DOI: 10.1002/gepi.22062

ABSTRACTS

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

The 2017 Annual Meeting of the International Genetic Epidemiology Society

INVITED SPEAKERS

1 | Multi-Omics Approaches in Genetic Epidemiology Studies

Katerina Kechris[1]

[1] *University of Colorado, Denver Colorado, United States of America*

Biomedical investigators are frequently using multiple high-throughput technologies to study their biological and disease system of interest. These technologies are used to generate various types of -omics profiles (e.g., transcriptomic, proteomic, metabolomic) for biomarker discovery and pathway identification. Using a multi-omics approach that integrates these profiles allows investigators to uncover potential interactions between different molecules and functional mechanisms. Furthermore, the identification of genetic variants that are associated with molecular features is important for guiding precision medicine by improving our understanding of the relationship between genetic variation and disease phenotypes. In the context of a genetic epidemiology study on Chronic Obstructive Pulmonary Disease (COPD), I will illustrate the statistical and computational challenges associated with a multi-omics approach, and highlight insights gained into the underlying genetic and molecular factors contributing to the phenotypic diversity and progression of COPD.

2 | Whole Genome Sequence-Based Association Studies for Complex Traits in Isolated Populations

Eleftheria Zeggini[1]

[1] *Wellcome Trust Sanger Institute, Hinxton, United Kingdom*

Rare variation has a key role in the genetic aetiology of complex traits. Genetically isolated populations have been established as a powerful resource for novel locus discovery and they combine advantageous characteristics that can be leveraged to expedite discovery. Genome-wide genotyping approaches coupled with whole genome sequencing efforts have transformed the landscape of disease genomics and highlight the potentially significant contribution of studies in founder populations.

3 | The 100,000 Genomes Project Transforming Healthcare

Mark Caulfield[1]

[1] *William Harvey Research Institute, Queen Mary University of London, London, United Kingdom*

The 100,000 Genomes project is using whole genome sequencing to bring diagnoses to patients with rare inherited disorders, identify drivers to cancer and response to therapy and drivers to antimicrobial resistance in pathogens. This is transforming the capability and capacity of the National Health Service (NHS) to apply genomic medicine for patient benefit who are now preparing to commission this in routine healthcare from March 2018. To do this we have created 13 NHS Genomic Medicine Centres across England to enable generation of clinical data and sample flows from NHS patients with broad consent for whole genome sequencing into our Genomics England Biorepository at the National Institute for Health Research Biosample centre. With our partner, Illumina we have one of the largest X Ten Next Generation Sequencing Centres in the World at Hinxton where we have sequenced 22,237 whole genomes. The value of this programme will be the alignment of the highest fidelity and most comprehensive whole genome DNA sequence produced from patients to date with high fidelity clinical data stored in pseudonymised format within a multi-petabyte data infrastructure. This will allow ongoing refreshment from primary, secondary and tertiary NHS care to offer a picture of life-course health and disease progression for participants. To drive up diagnoses for patients we have created the Genomics England Clinical Interpretation Partnership where 2600 clinicians and scientists will work on these data to enhance value for patients. Alongside this significant advance in NHS capability to utilise next generation sequencing for clinical care we have established over 700 person years of training to ensure we create the next generation of clinicians and scientists. This programme will ensure that our NHS has the capacity and capability to usher in a new era of Genomic Health and that the UK is amongst the most advanced in the world.

5 | Shrinkage Methods for Calculating Polygenic Risk Scores

Pak Chung Sham[1]

[1] Centre for Genomic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR

Polygenic risk scores (PRS) calculated from the results of genome-wide association studies (GWAS) are becoming increasingly popular for characterizing the genetic components of complex disorders. A simple method for calculating PRS weights the contributions of single nucleotide polymorphisms (SNPs) by their regression coefficients estimated from GWAS, for SNPs selected in having (1) in low levels of linkage disequilibrium with each other and (2) *P* values for association with the disease of below an arbitrary cutoff. The selection of SNPs with *P* values below a cutoff is a form of shrinkage – with the weight of a SNP being shrunk by a factor of 1 when its *P* value is below cutoff, and 0 when its *P* value is above cutoff. A variant of this method calculates PRS's under a range of *P* value cutoffs between 0 and 1 and chooses the cutoff value that results in the best discrimination between cases and controls. We have explored other shrinkage methods, by local false discovery rate, empirical Bayes, and LASSO-based on summary statistics. We show that these methods produce PRS's that perform at least as well as existing methods, without having to optimize over a range of *P* value cutoffs.

5 | Integrating Different Shapes and Sources of Genomic Data to Prioritise Disease-Candidate Genes

Chris Wallace[1]

[1] University of Cambridge, Cambridge, United Kingdom

Genome-Wide Association Studies (GWAS) have identified thousands of genetic variants associated with common complex diseases. However, most of these variants do not alter protein sequence, but instead are assumed to regulate nearby gene(s), perhaps in specific subsets of cells and cell-states. Different sources of genomics data have been integrated with GWAS results to attempt to link disease associated variants with the genes they regulate.

Two different statistical methods that have been used for integration of GWAS and eQTL (expression Quantitative Loci) studies will be compared: testing for colocalisation of disease and eQTL signals, and imputation of gene expression into GWAS datasets to perform association testing on imputed expression, so-called Transcriptome-Wide Association Analysis (TWAS). A newer approach will also be presented: integrating GWAS results with maps of interactions between gene promoters and regulatory elements defined using capture Hi-C. These three approaches, applied to the same autoimmune disease GWAS datasets, produce overlapping but different sets of candidate genes, and the relationship of these differences to the differences underlying the different techniques will be explored.

ORAL PRESENTATIONS

6 | FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation

Daniel Backenroth[1], Krzysztof Kiryluk[1], Valentina Boeva[2], Zihuai He[1], Lynn Pethukova[1], Ekta Khurana[3], Angela Christiano[1], Joseph Buxbaum[4], Iuliana Ionita-Laza[1]

[1] Columbia University, New York, New York, United States of America; [2] Institut Curie, Mines ParisTech, PSL Research University, F-75005, Paris, France; [3] Weill Medical College, Cornell University, New York, New York, United States of America; [4] Mount Sinai School of Medicine, New York, New York, United States of America

We propose a latent Dirichlet allocation model for predicting functional effects of noncoding genetic variants (FUN-LDA) by integrating diverse epigenetic annotations for specific tissues and cell types from large scale genomics projects such as ENCODE and Roadmap Epigenomics. Our approach allows joint modelling of data from multiple tissues and is easily extensible to data from additional tissues, not used to train the model. Using this unsupervised approach we predict tissue-specific functional effects for every position in the human genome. We demonstrate the usefulness of our predictions using several validation experiments. In particular, we provide a global view of the sharing of predicted functional variants across large number of tissues and cell types and demonstrate that functional variants in promoters are more likely to be shared across many tissues compared with variants that fall in enhancers. Using expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) project we show that eQTLs in specific GTEx tissues tend to be most enriched among the predicted functional variants in relevant tissues in Roadmap. Furthermore, we show how these integrated functional scores can be used to derive the most likely causal tissue-/cell-type for a complex trait using summary statistics from genome-wide association studies. Finally, using experimentally validated functional variants from the literature, we show that our proposed method has better accuracy and precision in predicting functional variants compared to state-of-the-art methods such as ChromHMM and GenoSkyline.

7 | Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations, and Applications to Meta-Analysis of Noncoding Variation in Metabochip Data

Zihuai He[1], Bin Xu[2], Seunggeun Lee[3], Iuliana Ionita-Laza[1]

[1]Department of Biostatistics, Columbia University, New York, New York, United States of America; [2]Department of Psychiatry, Columbia University, New York, New York, United States of America; [3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America

Substantial progress has been made in the functional annotation of genetic variation in human genome. Integrative analysis that incorporates such functional annotations into sequencing studies may aid the discovery of disease associated genetic variants, especially for those with unknown function and located outside protein coding regions. Direct incorporation of one functional annotation as weights in existing dispersion/burden tests can suffer substantial loss of power when the functional annotation is not predictive of the risk status of a variant. Here, we develop unified tests that can utilize multiple functional annotations at once for integrative association analysis with efficient computational techniques. Through extensive simulations, we show that the proposed tests significantly improve power when variant risk status can be predicted by functional annotations. Importantly, when functional annotations are not predictive of risk status, the proposed tests also incur improved power relative to existing dispersion/burden tests by learning a weight that approximates the underlying disease model in a data-adaptive manner. The tests can be constructed using summary statistics of existing dispersion/burden tests for sequencing data, therefore allowing meta-analysis of multiple studies without sharing individual level data, and leading to substantial computational improvements. We apply the proposed tests to a meta-analysis of noncoding rare variants in Metabochip data on 12,281 individuals from eight studies for lipid traits. By incorporating the Eigen functional score, we detect significant associations between noncoding rare variants in SLC22A3 gene and LDL (*low-density lipoprotein*) cholesterol, and total cholesterol, associations that are missed by standard dispersion and burden tests.

