary.

Costs of WGS and time required to generate data introduce practical complications. High cost induces incentives to limit the number of subjects with newly-generated WGS. Limiting the sample size may include oversampling cases while using previously sequenced subjects from other studies as controls, or sequencing only some subjects with genotype imputation into the rest. The time needed to generate data also leads to temporal or spatial batch-effects. We show in analysis of the Alzheimer's Disease Sequencing Project (ADSP) family sample that these issues lead to confounding that affects results.

We carried out association testing with WGS data in 110 ADSP discovery families. Initial WGS on 578 subjects were generated for an 8:1 case-control ratio. A second round of WGS emphasized controls, but still yielded an overall skewed case-control ratio. Only *p* values obtained from the initial WGS conformed to the expected distribution. The joint WGS from both rounds of sequencing lead to a genomic inflation factor, λ, of 25% in the $-\log_{10}$p, correctable with a batch covariate. Genotype imputation lead to severe ($\lambda > 200\%$) inflation factors, which was negatively impacted by batch effects. Inflation could be controlled using data driven individual measures of missing data rates and imputation deviance, and family-based WGS fractions from each batch. These results show that care is needed to avoid spurious conclusions from WGS data.

136 | Trans-ethnic meta-analysis of metabolic syndrome in a multi-ethnic study

Emileigh L. Willems[1], Jia Y. Wan[2], Trina M. Norden-Krichmar[2], Karen L. Edwards[2], Stephanie A. Santorico[1,3,4]

[1]Department of Mathematical and Statistical Sciences, University of Colorado, Denver, United States of America; [2]Department of Epidemiology, University of California, Irvine, United States of America; [3]Human Medical Genetics and Genomics Program, University of Colorado, Denver, United States of America; [4]Department of Biostatistics & Informatics, University of Colorado, Denver, United States of America

Metabolic Syndrome (MetS) is defined as a clustering of metabolic risk factors that includes: abdominal obesity, high triglyceride (TG) levels, high fasting glucose levels, low levels of high-density lipoproteins (HDL) cholesterol, and high blood pressure. Samples from the GENetics of NonInsulin-dependent Diabetes mellitus (GENNID) Study serve as a characterization of Type 2 Diabetes multiplex families across four diverse ethnic groups:

African American (#families = 73, #individuals = 288),
European-American (#families = 79, #individuals = 519),
Japanese-American (#families = 17, #individuals = 132),
Mexican-American (#families = 113, #individuals = 610).

Our study focuses on the quantitative traits that characterize MetS to better understand the individual genetic factors which when clustered can lead to a diagnosis of MetS.

The goal of our study is to identify trans-ethnic associations between common variants in the genome and eight quantitative MetS traits (body weight, waist circumference, systolic and diastolic blood pressure, triglycerides (TG), high-density lipoproteins (HDL), fasting glucose, and fasting insulin) in the GENNID Study using meta-analysis methods. Genome-wide association studies (GWAS) results from four ethnic groups were combined using four meta-analysis methods (TransMeta [Biometrics 72:945-954, 2016] and MR-MEGA [HMG 26:3639–3650, 2017] as well as fixed and random effects models). Preliminary trans-ethnic meta-analysis results from these methods support the presence of heterogeneity between the association results from ethnic groups. We compared the four methods when applied to the diverse GENNID Study, along with each method's ability to accommodate

heterogeneity between ethnic groups. Additionally, heterogeneity between ethnic groups' association results will be explored, and comparisons will be made between the trans-ethnic meta-analysis results and previous linkage analysis results from the GENNID Study. Understanding the heterogeneity present between GENNID groups will allow better insight into the genomic structure of diverse ethnic groups; we hope that this understanding will help to address the current disparity of care among diverse groups in the field of personalized medicine.

137 | Developing approaches to detecting UK primary care-treated respiratory infections for use in genetic studies

Alexander T. Williams[1], Nicola F. Reeve[1], Catherine John[1], Robert C. Free[3], David J. Shepherd[1], Louise V. Wain[1,3], David Michalovich[2], Martin D. Tobin[1,3]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Refractory Respiratory Inflammation Discovery Performance Unit, GlaxoSmithKline, Stevenage, United Kingdom; [3]NIHR Leicester Biomedical Research Centre – Respiratory, Glenfield Hospital, Leicester, United Kingdom

A typical genome-wide association study consists of gathering individuals according to the presence of some phenotype of interest. This can be an expensive and laborious approach even for some common diseases. Routinely collected medical records potentially allow studies to be conducted in larger numbers of individuals, with detailed, longitudinal medical history on all participants enabling complex phenotypes to be defined and studied.