8 | SEQSpark: A Complete Analysis Tool for Large-Scale Rare Variant Association Studies using Whole Genome and Exome Sequence Data

Di Zhang[1], Linhai Zhao[1], Biao Li[1], Zongxiao He[1], Gao T. Wang[2], Dajiang J. Liu[3], Suzanne M. Leal[1]

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America; [2]Department of Human Genetics and Statistics, University of Chicago, Chicago, Illinois, United States of America; [3]Department of Public Health Sciences, College of Medicine, Pennsylvania State University, Hershey, Pennsylvania, United States of America

Massively parallel sequencing technologies provide great opportunities for discovering rare susceptibility variants involved in complex disease etiology via large-scale imputation, exome, and whole genome sequence-based association studies. Due to modest effect sizes, large sample sizes of tens to hundreds of thousands of individuals are required for adequately powered studies. Current analytical tools are obsolete when it comes to handling these large datasets. To facilitate the analysis of large-scale sequence-based studies, we developed SEQSpark which implements parallel processing based on Spark to increase the speed and efficiency of performing data quality control, annotation and association analysis. To demonstrate the versatility and speed of SEQSpark, we analyzed whole genome sequence data from the UK10K, testing for associations with waist-to-hip ratios. The analysis which was completed in 1.5 hours, included loading data, annotation, principal component analysis, single variant and rare variant aggregate association analysis of >9 million variants. For rare variant aggregate analysis, an exome-wide significant association ($P<2.5 \times 10^{-6}$) was observed with *CCDC62* [SKAT-O ($P = 6.89 \times 10^{-7}$), Combined Multivariate Collapsing ($P = 1.48 \times 10^{-6}$) and Burden of Rare Variants ($P = 1.48 \times 10^{-6}$)]. SEQSpark was also used to analyze 50,000 simulated exomes and it required 1.75 hours for the analysis of a quantitative trait using several rare variant aggregate association methods. Additionally, the performance of SEQSpark was compared to Variant Association Tools and PLINK/SEQ. SEQSpark was always faster and in some situations computation was reduced to a hundredth of the time. SEQSpark will empower large sequence-based epidemiological studies to quickly elucidate genetic variation involved in the etiology of complex traits.

9 | Assessing the Causal Role of Body Mass Index on Cardiovascular Health in Young Adults: Mendelian Randomization and Recall-By-Genotype Analyses

Kaitlin H. Wade[1,2], Scott T. Chiesa[3], Alun D. Hughes[4], Nish Chaturvedi[4], Marietta Charakida[3], Alicja Rapala[3], Vivek Muthurangu[3], Tauseef Khan[3], Nicholas Finer[3], Naveed Sattar[5], Laura D. Howe[1,2], Abigail Fraser[1,2], Debbie A. Lawlor[1,2], George Davey Smith[1,2], John E. Deanfield[3], Nicholas J. Timpson[1,2]

[1]MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, Bristol, United Kingdom; [2]School of Social and Community Medicine, Faculty of Health Sciences, University of Bristol, Bristol, United Kingdom; [3]Vascular Physiology Unit, Institute of Cardiovascular Science, University College London, London, United Kingdom; [4]Cardiometabolic Phenotyping Group, Institute of Cardiovascular Science, University College London, London, United Kingdom; [5]Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, United Kingdom

Body mass index (BMI) and cardiovascular health are causally related in late life, but this has not been explored in younger ages. Using Mendelian randomization (MR) and “recall-by-genotype” (RbG) methodologies, we estimated the causal effect of BMI on cardiovascular health in young adults in the Avon Longitudinal Study of Parents and Children. For MR analyses, a genetic risk score (GRS) comprising 97 independent genetic variants was used to test the causal effect of each unit increase in BMI (kg/m^2) on selected cardiovascular phenotypes measured at age 17 (n = 7909). An independent sample from the same cohort participated in a RbG study at age 21, which enabled more detailed cardiovascular phenotyping (n = 418; 191/227 from the lower/upper ~30% of a genome-wide GRS). Difference in mean BMI between RbG groups was 3.85kg/m^2 (95% CI: 2.53, 4.63; $p = 6.09 \times 10^{-11}$). In both MR and RbG analyses, results indicated that higher BMI causes higher blood pressure and left ventricular mass (indexed to height$^{2.7}$, LVMI) in young adults (e.g. difference in LVMI per kg/m^2 using MR: 1.07g/m$^{2.7}$; 95% CI: 0.62, 1.52; $p = 3.87 \times 10^{-06}$ and per 3.58kg/m^2 using RbG: 1.65g/m$^{2.7}$ 95% CI: 0.83, 2.47; $p = 0.0001$). RbG results indicated a causal role of higher BMI on higher stroke volume (difference per 3.58kg/m^2: 1.49ml/m$^{2.04}$; 95% CI: 0.62, 2.35; $p = 0.001$) and cardiac output (difference per 3.58kg/m^2: 0.111/min/m$^{1.83}$; 95% CI: 0.03, 0.19; p = 0.01). Consistent with efforts to prevent or reverse obesity in the young, complementary MR and RbG causal methodologies showed that higher BMI is likely to cause worse cardiovascular health even in youth.

10 | Comparison of Methods for Transcriptome Imputation Through Application to Two Common Complex Diseases

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

Transcriptome imputation has become a popular method for integrating genotype and expression data to investigate the causal role of gene expression in complex traits. Here, we compare three approaches (PrediXcan, MetaXcan and TWAS/FUSION) via application to genome-wide association study (GWAS) data from the Wellcome Trust Case Control Consortium, focusing on Crohn's disease and type one diabetes (T1D). We investigate how the genes identified as significant by each approach compare with each other and with those identified through standard GWAS analysis, and how the effects used by the prediction models compare with known effects of genotype on expression. We find all approaches produce similar results when applied to the same data, although for a small subset of genes (mostly in the MHC) the approaches heavily disagree. We also observe that most associations detected by these methods occur near known GWAS risk loci, with few new discoveries found. PrediXcan and MetaXcan's models for predicting gene expression more consistently recapitulate known effects of genotype on expression, suggesting they are more robust than TWAS/FUSION. Application of these approaches to summary statistics from recent meta-analyses in Crohn's disease and T1D detects 53 significant associations of gene expression with Crohn's and 170 with T1D, providing insight into biological mechanisms underlying these diseases. We conclude that while current implementations of transcriptome imputation typically detect fewer associations than GWAS, they nonetheless provide an interesting way of examining GWAS results to determine causal genes, and that PrediXcan and MetaXcan are currently the most reliable ways to implement transcriptome imputation.

11 | Joint Fine Mapping of Multiple Related Diseases Increases Power Through Exploiting Shared Causal Variant Structure

Jennifer L. Asimit[1], Mary D. Fortune[1,2], Chris Wallace[1,2]

[1] MRC Biostatistics Unit; University Of Cambridge, Cambridge, United Kingdom; [2] Department of Medicine, University of Cambridge, Cambridge, United Kingdom

The underlying genetic contribution to many complex diseases and traits has been investigated with great success by genome-wide association studies (GWAS), which have led to the detection of hundreds of variants associated with a spectrum of diseases. Extended Linkage Disequilibrium (LD) and finite sample sizes complicate the fine mapping of causal variants. Fine mapping multiple causal variants is often approached through stepwise or, less often, through stochastic searches of the potential causal variant model space. We contrast these two approaches and provide examples where, as sample size increases, stochastic searches converge to the correct solution but stepwise do not; for example, this may occur when there is a single SNP in moderate LD with both of two distinct causal variants. This behaviour is explored via mathematical theory and extensive simulation studies, and we provide multiple real data examples.

However, current sample sizes remain a limiting factor for any high dimensional model search strategy. We propose to leverage information between diseases through joint analysis of data from related diseases in a novel Bayesian multinomial stochastic search framework, where prior model probabilities are formulated to favour combinations of models with a degree of sharing of causal variants between diseases. We use simulations and real data examples to illustrate the improved power in comparison to a marginal analysis of each disease.

12 | Estimating Indirect Effect when the Mediator is a Censored Variable in a Mediation Model

Jian Wang[1], Sanjay Shete[1,2]

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

A mediation model is a statistical method exploring the direct and indirect effects of an initial variable (X) on an outcome (Y) by including one or more mediators (M), which has been widely applied in many different fields. In practice, investigators can observe censored data. Currently, most approaches to mediation analysis with censored data focus on the censored Y but not censored M. In this study, we proposed an approach to estimate the indirect effect in a mediation model when the mediator is a censored variable, based on the accelerated failure time model and a multiple imputation approach. Using simulation studies, we first established the bias in estimating coefficients of different paths in the mediation model, including the effects of X on M $[a]$, of M on Y $[b]$ and of X on Y given mediator M $[c']$, as well as indirect effects when using the existing approaches, including a naïve approach, complete-case analysis, and the Tobit mediation model. We conducted simulation studies to investigate the performance of the proposed approach and compared it to that of the existing approaches. The proposed approach accurately estimates the coefficients of different paths, indirect effects and percentages of the total effects mediated. We applied the proposed approach, as well as the existing approaches, to investigate the mediation model of SNPs, age at menopause and fasting glucose levels.

13 | Identifying Positive Selection Associated with Antimalarial Drug Resistance in *Plasmodium falciparum* using Identity-By-Descent Analysis

Lyndal Henden[1,2], Stuart Lee[1,2], Ivo Mueller[1,2], Alyssa Barry[1,2], Melanie Bahlo[1,2]

[1] Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville Victoria, Australia; [2] Department of Medical Biology, University of Melbourne, Parkville Victoria, Australia

Drug resistance in microorganisms is a global health crisis and identifying the mechanisms underlying such resistance is crucial in advancing disease control and elimination efforts. Genes associated with resistance experience selective pressures, creating strong genetic signals in the microorganism's genome. Here we present a novel method for identifying loci under recent positive selection in microorganisms using identity by descent analysis. We apply our method to whole genome sequencing data of more than 2,000 *Plasmodium falciparum* isolates from Africa, Southeast Asia and Papua New Guinea. In doing so, we are able to identify many well-known signals associated with antimalarial drug resistance as well as several new loci suspected of being associated with resistance. Identity-by-descent analysis also allows us to explore population structure through relatedness networks, providing clues as to the number of haplotypes contributing to a selection signal and the distribution of these signals within and between countries. Furthermore, we are able to determine whether a haplotype conferring drug resistance has arisen independently between geographic locations or whether it has spread from other locations.