In the UK, 98% of the population is registered with a general practitioner (GP). Electronic recording of diagnoses and prescriptions was introduced in the 1990 s, meaning large-scale resources, such as UK Biobank, will eventually hold ~20 years of historical GP data on all participants. These data allow for difficult-to-study phenotypes to become more easily accessible to researchers. However, electronic healthcare records are stored using a complex clinical coding system, requiring careful code selection to properly define phenotype.

Genetic determinants of respiratory infections are understudied and remain poorly understood. Here, we describe approaches to detecting individuals with recurrent respiratory infections using linked primary care data from the Extended Cohort for E-health, Environment, and DNA (EXCEED) study. We compare the impact of different coding strategies to detect episodes of respiratory infection, including the impact of inclusion of different classes of clinical codes, including prescribing data, referrals and laboratory tests. The research will be extended to include recently generated genome-wide association data in EXCEED and linked primary care data in UK Biobank, to address research questions not currently attainable in standard observational study designs.

138 | Multiple-kernel learning for genomic data mining and prediction

Christopher M. Wilson[1], Brooke, L. Fridley[1], Xuefeng Wang[1]

[1]Department of Bioinformatics and Biostatistics, the Moffitt Cancer Center, Tampa, United States of America

There is a growing demand in cancer research for statistical methods and machine learning techniques that can fuse information from a variety of data sources, especially multiple types of Omics data. The ability to utilize information across a variety of Omics platforms allows for personalized treatment regimens and better prediction of outcomes. Support vector machines (SVM) can classify binary outcomes using linear or non-linear functions. Kernel selection is crucial to the success of SVM algorithms. On the other hand, multiple kernel learning (MKL) algorithms can use linear combinations of many candidate kernels. The use of multiple kernels can eliminate kernel selection and tuning of hyperparameters since a wide range of candidate kernels can be applied and the most suitable combination will be selected. Kernels can be used to represent a single, group, or all of the features in a data set, which may lead to more biologically relevant predictors. For instance, a kernel can be used to group genes or metabolites into pathways. We apply SVM and several different MKL methods to both simulated and cancer data, to provide suggestions for which types of kernels are appropriate for different types of Omics data. We demonstrate that MKL can provide a more accurate classification rule than SVM. Additionally, we provide results from Dual Augmented Lagrangian MKL algorithm, which is computationally efficient, has good performance for thousands of candidate kernels, and can be extended to continuous and survival outcomes.

139 | A Genome-wide association study of emotion recognition and theory of mind

Marc R. Woodbury-Smith[1,2], Peter Szatmari[2,3], Stephen W. Scherer[2,4], Andrew D. Paterson[2,5]

[1]Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, United Kingdom; [2]The Centre for Applied Genomics, Hospital for Sick Children, Toronto, Canada; [3]Division of Child and Youth Mental Health, University of Toronto, Toronto, Canada; [4]McLaughlin Centre for Molecular Medicine, University of Toronto, Toronto, Canada. [5]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Our capacity to negotiate the complex social world is mediated by numerous brain mechanisms working together. At a neuropsychological level, there is evidence for genetic contributions for several of these mechanisms. The Avon Longitudinal Study of Parents and Children (ALSPAC) has collected data on two such neuropsychological domains, emotion recognition and theory of mind (ToM), in more than 5,000 typically developing children (M:F ~ 1:1), who have also been genotyped with imputation to the 1000 genomes project. We have undertaken a genome-wide association study (GWAS) of both of these traits. Previously, genome-wide significant signals for ToM have been published using 23andMe data (N ~80,000).

Using non-imputed SNPs, genome-wide estimation of heritability was ~5–7% for both traits, with very little shared heritability. Three loci with suggestive evidence of association for emotion recognition were identified these did not overlap the published 23andMe results for ToM. GWAS of ToM in ALSPAC identified five signals at the suggestive level of significance; these also did not overlap with the published ToM results, or with the ALSPAC emotion recognition signals. Our results suggest unique rather than shared genetic mechanisms for these two traits, and a number of loci that may harbour genes underlying social cognition have been identified. Our study, therefore, has important implications for the understanding of the genetic underpinning of everyday social interaction.