14 | Disease-Informed Bayesian Association Scan Reveals Novel Loci Associated with Human Lifespan and Linked Biomarkers

Aaron F. McDaid[1,2], Peter K. Joshi[3], Ninon Mounier[1,2], Eleonora Porcu[2,4], Andrea Komljenovic[2,5], Bart Deplancke[2,6], Marc Robinson-Rechavi[2,5], Johan Auwerx[7], James F. Wilson[3,8], Zoltán Kutalik[1,2]

[1] Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital, Lausanne, Switzerland; [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3] Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom; [4] Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [5] Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; [6] Laboratory of Systems Biology and Genetics, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, Lausanne, Switzerland; [7] Laboratory of Integrative and Systems Physiology, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; [8] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom

The enormous variations in human lifespan are in part due to a myriad of sequence variants, only a few of which have been revealed to date. Since many life-shortening events are related to diseases, we developed a Mendelian randomization-based (MR) method combining 58 disease-related GWA studies to derive disease-informed longevity priors for all SNPs genome-wide. Our Bayesian association scan, informed by these priors, for parental age of death in the UK Biobank study ($n = 116{,}279$) revealed 16 independent SNPs with significant Bayes factor at a 5% false discovery rate (FDR), 12 of which are novel. Eleven of them replicate (5% FDR) in five independent longevity studies combined. While most of them have pleiotropic effects, three have not been associated with any human trait to date. Interestingly, all but three of them have life-shortening alleles that are depleted in older Biobank participants. Further MR analysis at these 16 loci revealed that lower expression levels of *RBM6, SULT1A1* and *CHRNA5* in the brain might be causally implicated in longevity. Our follow-up animal experiment showed, consistently with the human results, that lower mRNA level of *RBM6* in the prefrontal cortex at 72 days of age was a strong predictor of shorter lifespan in the LXS mouse lines ($r2 = 0.45$, $P = 4 \times 10-4$). Furthermore, we found that *SULT1A1* expression levels are down-regulated upon lifespan-extending caloric restriction diet. Finally, genome-wide analysis revealed significant enrichment for the lipoprotein metabolism pathway ($P = 3 \times 10{-}6$) and largely shared genetics with extreme longevity (LD-score regression $rg = 0.73$).

15 | DoriTool: A Bioinformatics Integrative Tool for Post-Association Functional Annotation

Isabel Martín-Antoniano[1,2], Lola Alonso[1], Miguel Madrid[3], Evangelina López de Maturana[1], Núria Malats[1]

[1] Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre, Madrid, Spain; [2] Instituto de Medicina Molecular Aplicada, Facultad de Medicina, Universidad San Pablo, Madrid, Spain; [3] Structural Computational Biology Group, Spanish National Cancer Research Centre, Madrid, Spain

The emergence of high-throughput data in biology has increased the need for functional ‘in silico’ analysis and promoted the development of integrative bioinformatics tools to facilitate the obtention of biologically meaningful data.

In this paper we present DoriTool, a comprehensive, easy, and friendly pipeline integrating biological data from different functional tools. The tool was designed with the aim to maximize reproducibility and reduce the working time of the researchers, especially those with limited bioinformatics skills, and to help them with the interpretation of the results.

DoriTool is based on an integrative strategy, which is implemented by following a modular design pattern. DoriTool combines up-to-date functional and genomic data as well as third-party bioinformatics tools in a pipeline to perform “in silico” analysis of annotations at mutation/variant, gene, pathway and network levels using scripts written in bash, Perl, and the R programming language. DoriTool uses GRCh37 human assembly and online mode. DoriTool also provides nice visual reports including variant annotation, linkage disequilibrium proxies, gene annotation, gene ontology analysis, expression quantitative loci (eQTL) results from Genotype-Tissue Expression (GTEx) and coloured pathways. Here we show also DoriTool functionalities by applying it to a dataset of 13 variants associated with prostate cancer. Project development, released code libraries, GitHub Repository (https://github.com/doritool) and documentation are hosted at https://doritool.github.io/.

DoriTool is, to our knowledge, a most complete bioinformatics tool offering functional ‘in silico’ annotation of variants previously associated with a trait of interest, shedding light on the underlying biology and helping the researchers in the interpretation and discussion of the results.

16 | Bayesian Generalized Least Squares in Multiethnic Fine-Mapping

Kan Wang[1], David V. Conti[1]

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

To follow up regions identified through GWAS, multiethnic fine-mapping can improve the ability to identify an underlying causal variant by leveraging different linkage disequilibrium (LD) structures across diverse populations. In this context, fixed-effect meta-analysis (FE) remains the most commonly used approach for its ease of interpretation and its ability to pinpoint the causal SNP, especially under the assumption of a common effect across populations. Alternative approaches mostly build upon the FE approach and leverage heterogeneity or incorporation of functional information. However, when applied in practice across numerous SNPs in a region, these approaches often lead to less clarity as evidence from SNPs with common effects are compared to SNPs with large heterogeneity. Here, we expand upon the FE and propose a Bayesian Generalized Least Squares (BGLS) that explicitly accounts for the covariance of estimates within each population. Determination of the most likely causal SNP is determined via a model averaging approach investigating all two-SNP combinations to yield SNP-specific posterior inclusion probabilities. We present a simulation study showing that for realistic effect sizes, BGLS out-performs FE by a large margin. The improvement is the most dramatic when the LD between the causal SNP and surrounding SNPs is modest to high. We investigate the performance for different study designs with varying numbers of populations, un-equal sample sizes, and multiple causal SNPs in a single region. Additionally, we demonstrate how external functional information can also be incorporated in BGLS to further improve its accuracy in prioritizing functional SNPs.

17 | Comparing the Effectiveness of Current Methods of Polygenic Score Measurement

Alexandros Rammos[1,2], Kevin J. Mitchell[1], Kristin K. Nicodemus[2,3]

[1] Smurfit Institute of Genetics and Institute of Neuroscience, Trinity College Dublin, Ireland; [2] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, United Kingdom; [3] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, United Kingdom

Polygenic risk scores (PRS) were introduced as a means to account for additive common variation and have been widely used in psychiatric disorder research. Over the years many different methods on the calculation of PRS have been proposed, but no consensus has been reached as to which method is optimal at capturing the variation while at the same time being biologically and statistically robust. By using simulated datasets based on an unrelated subset of the Generation Scotland dataset ($N = 7500$), we attempted to compare methodologies that are currently being used for PRS calculation including Linkage Disequilibrium (LD) Pruning, LD Clumping, weighted PRS on the basis of True Discovery Rate and finally the auto thresholding software PRSice. To investigate the differences between, as well as, within methods, we applied a number of different parameters, including sample size, LD structure and number of causal alleles. All methods underperformed and were unable to produce estimates close to the true value. LD pruning at low *P* value cut off outperformed LD Clumping and weighting. PRSice outperformed the other methods but after correcting for multiple testing failed to reach conventional levels of statistical significance. Sample size, LD structure and number of causal alleles heavily influenced estimates of the score. Optimal estimates were achieved in LD rich regions with larger sample sizes and with a smaller number of causal alleles with larger effects. These results suggest that PRS, as they are currently being calculated, might be underestimating the effects of common additive variation.

18 | Multivariate Generalized Linear Model for Genetic Pleiotropy

Daniel J. Schaid[1], Xingwei Tong[2], Anthony Batzler[1], Jason P. Sinnwell[1], Jiang Qing[2], Joanna M. Biernacka[1]

[1] Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America; [2] School of Statistics, Beijing Normal University, Beijing, China

When a single gene influences more than one trait, known as pleiotropy, it is important to detect pleiotropy to improve the biological understanding of a gene. Yet, most current multivariate methods to evaluate pleiotropy test the null hypothesis that none of the traits are associated with a variant; departures from the null could be driven by just one associated trait. A formal test of pleiotropy should assume a null hypothesis that one or fewer traits are associated with a genetic variant. We have developed statistical methods to analyze pleiotropy for analysis of binary, ordinal, or quantitative traits, or a mixture of these types of traits, based on generalized linear models and estimating equations. Our framework provides a sequential approach to test the null hypothesis that k+1traits are associated, given that the null of k traits are associated was rejected. This provides a method to determine the number of traits associated with a genetic variant, as well as which traits while accounting for correlations among the traits. By simulations, we illustrate the Type-I error rate and power of our new methods, describe how they are influenced by sample size, the number of traits, and the trait correlations, and apply the new methods to a genome-wide association study of multivariate traits. Our new approach provides a quantitative assessment of pleiotropy, enhancing current analytic practice.