140 | Maternal and fetal genetic interactions, imprinting, and risk of placental abruption

Tsegaselassie Workalemahu[1,2], Daniel A. Enquobahrie[1,3], Bizu Gelaye[4], Mahlet G. Tadesse[5], Sixto E. Sanchez[6,7], Fasil Tekola-Ayele[2], Anjum Hajat[1], Timothy A. Thornton[8], Cande V. Ananth[9,10], Michelle A. Williams[4]

[1]Department of Epidemiology, School of Public Health, University of Washington, Seattle, United States of America; [2]Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, United States of America; [3]Center for Perinatal Studies, Swedish Medical Center, Seattle, United States of America; [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, United States of America; [5]Department of Mathematics and Statistics, Georgetown University, Washington, D.C., United States of America; [6]Facultad de Medicina Humana, Universidad San Martín de Porres, Lima, Peru; [7]Asociación Civil PROESA, Lima, Peru; [8]Department of Biostatistics, University of Washington, Seattle, United States of America; [9]Department of Obstetrics and Gynecology, Roy and Diana Vagelos College of Physicians and Surgeons, Columbia University, New York City, United States of America; [10]Department of Epidemiology, Joseph L. Mailman School of Public Health, Columbia University, New York City, United States of America

Maternal genetic variations in mitochondrial biogenesis (MB) and oxidative phosphorylation (OP), have been associated with placental abruption (PA). However, investigations of maternal and fetal genetic interactions (MFGI) and parent of origin (imprinting) effects in PA are limited. We investigated MFGI in MB/OP, and imprinting effects in relation to risk of PA. Among Peruvian mother-infant pairs, single nucleotide polymorphisms (SNPs) were selected to characterize variations in MB/OP and imprinted genes. For each MB/OP SNP, four multinomial models corresponding to fetal allele effect, maternal allele effect, maternal and fetal allele additive effect, and maternal fetal allele interaction effect were fit. The Bayesian information criterion was used for model selection. Imprinting effect was tested using a likelihood ratio test. Bonferroni corrections were used to determine statistical significance. Models with MFGI effects provided improved fit than models with only maternal and fetal genotype main effects for rs12530904 (log-likelihood ratio = 18.2; *P* value = 1.2e-04) in *CAMK2B,* and rs73136795 (log-likelihood ratio = 21.7; *P* value= 1.9e-04) in *PPARG*, both MB genes. We identified 311 SNPs in 35 maternally-imprinted genes (including *KCNQ1*, *NPM*, and, *ATP10A*) associated with abruption. Top hits included rs8036892 (*P* value = 2.3e-15) in *ATP10A*, rs80203467 (*P* value = 6.7e-15) and rs12589854 (*P* value = 1.4e-14) in *MEG8*, and rs138281088 in *SLC22A2* (*P* value = 1.7e-13). We identified PA related maternal and fetal MB gene interactions and imprinting effects that highlight the role of the fetus on PA risk. Findings can inform mechanistic investigations to understand the pathogenesis of PA.

141 | Integration of mQTL data and enhancer-promoter interactions with GWAS summary results identifies novel genes

Chong Wu[1], Wei Pan[1]

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

Although genome-wide association studies (GWAS) have identified thousands of variants associated with complex diseases, most identified variants are located outside gene coding regions and a biological interpretation of their function is largely unknown. By noting that enhancer-promoter interactions play important roles in regulating gene expression, a new gene-based method called E+G (Wu and Pan, Genetics, early online, 2018) has been proposed to both boost statistical power and offer biological insights by integrating enhancer-promoter

interactions with GWAS association results. On the other hand, methylation quantitative trait loci (mQTLs) and expression quantitative trait loci (eQTLs) provide orthogonal ways of functionally annotating SNPs for complex traits and common diseases (Gamazon et al., Mol Psych 18, 340–346, 2013). Although some new gene-based methods, such as transcriptome-wide association study (TWAS) (Gamazon et al., Nat Genet 47, 1091–1098, 2015; Gusev et al., Nat Genet 48, 245–252, 2016), have been proposed to integrate eQTL data with GWAS results, the topic of integrating mQTL data with GWAS summary results remains largely untouched. Here, we propose integrating enhancer-promoter interactions and mQTL data with GWAS summary results to enhancer mechanistic interpretability and boost statistical power. Through an application to two large-scale schizophrenia (SCZ) GWAS summary datasets, we demonstrate that the proposed method could identify some novel SCZ-associated genes (containing no significant SNPs nearby). For example, for the larger SCZ data set with 36,989 cases and 113,075 controls, our proposed method identified 12 novel genes, all missed by TWAS, E + G, and the standard gene-based method. We conclude that our proposed method is potentially useful, complementary to TWAS, E+G, and other existing standard methods.

142 | Causal inference with GWAS-based Mendelian randomization for CAD and T2DM

Hongyan Xu[1]

[1]*Augusta University, Augusta, United States of America*

Mendelian randomization (MR) is an effective approach for causal analysis of risk factors for complex diseases. It is critical to use appropriate genetic variants as instrumental variables for valid and efficient Mendelian randomization analysis using data from genome-wide association studies (GWAS). We developed an approach to systematically select single nucleotide polymorphisms (SNPs) to investigate the causality of blood lipids on coronary heart disease (CAD) and type 2 diabetes mellitus (T2DM). The SNP selection considered statistical significance, sample size and physical distance for each SNP. Using the 338 SNPs selected for multivariable MR analysis of CAD, we found that low density lipoprotein cholesterol and triglycerides increased risk for CAD ($\beta = 0.424$ and 0.200; p = 6.23E-22 and 1.95E-07, respectively), high density lipoprotein cholesterol (HDL-c) had protective effect against CAD ($\beta = -0.286$; p = 2.21E-13). With 309 SNPs for similar analysis of T2DM, HDL-c was found to decrease risk for diabetes ($\beta = -0.1677$, p = 5.06E-07). MR-Egger analysis gave consistent results. T2DM was significantly associated with increased risk for CAD ($\beta = 0.11499$, p = 2.93E-10) with MR analysis. Our results provided evidence of casual effect of blood lipids on CAD and T2DM and T2D on CAD.

143 | Role of mismatch repair genes in colorectal cancer: a study of North Indian population

Alka Yadav[1], **Mayank Jain**[1], **Ashok Kumar**[1], **Rajan Saxena**[1], **Niraj Kumari**[2], **Narendra Krishnani**[2]

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

Colorectal cancer (CRC) involves malignant neoplasm of the colon and rectum. It is the second and third most commonly diagnosed cancer among women and men worldwide respectively and the fourth leading cause of cancer-related deaths. It is estimated that worldwide, there are approximately 1,400,000 new cases and about 700,000 deaths per year. In this study we looked for mismatch repair (MMR) Protein expression (MLH1, MSH2, PMS2 and MSH6) by immunohistochemistry in CRC patients and its clinicopathological correlation. A prospective study was conducted on histologically proven CRC patients in a tertiary care hospital.

Fifty-two patients (38 males and 14 females) underwent resection for CRC, with the median age of 52 years (16–81 years). Eighteen patients were younger than 50 years of the age and three patients had associated history of malignancy in the family. Twenty-nine (56%) patients had right colon cancer, nine patients (17%) had left colon cancers and 14 patients (27%) had rectal cancer. Histology revealed well differentiated tumours in 16 patients, moderately differentiated tumours in 10 patients and poorly differentiated tumours in 26 patients. MMR protein loss was seen in 15 (29%) patients. Seven (46%) of these patients were <50 years of age and a combined loss of MSH2 and MSH6 was found in 6 patients. Twelve (80%) patients with MMR protein loss had tumours located proximal to the splenic flexure as compared to 3 (20%) patients who had tumours located in the distal to the splenic flexure. There was no difference in MMR protein loss based on patient's age, gender, degree of differentiation, stage and tumour histological characteristics.

This study revealed that there was less than 30% MMR protein loss in CRC patients in north Indian population and that the loss was most commonly seen in right sided colon cancer than left sided colon cancer.

144 | Performance of polygenic risk scores in correlated quantitative target traits

Summaira Yasmeen[1], Heike Bickeböller[1]

[1]*Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany*

Polygenic risk scores (PRSs), based on consortium summary statistics for a discovery trait, have been commonly used to estimate genetic risks of correlated target traits in much smaller samples assuming a substantial overlap in genetic aetiology.