19 | Discovery and Fine-Mapping of Type 2 Diabetes Susceptibility Loci Across Diverse Populations

Jennifer E. Below[1], Hidetoshi Kitajima[2], Anubha Mahajan[2], Xueling Sim[3], Maggie Ng[4], Weihua Zhang[5], Daniel Taliun[6], Kyle J. Gaulton[7], Andrew P Morris[1,8], on behalf of the DIAMANTE Consortium

[1] Vanderbilt Genetics Institute, Vanderbilt, University of Texas Health Science Centre at Houston, Houston, Texas, United States of America; [2] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [3] Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore; [4] Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina, United States of America; [5] School of Public Health, Imperial College London, London, United Kingdom; [6] Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America; [7] Department of Pediatrics, University of California San Diego, La Jolla, California, United States of America; [8] Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom

We conducted trans-ethnic meta-analysis of genome-wide association studies of type 2 diabetes (T2D) in 99,265 cases and 545,212 controls from diverse populations. We identified 110 loci at genome-wide significance ($p<5 \times 10^{-8}$), including 37 mapping outside regions previously implicated in the disease, with the strongest novel associations at/near *INHBB* (rs58884021, $p = 2.8 \times 10^{-12}$), *PLEKHA1* (rs2421016, $p = 3.2 \times 10^{-12}$), and *EIF5A2* (rs6804915, $p = 3.8 \times 10^{-12}$). We identified 156 distinct association signals ($p<10^{-5}$) across the 110 loci, including 11 at *KCNQ1*, 5 at *INS-IGF2*, and 4 each at *CDKN2A-B* and *CCND2*. Whilst allelic effects on T2D risk of index variants were predominantly consistent across populations, for the first time we observed strong evidence of heterogeneity that was correlated with ancestry at *LEP* (rs7778167, $p_{\text{HET}} = 8.2 \times 10^{-16}$, East Asian specific), *UBE2E2* (rs35352848, $p_{\text{HET}} = 4.2 \times 10^{-11}$, effect strongest in East Asians), and *KCNQ1* (rs11819853, $p_{\text{HET}} = 2 \times 10^{-10}$, varying direction and magnitude of effect between ethnic groups). Fine-mapping analyses substantially improved localisation of potential causal variants compared with previous efforts, highlighting 17 signals for which a single variant accounted for >99% of the posterior probability of driving the association, with the most precise resolution at *JAZF1* (rs10226758), *CDC123-CAMK1D* (rs11257655), *TCF7L2* (rs7903146), and *KCNQ1* (rs2237884 and rs2237895). Integration of fine-mapping data and annotation revealed the posterior probability (π) to be significantly enriched in coding exons ($p = 1.4 \times 10^{-5}$), for the first time including an index variant at the *APOE-TOMM40* locus, *APOE* p.Cys130Arg (rs429358, $\pi = 99.2\%$). After accounting for coding variation, the posterior probability was also significantly jointly enriched for transcription factor binding sites for PDX1 ($p = 2.6 \times 10^{-6}$) and FOXA2 ($p = 1.8 \times 10^{-5}$), highlighting potential regulatory signatures that are predictive of causal variants for T2D in non-coding sequence.

20 | The Role of Coding and Low-Frequency Variants Contributing to Anthropometry

Anne E. Justice[1] on behalf of the GIANT, CHARGE, BBMRI-NL, and GoT2D Consortia

[1] Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Anthropometric traits, such as body mass index (BMI), waist-to-hip ratio (WHR), and height, are highly heritable, polygenic traits. There are nearly 1000 common (minor allele frequency [MAF]>5%) GWAS-identified variants across these traits, but the genetic underpinnings may include rare and protein-coding variants with large effects. We aimed to identify protein-altering (coding/splicing), low-frequency variants (LFVs) (MAF<5%), and genes influencing these three anthropometric traits using exome array data. We meta-analyzed study-specific association results for 216,883 LF, coding variants from up to 526,508 individuals. We took forward variants with P value$<2 \times 10^{-6}$ for validation in two independent studies (deCODE, UKBiobank) and subsequently meta-analyzed all samples ($N_{max} = 718{,}734$, ∼89% European). We conducted gene-based tests using the SKAT method including 16,222 genes defined by LFVs predicted as damaging. We identified 83 (72 novel) LFVs associated with height, 16 (11 novel) with BMI, and 13 (nine novel) with WHR that achieved array-wide significance (P value$<2 \times 10^{-7}$, 0.05/#variants). In general, LFVs exhibit larger effect estimates than previously recorded, in some cases exhibiting a 10-fold increase in effect estimates compared to the average common GWAS SNP. We identified 10 genes associated with height and one, *GIPR,* associated with BMI. Given the samples sizes available for these studies, it is not surprising that we present the largest set of validated coding and LFVs associated with complex human traits. In light of our very large sample and comprehensive study design, we will reflect upon the promise and potential limits of further studies examining the role of coding and LFVs for complex traits.

21 | X-Chromosome Association on Microbiome Data

Osvaldo Espin-Garcia[1,2], Wei Xu[1,3]

[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 Biostatistics, Princess Margaret Cancer Centre, Toronto, Canada

Analysis of the X-chromosome (X-chr) has been largely neglected in genetic studies mainly due to the complex underlying biological mechanisms. We propose a novel approach to infer association between microbiome data with excess of zeros and host genetic variants in the X-chr. The method accounts for X-chr inactivation (XCI), escape of XCI (XCI-E) and skewed XCI (XCI-S). Inference is carried out via semi-parametric maximum likelihood (SPML) in which the "true" biological mechanism is treated as an unobserved missing category. An expectation-maximization (EM) algorithm on zero-inflated models is implemented to estimate genetic effects. We perform simulations to assess the performance of the SPML approach compared to Clayton-like (CL) or PLINK-like (PL) analyses. Briefly, CL assumes XCI for every genetic marker in the X-chr, i.e. codes males as homozygous females, whereas PL analyzes every genetic marker under XCI-E, i.e. codes males as heterozygous females. Preliminary results suggest that the proposed method can render reduced bias compared to CL or PL. We aim to further explore tests of hypothesis under an efficient score statistic. The proposed method has far-reaching applications. In particular, we illustrate its usage on a large-scale human microbiome study, the GEM project, to explore X-chr wide genetic association.

22 | A Bayesian Hierarchical Model for Pathway Analysis with Simultaneous Inference on Pathway-Gene-SNP structure

Lei Zhang[1], Pankaj K. Choudhary[1], Swati Biswas[1]

[1] Department of Mathematical Science, University of Texas at Dallas, Richardson, Texas, United States of America

Pathway analysis is an approach that allows joint consideration of multiple SNPs belonging to multiple genes, which in turn belong to a biologically defined pathway. This analysis is usually more powerful than single-SNP analyses for detecting joint effects of multiple variants in a pathway, each with a modest effect. We develop a Bayesian hierarchical model that fully models the three-level hierarchy, namely, SNP-gene-pathway that is naturally inherent in the structure of the pathways, unlike the current methods that use ad hoc ways of combining such information. To handle the high dimensionality involved in such modeling, we regularize the effects at each level through appropriate choice of hierarchical priors. A key advantage of the joint modeling is that not only can we find associated pathways but also the associated genes within the significant pathways, and the associated SNPs within the significant genes. Such a formal mechanism for testing of components of a significant pathway is not available in the current methods yet is useful for follow-up studies. Moreover, we can test multiple pathways through one single joint model. We use Hierarchical False Discovery Rate for multiplicity adjustment of the entire inference procedure. To study the proposed approach, we conduct simulations with samples generated under realistic linkage disequilibrium patterns obtained from the HapMap project. We find that our method has higher power than some current approaches for identifying pathways with multiple modest-sized variants. In some settings, it has reasonable power to detect associated genes, a feature unavailable in other methods.

23 | Improved Genotype imputation in Disease-Relevant Regions with Inclusion of Patient Sequence Data: Lessons from Cystic Fibrosis

Naim Panjwani[1], Bowei Xiao[1], Lizhen Xu[2], Jiafen Gong[1], Katherine Keenan[3], Fan Lin[1], Gengming He[1], Zeynep Baskurt[1], Lin Zhang[4], Sangook Kim[5], Mohsen Esmaeili[1], Scott Blackman[6], Harriet Corvol[7,8], Mitchell Drumm[9,10], Michael Knowles[11], Garry Cutting[6,12], Johanna M. Rommens[1,13], Lei Sun[4,5], Lisa J. Strug[1,2,4]

[1]Program in Genetics and Genome Biology, The Hospital for Sick Children Toronto, Canada; [2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada; [3]Program in Physiology and Experimental Medicine, The Hospital for Sick Children, Toronto, Canada; [4]Department of Statistics, University of Toronto, Toronto, Canada; [5]Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [6]Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America; [7]Pediatric Pulmonary Department, Hospital Trousseau, Assistance Publique-Hôpitaux de Paris (AP-HP), Institut National de la Santé et la Recherche Médicale (INSERM), U938, Paris, France; [8]Pierre et Marie Curie University–Paris 6, Paris, France; [9]Department of Pediatrics, Case Western Reserve University, Cleveland, Ohio, United States of America; [10]Department of Genetics, Case Western Reserve University, Cleveland, Ohio, United States of America; [11]Cystic Fibrosis Pulmonary Research and Treatment Center, University of North Carolina, Chapel Hill, North Carolina, United States of America; [12]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America; [13]Department of Molecular Genetics, University of Toronto, Toronto, Canada

Genotype imputation improves fine mapping and enables meta-analysis of cohorts genotyped on different platforms. Traditional imputation uses whole genome sequence (WGS) reference panels e.g. 1000 Genomes Project (1KGP, n = 2,504). However, imputation with the 1KGP failed to impute variants in the cystic fibrosis (CF) transmembrane conductance regulator (*CFTR*) for the International CF Gene Modifier Consortium, genotyped on the Illumina Human660W-Quad BeadChip (n = 1,995). *CFTR* displays significant allelic heterogeneity and is associated with CF co-morbidities of complex inheritance such as intestinal obstruction. Using larger reference panels such as the Haplotype Reference Consortium (HRC; n = 32,470), or alternatives that incorporate in-sample WGS may be preferable. We compare imputation in *CFTR* using the HRC to a composite reference combining the 1KGP with WGS from 101 CF patients; the latter enriching the reference with study-specific haplotypes. The 1KGP, HRC and composite reference provide, respectively, 1,438, 2,164 and 2,439 biallelic variants for imputation in *CFTR*. The composite reference results in more variants imputed; greater concordance between imputed and known CF-causing mutations, e.g. W1282X with 98.3% vs. 96.6% in HRC; and imputation of the most common CF mutation (p.Phe508del) missed by HRC which does not support indels, a consequence of low coverage sequencing. Notably, the composite reference displayed greater ability to detect an association between *CFTR* and intestinal obstruction over 1KGP or HRC panels. Results suggest that traditional imputation can omit the most disease-relevant genotypes when there is allelic heterogeneity at causal loci, but incorporating WGS on a subset of the study population can improve imputation and causal variant identification.

24 | Summary Statistic GWAS Joint Re-Analyses Across 30+ Traits

Carla Lasry[1], Vincent Guillemot[1], Pierre Lechat[1], Herve Menager[1], B.J. Vilhjalmsson[2,3], Hugues Aschard[1,3]

[1]Département de Génomes et Génétique, Centre de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur, Paris, France; [2]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Genome-wide association studies (GWAS) have proven successful in identifying thousands of significant genetic associations for multiple traits and diseases. This success is largely thanks to the dramatic increase in sample size achieved by GWAS meta-analysis consortia, now allowing for the detection of genetic variants explaining as low as 0.02% of quantitative outcomes. However, GWAS meta-analyses across different diseases and traits have received limited attention, even though multivariate analyses have the potential to improve the detection of genetic variants. Here we propose Joint Analysis of Summary Statistics (JASS), a computationally efficient framework for the joint analysis of multiple phenotypes based on GWAS summary statistics. Our framework solves several practical and methodological issues which have been overlooked in previous studies. In particular, we identify realistic situations resulting in biased GWAS covariance estimates, demonstrate how this can lead to increased false positive rate in practice, and propose a set of guidelines and statistics to avoid such pitfalls. We applied JASS for the joint analysis of publicly available GWAS of more than 30 phenotypes, altogether, or using sub-groups of phenotypes based on their clinical or pathological features. We detected dozens of genome-wide significant variants missed by univariate screenings. The identified variants include a number of established candidates and SNPs recently discovered in sample size larger than those available in our analysis, thus confirming the capabilities and validity of our approach. Finally, we present a publicly available online implementation of JASS, allowing researchers to conduct the joint analysis of any specific sub-set of phenotypes.

25 | Multi-Phenotype Genome-Wide Meta-Analysis of Lipid Levels and BMI in 64,736 Europeans Suggests Shared Genetic Architecture

Marika Kaakinen[1,2], Reedik Mägi[3], Vasiliki Lagou[4,5], Annique Claringbould[6], Kyle Gaulton[7], BIOS Consortium, Krista Fischer[3], Andrew Morris[8], Inga Prokopenko[1], for the ENGAGE Consortium

[1]*Department of Medicine, Division of Experimental Medicine and Toxicology, Imperial College London, United Kingdom;* [2]*Department of Genomics of Common Disease, Imperial College London, United Kingdom;* [3]*Estonian Genome Center, University of Tartu, Tartu, Estonia;* [4]*VIB Center for Brain and Disease Research, Leuven, Belgium;* [5]*KU Leuven, Department of Microbiology and Immunology, Leuven, Belgium;* [6]*Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands;* [7]*Department of Genetics, Stanford University, Stanford, California, United States of America;* [8]*Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom*

Serum lipid levels and obesity share biochemical pathways, suggesting shared genetic factors. Genome-wide association studies (GWAS) of correlated phenotypes could be performed simultaneously to identify such shared genetic effects with increased power. We performed a multi-phenotype GWAS (MP-GWAS) on three blood lipids (high-/low-density lipoprotein cholesterol and triglycerides [HDL-C/LDL-C/TG]) and body-mass index (BMI). We imputed each of 22 European-ancestry contributing studies to the 1000 Genomes reference panel (Phase 1). We performed the MP-GWAS in up to 64,736 individuals by fitting a "reverse regression" model between each SNP and the linear combination of HDL-C/LDL-C/TG and BMI using the SCOPA software, i.e. $\mathrm{SNP_i} = \beta_{1\mathrm{i}} \times$ HDL-C$+\beta_{2\mathrm{i}} \times$ LDL-C$+\beta_{3\mathrm{i}} \times$ TG$+\beta_{4\mathrm{i}} \times$ BMI$+\varepsilon_{\mathrm{i}}$, where $i = 1,\ldots,\mathrm{n}$, and n is the maximum number of SNPs tested and $\varepsilon_{\mathrm{i}} \sim \mathrm{N}(0,\sigma^2)$. Study-specific variance-covariance matrices for each variant were combined in a meta-analysis using the META-SCOPA software. Enhanced by the improved power from joint analysis, we identified 14 novel common variant loci at genome-wide significance ($P<5 \times 10^{-8}$), and detected 41/nine established lipid/BMI loci, respectively. The *SDC1*, *SLC8A1*, *EPHA6*, *SPATA4*, *MAGI2*, *CTSB, BC014119, SMCO4* and *CNTN5* loci showed effects on both lipids and BMI in the joint model, suggesting shared genetic architecture. We supported this observation through hierarchical cluster analysis, which resulted in three clades representing a mixture of lipid- and BMI-associated variants. We detected significant expression quantitative loci (eQTL) effects in whole blood (N = 2,114) at six novel loci and enrichment of association signals at HDAC6 binding sites, indicating a critical role of associated loci in various cellular events. MP-GWAS enables detection of multi-phenotype effects and has increased power compared to single-phenotype GWAS.

26 | Novel Agglomerative Partitioning Framework for Dimension Reduction of High-Dimensional Genomic Datasets

Joshua Millstein[1], Duncan Thomas[1], Yang Yu[1], Wendy Cozen[1]

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

A common feature across genomic data types, including genome, epigenome, transcriptome, microbiome, metabolome, etc., is dependencies among variables. Improvements in genomic technologies accompanied by decreasing costs have led to vastly increasing amounts of information collected from individual tissue samples. However, this increase in information is often accompanied by increasing dependencies among variables. This dynamic has fueled the need for methods to reduce dimensionality of datasets by summarizing multiple dependent variables into fewer and less dependent variables. Dimension reduction has multiple benefits including reduced computational demands, reduced multiple-testing challenge, and better-behaved data. However, few existing dimension reduction methods meet several critical criteria, (i) minimal information loss, (ii) each group of dependent variables in the full dataset results in a single variable in the reduced dataset, (iii) the maximum amount of information loss from the formation of a single variable from multiple variables is specifiable by the researcher, and (iv) the approach is scalable to high dimensions. We propose a formal framework that accommodates these criteria. Two novel computationally efficient algorithms are described based on an agglomerative strategy, forming a partition of the variables and summarizing each into a new variable with constrained loss of information. In simulated data with dependencies, we found that reducing dimensionality resulted in substantially *increased* power to detect associations with external variables. An application to real human gut microbiome data identified associations with diet and demonstrated high interpretability of summary variables.

27 | JEM: A Joint Test to Estimate the Effect of Multiple Genetic Variants on DNA Methylation

Chloé Sarnowski[1], Tianxiao Huan[2,3], Chunyu Liu[2,3], Chen Yao[2,3], Roby Joehanes[2,3,4], Daniel Levy[2,3], Josée Dupuis[1,2]

[1]*Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America;* [2]*The Framingham Heart Study, Framingham, Massachusetts, United States of America;* [3]*The Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, United States of America;* [4]*Hebrew SeniorLife, Harvard Medical School, Boston, Massachusetts, United States of America*

Several studies have investigated the influence of individual SNPs on DNA methylation (DNAm) across the genome. However, few methods have explored the joint regulation of DNAm by multiple SNPs at one CpG site.

We extended a hierarchical model based on a hybrid Bayesian pseudo-likelihood to propose a new CpG-based test (JEM) to assess the combined effects of SNPs on DNAm. We evaluated the performance of JEM using simulated data based on six chromosome 21 probes. We randomly sampled real methylation values for 2,000 individuals from 2,639 Framingham Heart Study (FHS) participants for each simulation replicate to form 450 nuclear families and simulated genotypes based on 1000 Genomes haplotypes. We restricted our analyses to low-frequency and common SNPs and used a 10kb window to assign SNPs to probes. Under H_1, 5% of SNPs were randomly selected to be causal per probe. Performance of JEM was compared to famSKAT. We also applied JEM to candidate methylated genes in FHS families.

Both methods had correct type I errors and an overall high power for the scenarios simulated. Performances were comparable for probes with a high number of SNPs ($N \geq 100$), but JEM outperformed famSKAT for probes with few SNPs ($N<30$). Estimated JEM bias was low (mean squared error = 2.4×10^{-5}).

JEM provides a global test of association and estimates the individual contribution of each SNP on DNAm. It is a flexible approach for binary and quantitative traits that can incorporate covariates and annotation features. Future work includes assessment of JEM performance under various scenarios.

28 | Population-Wide Whole-Genome Sequencing in an Isolated Cohort Reveals Rare Variant Burdens Associated with Multiple Quantitative Traits

Arthur Gilly[1], on behalf of HELIC Investigators[2]

[1]Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom; [2]A list of HELIC Investigators is available at https://www.helic.org/team.html

Low-depth sequencing and imputation can provide an accurate picture of common variation genome-wide, while population-wide very low-depth sequencing can further interrogate low-frequency variation. However deeper whole-genome sequencing data are needed to fully inform genetic association studies at the rare end of the allelic spectrum. Isolated populations offer power gains in detecting associations in rare and low-frequency variants. Here, we sequence 1,457 individuals at an average depth of 22.6 × from an isolated cohort from Crete, Greece for which low-depth and very-low depth whole genome sequencing is also available. We test 13,449,852 SNPs with minor allele count (MAC) >10 for association with 48 quantitative traits, and report 29 independent signals across 24 traits at the 5×10^{-8} significance level, including 6 previously reported associations with haematological and lipid traits. For gene-based approaches, we test exonic and regulatory variants associated with 19,025 protein-coding genes reported by GENCODE V25 (GRCh38). We benchmark 12 different pipelines using different regions of interest (exonic, exonic and regulatory and regulatory only), variant weights and filters. We report 29 genome-wide significant ($P<1.3 \times 10^{-7}$) rare variant burden signals not driven by a single SNP, including for known loci, such as *ADIPOQ* for adiponectin ($P = 9.1 \times 10^{-8}$), *APOA1* and *APOC3* for HDL ($P = 2.12 \times 10^{-20}$ and 3.96×10^{-20}, respectively), *UGT1A10* for bilirubin ($P = 1.2 \times 10^{-8}$) as well as *HBB* and *HBE1* for multiple haematological traits ($P<10^{-50}$). Changing the region of interest gives rise to different signals, highlighting the importance of running genome-wide burden tests under multiple conditions which need to be carefully reproduced when seeking replication in external cohorts.