To investigate variance explained (R^2) by PRSs in target traits, we conducted a simulation study with 34K and thereof 100 replicates of 200 individuals for discovery and target sample. Based on the European Hapmap reference population of Utah Residents with Northern and Western European Ancestry (CEPH-CEU), we simulated with Hapgen2.0 50 K SNPs keeping the Linkage Disequilibrium pattern. Discovery trait and its effect sizes for 20 causal SNPs within an additive genetic model were normally distributed; effect sizes were based on 80% SNP heritability. We simulated three target traits T1, T2, and T3. T1 is based on the same generation model as the discovery trait; T2 and T3 have 80% and 60% correlation with T1.

The PRS was constructed using *P value* thresholds ranging from 0.0 to 0.05 in 0.01 increments. The optimal threshold distribution across replicates was right skewed U-shaped for T1, more symmetric U-shaped for T2 and left skewed for T3. R^2 values are normally distributed across all thresholds. For the smallest threshold 0.01 the mean R^2 (SD) was 0.32 (0.08), 0.21 (0.06) and 0.12 (0.04) for T1, T2 and T3, respectively. With decreasing correlation, R^2 decreases and the peak of the optimal threshold distribution gradually shifts to the left. Thus the best possible R^2 is reached with a smaller total SNP number SNPs included.

145 | The role of human mitochondrial DNA variants in common complex phenotypes

Ekaterina Yonova-Doing[1], Claudia Calabrese[2], 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 produce about 90% of the cellular energy needed to survive via the oxidative phosphorylation cycle (OXPHOS). Perturbations in OXPHOS can lead to mitochondrial dysfunction, generation of reactive oxygen species, apoptosis, premature aging and various age-related diseases. There are over 100 proteins involved in OXPHOS, 13 of which are encoded by the maternally-inherited mitochondrial DNA (mtDNA), while the rest are encoded by the nuclear genome. The shared genetic control permits a quicker and more efficient OXPHOS regulation in individual mitochondria. Despite the pivotal role of mitochondria in human health and disease, the role of mtDNA sequence and copy number variants is largely unexplored at scale in well-powered studies.

We have addressed this using the UK Biobank ($N = 500{,}000$) and INTERVAL ($N = 50{,}000$) studies, both of which were genotyped on the UK Biobank array. We performed quality control of the mitochondrial variants including genotype re-calling and then imputed mtDNA variants to a reference panel to maximise the number of variants. We have uncovered mtDNA-related genetic structure, after accounting for nuclear DNA ancestry. We performed single-variant analysis, aggregate tests and undertook TreeWAS analyses to understand how mtDNA sequence variants affect >4,000 traits including cardio-metabolic conditions and endophenotypes (e.g. metabolites and nuclear encoded plasma proteins). We have identified 100 s of variant-trait associations ranging from longevity to cellular phenotypes for example, white blood cell count.

The results of this study will improve understanding of disease pathophysiology and may help prioritise novel drug targets for a verity of complex diseases.

146 | Genetic association of arterial stiffness index with incident coronary artery disease and congestive heart failure

Seyedeh Maryam Zekavat[1,2,3,4], Mary Haas[1], Krishna Aragam[1,4,7], Connor Emdin[1,5], Amit V. Khera[1,4,6,7], Derek Klarin[1,4], Hongyu Zhao[3,8], Pradeep Natarajan[1,4,6,7]

[1]*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, United States of America;* [2]*Yale School of Medicine, New Haven, United States of America;* [3]*Computational Biology & Bioinformatics Program, Yale University, New Haven, United States of America;* [4]*Center for Genomic Medicine, Massachusetts General Hospital, Boston, United States of America;* [5]*Harvard Medical School, Boston, United States of America;* [6]*Department of Medicine, Harvard Medical School, Boston, United States of America;* [7]*Cardiovascular Research Center, Massachusetts General Hospital, Boston, United States of America;* [8]*Department of Biostatistics, Yale School of Public Health, New Haven, United States of America*

Arterial stiffness index (ASI) is a noninvasive, rapid measurement extracted from finger infrared analysis. ASI has been independently associated with cardiovascular disease risk in multiple epidemiological studies; however,

it is unknown whether these associations represent causal relationships.