29 | Semiparametric Methods for Estimation of a non-Linear Exposure-Outcome Relationship Using Instrumental Variables in Mendelian Randomization

James R. Staley[1,2], Stephen Burgess[2,3]

[1]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; [2]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [3]MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom

Mendelian randomization, the use of genetic variants as instrumental variables (IV), can test for and estimate the causal effect of an exposure on an outcome. Most IV methods for estimating the association between an exposure and outcome implicitly assume that the relationship is linear. However, in practice, this assumption may not hold. Indeed, often the primary question of interest is to assess the shape of this relationship. We present two novel IV approaches for investigating the shape of the exposure-outcome relationship in individual-level concomitant data: a fractional polynomial method and a piecewise linear method. These methods rely on dividing the population into strata using the exposure distribution, and estimating a causal effect, referred to as a Localized Average Causal Effect (LACE), in each stratum of the population. The fractional polynomial method performs meta-regression across these LACE estimates. The piecewise linear method estimates a continuous piecewise linear function, the gradient of which is the LACE estimate in each stratum. Both modelling approaches performed well in simulations, yielding reasonable model fits to a variety of underlying data generating models. Using these methods, we identified strong non-linear causal effects of body mass index on diastolic and systolic blood pressure in UK Biobank. In summary, these novel IV approaches can be used to investigate the shape of exposure-outcome relationships in the context of Mendelian randomization, and are available in the nlmr R package (https://github.com/jrs95/nlmr).

30 | Analyses of Copy Number Variation in Cutaneous Melanoma Implicates its Functional Role in Gene Expression Regulation

Feifei Xiao[1], Xizhi Luo[1], Jeffrey E. Lee[2], Qingyi Wei[3], Guoshuai Cai[4], Christopher I. Amos[5]

[1]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, United States of America; [2]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America; [3]Department of Medicine, School of Medicine, Duke University, Durham, North Carolina, United States of America; [4]Department of Genetics, Geisel School of Medicine at Dartmouth College, Lebanon, New Hampshire, United States of America; [5]Department of Biomedical Science, Geisel School of Medicine at Dartmouth College, Lebanon, New Hampshire, United States of America

Cutaneous Melanoma (CM) is the most aggressive form of skin cancer and accounts for the majority of deaths from skin cancer worldwide. Genome-wide association studies have identified various susceptibility single nucleotide polymorphisms for CM but little is known about potential role of Copy Number Variations (CNVs). CNVs are a substantial source of genetic variation and their important roles have been directly implicated in several cancer types such as sporadic pancreatic adenocarcinoma. In this study, we hypothesize that some variability in cutaneous melanoma results from genetic variation mediated by CNVs. First, we examined the regulatory potential of aberrant copy numbers as a driver for CM risk. We test this using genetic intensities and phenotype data from a large case-control study composed of 2830 European Americans. CNVs were called genome-wide by a change-point model based statistical method for array-based CNV identification, modSaRa. Using gene-based collapsing test, we detected CNVs in several genes previously implicated in CM or DNA repair pathways, including *CELF1*, *ASXL3*, *CYP26B1*, *PLA2G6* and *FANCC*. Second, to elucidate the regulatory potential of copy numbers, we used TCGA melanoma data to examine the expression patterns of these genes in tumor tissues. Tumor samples with deletion in *CYP26B1* gene were significantly down regulated in gene expression compared to those with diploid. Both *PLA2G6* and *FANCC* presented up-regulation for samples with copy number gain and down-regulation for copy number loss. In addition to providing a comprehensive analysis of the relationship between CNVs and CM in this population, our study provides a unique source of information about CNVs important functional roles in cancer etiology through gene expression regulation.

31 | Influence of Lung Development Genes on Lung Function in Adults: Application of a Bayesian Model to UK Biobank Data

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

[1]National Heart and Lung Institute, Imperial College London, London, United Kingdom; [2]Department of Health Sciences, University of Leicester, Leicester, United Kingdom

Low lung function, as measured by the Forced Vital Capacity (FVC) and the ratio of Forced Expiratory Volume in 1 sec (FEV1) over FVC, have been associated with increased risk of mortality in adults. Epidemiological studies have suggested a role of early life factors as determinants of lung function in adulthood and familial studies have reported a heritability of ∼40% for lung function. However, current findings from GWAS account for 6.4% and 14.3% of the expected heritability for FVC and FEV1/FVC, respectively.

To improve statistical power to detect novel variants, we focus on a set of 403 genes involved in lung development. We apply a Bayesian method that we previously developed which integrates external biological information in a joint SNP analysis and is being implemented as BioShrink, a R Shiny-based tool.

The method was applied to UK Biobank data on ∼112,277 participants and 177,880 SNPs. Standard association analysis was performed to obtain the top signals based on the lowest *P*-value and the Bayesian joint SNP analysis was performed to the top 20,000 SNPs after retrieving biological information.

Preliminary results identified 26 independent regions in 28 lung development genes associated with FVC with a P-value$<10^{-6}$, with 14 regions in 10 genes with a P-value $<5 \times 10^{-8}$. For FEV1/FVC, we found 45 independent regions in 34 genes with a P-value$<10^{-6}$, with 16 regions in 19 genes with a P-value $< 5 \times 10^{-8}$.

Currently, we are replicating the signals identified using the Bayesian method in ∼20,000 subjects from three independent cohorts and integrating the results with the standard SNP analysis.

32 | Whole Genome Sequence Association Analysis of Type 2 Diabetes and Glycemic Traits in Trans-omics for Precision Medicine (TOPMed)

Jennifer Wessel[1], Jennifer Brody[2], Bertha Hidalgo[3], Alisa Manning[4,5], on behalf of the Trans-Omics for Precision Medicine (TOPMed) Program Diabetes Working Group

[1]Departments of Epidemiology and Medicine, Indiana University, Indianapolis, Indiana, United States of America; [2]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [3]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, Alabama, United States of America; [4]Department of Medicine, Harvard University, Boston, Massachusetts, United States of America; [5]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, Massachusetts, United States of America

The majority of genetic variants significantly associated with type 2 diabetes (T2D) and glycemic traits reside in the non-coding genome, with many causal variants still unknown. We leveraged Whole Genome Sequence (WGS) phase 1 data from TOPMed to perform a (1) T2D WGS association (WGSA) pooled analysis and (2) meta- and pooled-analysis of fasting glucose (FG). WGSA analyses included samples with deep (>30 ×) sequence coverage in 5 cohorts, three

European-ancestry: Framingham Heart Study (N_{FG} = 3209, N_{T2D} = 4007); Old Order Amish Studies, N_{FG} = 980, Cleveland Family Study (CFS, N_{FG} = 197, N_{T2D} = 357), and two African American: Jackson Heart Study (N_{FG} = 2487, N_{T2D} = 3343), CFS (N_{FG} = 248, N_{T2D} = 332). We used mixed effects models adjusting for sex, age, with empirical kinship and/or principal components to adjust for relatedness and population structure. In multi-ethnic analyses, common (minor allele frequency [MAF]>0.05) variant associations (*P* value<5E−8) were identified at known loci with T2D: *TCF7L2* (rs7903146, *P*-value = 2.5E−11 and 7 additional variants); and with FG: *MTNR1B* (rs10830963, *P*-value = 2.5E−16; rs12792753, *P*-value = 1.4E-8), *GCK* (rs4607517, *P*-value = 1.16E−10, and 13 additional variants), and *G6PC2* (rs560887, *P*-value = 5.4E−10). Additional associations with T2D included 12 rare variants (MAF<.01) in six loci not previously described; including intergenic rs778917988 (MAF = 0.0003, *P*-value = 2.0E−8) near *SESN3*, a known glucose-homeostasis gene. Preliminary results suggest multi-ancestry WGSA can discover novel loci for complex traits. Work is ongoing to refine annotation for gene-based tests, perform fine-mapping, and extend into phase 2 and 3 data (N = 87,724).

POSTER PRESENTATIONS

33 | Using *IPCAPS* to Identify Fine-Scale Population Structure

Kridsadakorn Chaichoompu[1], Fentaw A. Yazew[1,2], Sissades Tongsima[3], Philip J. Shaw[4], Anavaj Sakuntabhai[5,6], Bruno Cavadas[7,8], Luísa Pereira [7,8], Kristel Van Steen[1,2]

[1]GIGA-R Medical Genomics - BIO3, University of Liege, Liege, Belgium; [2]WELBIO (Walloon Excellence in Lifesciences and Biotechnology), Brussels, Belgium; [3]Genome Technology Research Unit, National Center for Genetic Engineering and Biotechnology, Pathum Thani, Thailand; [4]Medical Molecular Biology Research Unit, National Center for Genetic Engineering and Biotechnology, Pathum Thani, Thailand; [5]Functional Genetics of Infectious Diseases Unit, Institut Pasteur, Paris, France; [6]Centre National de la Recherche Scientifique, Paris, France; [7]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal; [8] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal

SNP-based information is used in several existing clustering methods to detect shared genetic ancestry or to identify population substructure (Price 2006, Raj 2016). Here, we present an unsupervised clustering algorithm called the *iterative pruning to capture population structure* (*IPCAPS*). Our method supports ordinal data which can be applied directly to SNP data to identify fine-level population structure and it is built on the *iterative pruning Principal Component Analysis* (*ipPCA*) algorithm (Intarapanich 2009). The *IPCAPS* involves an iterative process using multiple splits based on multivariate Gaussian mixture modeling of principal components and Clustering expectation–maximization estimation as in Lebret et al. (2015). In each iteration, rough clusters and outliers are also identified using our own method called *RubikClust*. To evaluate the performance of our method, we tested different simulated data sets of 2–3 populations, 250 individuals per population, 10,000 independent SNPs in Hardy–Weinberg equilibrium, and F_{ST} = [0.0008,0.005], with 100 replicates for each data set. For real-life data sets, we applied the *IPCAPS* to Thai (Wangkumhang 2013), African (Triska 2015), and HapMap populations. Our method showed that a population classification accuracy was superior to the *ipPCA* in simulated scenarios of extremely subtle structure (F_{ST} = [0.001,0.005]). In case of Thai population, results to detect fine-level structure were obtained as well as in case of African and HapMap populations. We are convinced that the *IPCAPS* has a potential to detect fine-level structure and it will be important in molecular reclassification studies of patients once underlying population structure has been removed.

34 | A Fast and Accurate Algorithm to Test for Binary Phenotypes and its Application to PheWAS

Rounak Dey[1,2], Ellen M. Schmidt[1,2], Goncalo R. Abecasis[1,2], Seunggeun Lee[1,2]

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

The availability of Electronic Health Record (EHR)-based phenotypes allows for genome-wide association analyses in thousands of traits and has great potential to identify novel genetic variants associated with clinical phenotypes. We can interpret the Phenome-Wide Association Study (PheWAS) result for a single genetic variant by observing its association across a landscape of phenotypes. Because PheWAS can test 1000s of binary phenotypes, and most of them have unbalanced (case:control = 1:10) or often extremely unbalanced (case:control = 1:600) case-control ratios, existing methods cannot provide an accurate and scalable way to test for associations. Here we propose a computationally fast score test-based method that estimates the distribution of the test statistic using the saddlepoint approximation. Our method is much faster than the state of the art Firth's test (∼ 100 times). It can also adjust for covariates and control type I error rates even when the case-control ratio is extremely unbalanced. Through application to PheWAS data from the Michigan Genomics Initiative, we show that the proposed method can control type I error rates while replicating previously known association signals even for traits with a very small number of cases and a large number of controls.

35 | Multi-SKAT: A Generalized Framework for Testing Pleiotropic Associations of Rare Variants

Diptavo Dutta[1], Seunggeun Lee[1]

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

In genetic association analysis, a joint test of related multiple phenotypes can provide novel insights into the genetic architecture of complex diseases. Although a number of methods have been developed for multiple phenotype tests for common variants, only a few exist for rare variants. Here we develop a generalized framework based on multivariate linear mixed model (Multi-SKAT: Sequence Kernel Association Tests with Multiple phenotypes) for testing such pleiotropic associations of rare variants. Multi-SKAT models the distribution of effect sizes of the variants on the phenotypes through a particular kernel matrix and performs a variance component test of association. We also provide two efficient methods to combine the result across different kernels through the minimum *P*-value and weighted sum of squares approaches that handles model misspecification efficiently. In addition, the assumption of independent samples can be relaxed in Multi-SKAT by incorporating the kinship information in the model. From extensive simulation studies, we show that Multi-SKAT can improve power over a standard approach of aggregating single-phenotype test *P*-values while maintaining type-1 error rate. The relative performance of these tests depends on the number of associated phenotypes and correlation patterns however the omnibus tests had robust power regardless of the genetic model. Most of the published methods can be proved to be a special case of Multi-SKAT. Our method is computationally efficient and is applicable to genome-wide datasets. Further, real data analysis of 9 amino acid phenotypes identified newer associations in addition to those discovered by other competing approaches.

36 | Polymorphism of Genes Related to Hypertension: A Hospital-Based Study

Erlin Listiyaningsih[1], Nadira A. Putri[2], Anwar Santoso[1,3], Hananto Adriantoro[1,3]

[1]Harapan Kita National Cardiovascular Center Hospital, Jakarta, Indonesia; [2]Faculty of Biology, Universitas Gadjah Mada, Yogyakarta, Indonesia Department of Cardiology-Vascular Medicine; Faculty of Medicine – Universitas Indonesia, Jakarta, Indonesia

Single Nucleated Polymorphism (SNP) in several genes play a role in hypertension. These include rs17249754 and rs7136259 in ATPase, Ca+++ transporting, plasma membrane 1 (ATP2B1), rs1004467 in Cytochrome P450, family 17, subfamily A, Polypeptide 1 (CYP17A1), and rs11191548 in 5'-Nucleotidase Cytosolic II (NT5C2), rs2285666 in Angiotensin Converting Enzyme (ACE), and rs1801253 and rs1801252 in ß1-adrenergic receptor. We analyzed their association with hypertension in patients diagnosed with angina pectoris symptom who were suggested to undergo *Computed tomography coronary angiography (CTCA)* in the Harapan Kita Cardiovascular Center Hospital, Jakarta, Indonesia.

Polymorphisms were detected using TaqMan® SNP Genotyping Assay. Statistical associations were estimated using multiple logistic regression and their statistical significances were examined using chi-squared tests.

The crude odds ratio (OR) of minor allele of polymorphisms ATP2B1 rs17249754, ATP2B1 rs713625, CYP1741 rs1004467, NT5C2 rs11191548, ACE2 rs2285666, ADRB1 rs1801253, and ADRB1 rs1801252 were 1.21, 0.74, 0.71, 0.61, 4.28, 0.79, and 0.74, respectively. Odds ratios adjusted for age, gender, body mass index (BMI) class, smoking, and exercise were calculated in a multivariate analysis, and only age factor give slightly different result.

Carriers of a minor allele in the ACE2 SNP rs2285666 were at a significantly higher risk for hypertension than non-carriers (odds ratio = 4.3), while carriers of a minor allele in the ATP2B1 SNP rs17249754 had a considerably smaller risk (odds ratio = 1.2) than non-carriers. Surprisingly, minor alleles in other suspected genes provide protection for hypertension in the study population.

37 | A Meta Genome-Wide Association Study Identifies a Novel Locus for Cardiovascular Disease in Type 1 Diabetes

Sareh Keshavarzi[1], Angelo Canty[2], Barbara E.K. Klein[3], Orchard Trevor J.[4], Tina Costacou[4], Ronald Klein[3], Janet Snell-Bergeon[5,6], David Maahs[6,7], Rachel Grace Miller[4], Kristin E. Lee[3], Andrew D. Paterson[1]

[1]Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada; [2]Departments of Mathematics and Statistics, McMaster University, Hamilton, Canada; [3]Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [4]Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America; [5]Department of Pediatric Endocrinology, University of Colorado, School of Medicine, Aurora, Colorado, United States of America; [6]Barbara Davis Center for Diabetes, University of Colorado School of Medicine, Aurora, Colorado, United States of America; [7]Department of Paediatrics, Stanford School of Medicine, Stanford, California, United States of America

People with Type 1 Diabetes (T1D) are at high risk for Cardiovascular Disease (CVD) compared to those without diabetes, accounting for the majority of premature mortality. It has been estimated that heritable factors account for 30–60% of the variation in CVD risk.

To identify genetic variants associated with CVD in individuals with T1D, we carried out a meta-analysis of five genome-wide association studies (GWAS) with ~3200 T1D individuals of European descent with ~24 years of T1D duration (627

CVD cases, 2570 controls). We investigated genetic association of ~9.5M genotyped and imputed autosomal variants with any CVD event including fatal or non-fatal myocardial infarction, stroke, confirmed angina, revascularization or congestive heart failure. We computed association test results by logistic regression and Cox proportional hazard models assuming additive allelic effects.

We identified a new locus at 14q21.2 tagged by rs61998300 near *LINC00871* for CVD risk [P = 1.13×10^{-9}, OR (95% CI) = 2.74(1.97, 3.83)]. The effect sizes for this locus were similar in all studies, with no significant evidence for heterogeneity. The meta-GWAS of time to first CVD event also confirmed the significant effect of rs61998300 [P =4.6×10^{-10}, Hazard Ratio (95% CI) = 2.21(1.71, 2.86)]. These associations remained significant under a dominant model. However, this locus was not associated with CVD risk in subjects without diabetes [a meta-GWAS of ~185,000 cases and controls, P = 0.85(Nikpay M, et.al, 2015)].

These findings provide evidence for a T1D-specific association of rs61998300 with CVD risk. This association should be investigated in additional large diabetes cohorts.