Here, we used Mendelian randomization to determine whether a genetic predisposition to increased arterial stiffness is associated with increased risk for incident congestive heart failure (CHF) and coronary artery disease (CAD). We first performed a genome-wide association analysis of arterial stiffness index in 131,686 participants from the UK Biobank. Using these results, we developed a 744-variant polygenic risk score that robustly associated with ASI ($P < 1 \times 10^{-300}$, F-statistic = 10,020), and tested its association with incident CHF and CAD through Mendelian randomization. Across nearly 400 K participants of the UK Biobank and 200 K participants of the CARDIOGRAMplusC4D consortium, we do not find evidence supporting a causal association with incident CAD ($P > 0.05$). However, we do observe evidence supporting a causal association of ASI with incident CHF (HR = 1.23 per SD increase in genetically-mediated ASI, $P = 5.7 \times 10^{-3}$), independent of other cardiometabolic risk factors. Furthermore, we find that the association with incident CHF is present only among individuals without hypertension (HR = 1.52 per SD increase in genetically-mediated ASI, $P = 1.3 \times 10^{-4}$) and not among those with hypertension ($P > 0.05$) ($P_{\text{interaction}} = 9.6 \times 10^{-3}$).

These results are consistent with a causal association between ASI and CHF but not between ASI and CAD, and support ASI as an orthogonal clinical tool particularly for individuals without hypertension. Furthermore, these data suggest that reducing ASI, particularly among those without hypertension, may reduce risk for CHF.

147 | LASSO with custom penalization based on external information

Chubing Zeng[1], Duncan C. Thomas[1], Juan Pablo Lewinger[1]

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

The Least Absolute Shrinkage and Selection Operator (LASSO) is a popular technique for fitting regression models to high-dimensional genomic data. However, the single penalty parameter in the LASSO applies equally to all regression coefficients, potentially over-shrinking important coefficients and under-shrinking unimportant ones. With the goal of improving prediction, we extend the LASSO to allow the regression coefficients to be differentially penalized based on external information such as Gene Ontology annotations or previous studies. In our model, each coefficient has its own "custom" individual penalty that is in turn modeled as a log-linear function of the external information. The individual penalties can be interpreted as variance terms of the double exponential prior in a Bayesian formulation of the LASSO and can be estimated using Empirical Bayes (EB) based on an approximation of the marginal likelihood. Through simulations, we show that the LASSO with custom penalties can greatly outperform the standard LASSO in terms of prediction when the external data is informative for the regression effect sizes while having negligible loss of performance when the external data is not informative. We demonstrate our approach with applications to gene expression studies of bone density and breast cancer.

148 | A new statistical method to detect novel disease associated genes using publicly available GWAS summary data

Jianjun Zhang[1], Zihan Zhao[2], Xuan Guo[3], Bin Guo[4], Baolin Wu[4]

[1]Department of Mathematics, University of North Texas, Denton, United States of America; [2]Texas Academy of Mathematics & Science, University of North Texas, Denton, United States of America; [3]Department of Computer Science and Engineering, University of North Texas, Denton, United States of America; [4]School of Public Health, University of Minnesota, Minneapolis, United States of America

To date, a large number of common variants underlying complex diseases have been identified by genome-wide association study (GWAS). More and more summary data has been posted for public access. However, most of the common variants are identified by the single-marker tests, which tests one single nucleotide polymorphisms (SNP) at a time. A gene, rather than a SNP is the basic functional unit of inheritance. Results obtained at the gene level can be more readily extended to and integrated with downstream functional analysis. We propose a gene-based P value adaptive combination approach that can incorporate association evidence from summary statistics, either P values or other statistical values from studies with continuous or binary traits. Our simulations show that the proposed method controls the type I errors well. Application to the GWAS meta-analysis results of fasting glucose from the MAGIC consortium suggests that the proposed method has improved statistical power over single-marker analysis and other existing methods in most of the cases. The proposed method identified some novel glycemic associated genes, which can improve our understanding of the mechanisms involved in β-cell function and glucose homeostasis.

149 | Incorporating prior information into set-based analyses using higher criticism statistics with an application to amyotrophic lateral sclerosis

Mengqi Zhang[1,2,3], Sahar Gelfman[4], David B. Goldstein[4], Andrew S. Allen[1,2,3]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, United States of America; [2]Center for Genomic and Computational Biology, Duke University, Durham, United States of America; [3]Center for Statistical Genetics and Genomics, Duke University, Durham, United States of America; [4]Institute of Genomic Medicine, Columbia University, New York City, United States of America

Gene-set-based analyses are often used to identify pathways where genetic variation may lead to disease. These analyses combine signal across the genes in the set and attempts to identify any evidence of non-null effects within the set. Currently, these analyses typically treat genes within the sets similarly, even though there is substantial information concerning the likely importance of each gene within the sets. For example, variants within hub genes (measured by network centrality) are more likely to substantially perturb a genetic pathway. Here we improve gene-set analyses by incorporating such prior information into a higher criticism-based gene-set analysis. Our simulation results show that when the prior information is correlated with disease related genes, our approach leads to a significant increase in power. We illustrate our approach with an analysis of amyotrophic lateral sclerosis (ALS) and show that our approach leads to the prioritization of ALS related pathways.

150 | A fast family-based quantitative trait association test for nuclear and extended pedigrees for the analysis of whole genome sequence data

Zhihui Zhang[1], Suzanne M. Leal[1]

[1]Molecular and Human Genetics, Baylor College of Medicine, Houston, United States of America

Whole genome sequence data is being used to elucidate the genetic etiology of complex traits. One strategy is to perform association study using families because pathogenic variants within families are likely to have larger effect sizes than those for sporadic cases. Although there are several family-based association methods for binary traits, there are only a couple for analyzing quantitative traits, for example, famSKAT and FFBSKAT. We recently extended the Rare variant (RV)-generalized disequilibrium test (GDT) to analyze quantitative traits RV-QGDT. This method analyzes rare variants in aggregate and can be used for any pedigree type including with missing data. We evaluated speed, type I error and power for nuclear and extended pedigrees (10 members). We found that RV-QDGT is five times faster than famSKAT and comparable to FFBSKAT. FFBSKAT and famSKAT give almost identical results because they implement the same algorithm. We evaluated power for $\alpha = 2.5\ 10^{-6}$ for a variety of pedigree types. For 1000 sib-pairs power was 52.5% (FFBSKAT) and 69.2% (RV-QGDT) for no missing data and when 50% of the founders are missing sequence data power is >15% higher for RV-QGDT than FFBSKAT. For 1000 multiplex families, the power gain is 6% for RV-QGDT compared to FFBSKAT when none of the family members are missing data and increases to 12.5% when 50% of the parents or grand-parents are missing their sequence data. We will further demonstrate the power and utility of RV-QGDT method using simulated and late-onset Alzheimer's disease families, to elucidate genetic factors underlying complex disease etiology.

151 | Partial nearest-neighbor prediction correlation test: an effective method for finding correlations between two continuous variables

Shaodong Liu[1], Yujie Zhou[1], Dayuan Zheng[2], Jianzhong Li[1,3], Zhaogong Zhang[1]

[1]School of Computer Science and Technology, Heilongjiang University, Harbin, China; [2]School of Data Science and Technology, Heilongjiang University, Harbin, China; [3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Determining whether two continuous variables are relevant, either linear correlated or non-linear correlated is a basic statistical problem. Two continuous variables (X, Y), often are assumed linearly correlated because it is simple and has a relative complete solution. To judge whether two continuous variables are in nonlinear correlation is far more difficult. We proposed novel Partial Nearest-Neighbor Prediction correlation Test (PNNPT) to determine whether two continuous variables are nonlinearly correlated. PNNPT can effectively solve the nonlinear problem which existing methods fail - such as "class ring" development track. In the PNNPT framework, we used the value of variable X to construct a partial neighborhood structure. We then used the corresponding y for the two neighbors closest to x based on the partial neighborhood structure. The sum of square errors was calculated to measure how well Y was predicted by X. Finally, we used permutation tests based on the sum of square errors to determine whether two continuous variables are relevant. To evaluate the strength of PNNPT compared to eight existing methods, we performed extensive simulations to explore the relationship between various methods and compared the statistical

power. Simulation results demonstrate that PNNPT is an efficient method to test non-linear correlations in real-world applications. It can apply to detect for gene-gene effect.

152 | A rare variant non-parametric linkage method for nuclear and extended pedigrees with application to exome and whole genome sequence data

Linhai Zhao[1], Di Zhang[1], Carl A. Broadbent[1], Gao T. Wang[2], Badri N. Vardarajan[4], Alison M. Goate[3], Richard Mayeux[4], Suzanne M. Leal[1]

[1]Center for Statistical Genetics, Baylor College of Medicine, Houston, United States of America; [2]Department of Human Genetics, The University of Chicago, Chicago, United States of America; [3]Department of Neuroscience and Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, 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, United States of America

Nonparametric-linkage (NPL) methods were applied to common variants to analyze complex familial traits. Although linkage could be detected, gene identification was usually not successful due to large mapped regions, for example, >50 Mb. Motivated by rare-variant (RV) aggregate methods used to analyze complex trait sequence data, we developed the RV-NPL for analysis of pedigrees. The RV-NPL collapses variants within a region for example, gene and tests for sharing of rare-haplotypes. The RV-NPL has some definite advantages compared to population and pedigree RV association methods: 1.) increased power to detect causal-variants with familial aggregation due to larger effect sizes; 2.) only necessary to analyze affected individuals which increases power, because unaffected individuals can be susceptibility variant carriers; 3.) can be applied to noncoding regions; and 4.) robust to population substructure, locus and allelic heterogeneity, and inclusion of nonpathogenic variants. To evaluate the RV-NPL method, we simulated exome data using non-Finnish European minor allele frequencies (MAFs) from the Exome Aggregation Consortium (ExAC). Power was estimated by the proportion of tested genes that were significant ($\alpha = 2.5 \times 10^{-6}$). For all tested scenarios, the RV-NPL was more powerful than the NPL, for example, for 100 extended-pedigrees power is 86% (RV-NPL) and 74% (NPL) and for 2,000 affected-sib-pairs power is 87% (RV-NPL) and 65% (NPL) when variants with MAF ≤ 0.01 with odds ratio = 5 were analyzed. Results from analyzing Alzheimer's disease pedigrees and extensive simulation studies demonstrate the power of RV-NPL and its ability to detect linkage to individual genes, making it an ideal method to elucidate the genetic etiology of complex familial diseases.

153 | Multivariate genome-wide association study for volumes of structural MRI regions of interest measures via a genetic correlation network modular analysis

Xiaofeng Zhu[1], Jingjing Liang[1], Alzheimer's Disease Neuroimaging Initiative

[1]Department of Population Quantitative Health Sciences, Case Western Reserve University, Cleveland, United States of America

Challenges in imaging genetic studies are the high dimensionality, multi-modality, and high noise data with relative small sample size in conjunction with increased evidence of polygene and pleiotropy. Multivariate and network approaches, which can accommodate correlated traits, are powerful to identify potentially weak complex effects buried in high dimensional datasets and extract independent components or module in both imaging and genetic data analysis.

We calculated the pairwise genetic correlation among 145 baseline structural magnetic resonance imaging (MRI) regions of interest (ROIs) from genome-wide association studies (GWAS) summary statistics of volumes of ROIs in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants. The genetic correlation was used to define a brain imaging genetic correlation network. In this network, the nodes are 145 baseline structural MRI ROIs spanning the entire brain, while edges are genetic correlation estimated by linkage disequilibrium score regression (LDSC) method reflecting the correlation between genetic influences on ROIs traits. We adapted a weighted gene co-expression network analysis (WGCNA) framework and used the method of topological overlap matrix elements in hierarchical clustering to identify module structures. We next performed multivariate GWAS using cross-phenotype association analysis (CPASSOC) program to combine GWAS summary statistics of all ROIs as well as within each of identified modules. We observed two variants in genes *SGCZ* and *TOMM40* (region harboring APOE gene) significantly associated with ROIs of volumetric measures of occipital pole and right/left hippocampus ($P < 5 \times 10^{-8}$). In the within-module CPASSOC analysis, we observed 10 independent variants in 7 loci (*TOMM40/APOE, APCS, APOC1, ADGRL3, CLSTN2, PSD3, SSBP2*) with $P < 5 \times 10^{-8}$, in which three of them (*TOMM40/APOE, APCS, SSBP2*) reached experimental wide significance ($P < 2.27 \times 10^{-9}$). Our results suggest that incorporating a brain genetic correlation network and multivariate analysis of GWAS summary statistics of brain structural imaging ROIs improves power to detect genes in imaging GWAS.