38 | Reconstructing a Melanoma Data Set for Evaluating Differential Treatment Benefit According to Biomarker Subgroups

Jaya M. Satagopan[1], Alexia Iasonos[1], Joseph G. Kanik[1]

[1] *Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America*

Cancer is a disease of the genome. Recent work to understand the genomic landscape of different types of cancers and the role of environmental exposures on genomic alterations has led to the development of numerous targeted therapies approved by regulatory agencies. Clinical studies of some targeted therapies in relation to time to event outcomes have demonstrated that treatment-related improvement in outcomes occurs only after a certain time lag, resulting in failure of the proportional hazards assumption, and the magnitude of improvement may vary according to certain biomarkers such as specific somatic mutations. These characteristics warrant the development of novel statistical methods to measure treatment benefit and to evaluate the variation in treatment benefit across biomarker subgroups in order to inform patient management and the design and analysis of future studies. Statisticians need access to clinical data to develop and validate novel methods that are consistent with the observed characteristics of the data. However, clinical trial data are typically not available to researchers outside the study team, creating a barrier to methodology development for addressing the emerging needs of patient-oriented research. This barrier can be addressed to some extent by digitally reconstructing patient-level data from published studies. We describe the software packages and digitization steps required to reconstruct the time to event outcome and covariates such as treatment and biomarker status. We illustrate these steps using published Kaplan-Meier figures from a Phase III clinical trial of monotherapies with nivolumab and ipilimumab and their combination therapy and PD-L1 expression in metastatic melanoma.

39 | Joint Statistical Modeling of Multiple Phenotypes in Samples with Related Individuals

Zuoheng Wang[1]

[1] *Department of Biostatistics, Yale University, New Haven, Connecticut, United States of America*

Genetic association studies have routinely been conducted to search for variants associated with diseases and quantitative phenotypes. Clinical and epidemiological studies typically collect data on a set of correlated phenotypes that may share common environmental and/or genetic factors. Such phenotypes contain more information than univariate phenotypes. Thus joint modeling of multiple phenotypes can potentially have increased power to detect association and increased precision of parameter estimation than univariate analysis. In this study, we develop novel statistical methods for multivariate association mapping in samples that contain arbitrarily related individuals. The proposed methods are based on retrospective analysis that is less dependent on model assumptions on phenotypes, thus they are robust to trait model misspecification. The new methods can potentially accommodate the external biological information by integrating graphical models and multivariate analysis and are computationally affordable. We comprehensively evaluate the proposed methods by simulation studies and real data application.

40 | A Parallel Algorithm to Construct Whole-Genome Network of Gene Regulation

Chen Chen[1], Min Zhang[1], Dabao Zhang[1]

[1] *Department of Statistics, Purdue University, West Lafayette, Indiana, United States of America*

Genetical genomics experiments have been commonly conducted to understand gene-gene interactions. Genome-wide association studies of gene expressions lead to identification of cis-eQTL and trans-eQTL, an indirect interrogation of the functional landscape of gene regulation. Recently, structural equations have been proposed to model the linear interactions between genes, with a potential to directly reveal whole-genome networks of gene interactions. However, no algorithm is available to directly build a whole-genome network of gene regulation. We propose a two-stage penalized least squares method for such a purpose. The algorithm fits one linear model for each gene at each stage, i.e., predicting the gene expression via genotypic values at the first stage and identifying the regulatory genes by associating target gene expression with the predicted gene expressions. Independent tasks of fitting linear models at each stage allow for parallel computation, making it computationally fast to construct whole-genome networks. We demonstrated the effectiveness of the method by conducting simulation studies, showing its improvements over other methods. Our method was applied to construct some gene regulatory networks with real genetical genomics data.

42 | Exact Tests of Zero Variance Component in Presence of Multiple Variance Components with Application to Longitudinal Microbiome Study

Jing Zhai[1], Kenneth Knox[2], Homer Twigg[3], Hua Zhou[4], Jin J. Zhou[1]

[1]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona, United States of America; [2]Division of Pulmonary, Allergy, Critical Care, Sleep Medicine, Department of Medicine, University of Arizona, Tucson, Arizona, United States of America; [3]Division of Pulmonary, Critical Care, Sleep, and Occupational Medicine, Indiana University Medical Center, Indianapolis, Indiana, United States of America; [4]Department of Biostatistics, University of California Los Angeles, Los Angeles, California, United States of America

In the metagenomics studies, testing the association of microbiome composition and clinical conditionals has been translated into testing the nullity of variance components. Such regression based methods that directly regresses the outcome on the microbiome profiles while allowing for easy covariate adjustment enjoys great popularity. Score-based tests with asymptotic null distribution are major tools due to its computational efficiency. However such tests were only correct when the model under the null hypothesis has only one error variance parameter and when sample sizes are large. When the null model has more than one variance parameter and when sample sizes are limited, e.g., longitudinal metagenomics studies, testing zero variance components remains an open challenging. In this paper, we propose a series of efficient exact tests (score test, likelihood ratio test, and restricted likelihood ratio test) of testing zero variance components when multiple variance components present. Our tests can be used to test the association of overall longitudinal microbiome profile and outcome in a longitudinal design. They can also be used to detect association of one specific microbiome cluster while adjusting for the effects from related clusters. This can be particularly useful to find causal microbiome clusters. Our approach combines the previous proposal of exact tests with single variance component under the null hypothesis with the strategy of reducing the multiple variance components to single one. It does not rely on the asymptotic theory of large samples, which can significantly boost the power of detecting association for small sample studies. Through simulation, we demonstrate our method has correct Type I error rate as well as superior power compared to existing methods. Finally, we apply our exact tests to a longitudinal pulmonary microbiome study of Human Immunodeficiency Virus (HIV) patients. We identify two interesting genera $Prevotella$ and $Veillonella$ associated with forced vital capacity, which shed lights on biological mechanisms. The software package is implemented in the open source, high-performance technical computing language {\sc Julia} and is freely available at \url{https://github.com/JingZhai63/VCmicrobiome}}

43 | Disease as Collider: A New Case-Only Method to Discover Environmental Factors in Complex Diseases with Genetic Risk Estimation

Félix Balazard[1,2], Sophie Le Fur[2,3], Pierre Bougnères[2,3], Alain-Jacques Valleron[2], the Isis-Diab Collaborative Group

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Paris, France; [2]INSERM U1169, Hôpital Bicêtre, Université Paris-Sud, Kremlin-Bicêtre, France; [3]Department of Pediatric Endocrinology, Hôpital Bicêtre, Kremlin-Bicêtre, France

Genetic risk scores can quantify part of the predisposition of an individual to a disease. The identification of environmental factors is more challenging. Collider bias appears between two causes (e.g. gene and environment) when conditioning on a shared consequence (the collider, disease).

We introduce Disease As Collider (DAC), a new case-only methodology to validate environmental factors using genetic risk. A complex disease is a collider between genetic and environmental factors. Under reasonable assumptions, a negative correlation between genetic risk and environment in cases provides a signature of a genuine environmental risk factor. Simulation of disease occurrence in a source population allows to estimate the statistical power of DAC as a function of prevalence of the disease, predictive accuracy of genetic risk and sample size. We illustrate DAC in 831 Type 1 Diabetes (T1D) patients.

The power of DAC increases with sample size, prevalence, and accuracy of genetic risk estimation. For a prevalence of 1% and a realistic genetic risk estimation, power of 80% is reached for a sample size under 3000. Power was low in our case study as the prevalence of T1D in children is low (0.2%).

DAC could provide a new line of evidence for discovering which environmental factors play a role in complex diseases, or validating results obtained in case-control studies. We discuss the circumstances needed for DAC to participate in the triangulation of environmental causes of disease. We

highlight the link with the case-only design for gene environment interaction.

44 | Electronic Health Record: An Untapped Resource for Family-Based Genetic Epidemiologic Research

Xiayuan Huang[1], Robert C. Elston[2], Guilherme J. Rosa[3], John Mayer[4], Zhan Ye[4], Terrie Kitchner[5], Murray H. Brilliant[5,6], David Page[1,7], Scott J. Hebbring[5,6]

[1] Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [2] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, United States of America; [3] Department of Animal Science, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [4] Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America; [5] Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America; [6] Department of Medical Genetics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America; [7] Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States of America

Pedigree analysis is a longstanding and powerful approach to gain insight into the underlying genetic factors in human health but identifying and recruiting families can be a challenge. Family-based studies have largely given way to population-based studies, but population-based studies have their own significant limitations. Development of high-throughput methods to foster family-based studies is necessary. To address this need, we developed a method that allowed us to identify 580,718 individuals linked to 173,368 pedigrees with high probability using data available in most electronic-health-record (EHR) systems. To demonstrate utility of such family data, we 1) quantified heritability for EHR extracted phenotypes, 2) developed logistic/linear regression of familial relatedness (LRFR) to assess the relationship between phenotypic concordance and genetic relatedness in EHR-linked families with significant data sparsity, 3) simulated the relationship between heritabilityand LRFR beta-coefficients, and 4) evaluated the utility of EHR-linked families for genetic mapping. Results indicate that heritabilityof human height was consistent with previous reports ($h^2 = 0.71$) but binary traits are a challenge. Conversely, studies show that LRFR beta-coefficients are highly correlated with h^2 and can effectively differentiate genetic from non-genetic binary phenotypes with sparse data ($P = 0.0051$). Lastly, we demonstrate that by using EHR-linked families connected to a biobank where only probands are recruited and genotyped provides added power for genetic mapping. The totality of these results emphasizes that EHR-linked families can uniquely enable classical genetic analyses in a high-throughput manner and supports the potential for a refocusing back towards family-based study designs for genetic epidemiologic research.

45 | Understanding Source of Prediction Power of Neural Network for Complex Disease in Omics Genetic Data through Stratified and Randomized Experimental Design

Xiaoxuan Xia[1,2], Rui Sun[1,2], Ruoting Men[1], Haoyi Weng[1,2], Benny Chung-Ying Zee[1,2], Maggie Haitian Wang[1,2]

[1] Centre for Clinical Research and Biostatistics, Division of Biostatistics, the Chinese University of Hong Kong, Shatin, Hong Kong SAR; [2] CUHK Shenzhen Research Institute, Shenzhen, China

Artificial neural networks have achieved impressive prediction power in many complex disease predictions. Many of the prediction algorithm consist of multiple layers and making use of omics data. Oftentimes it is not straight forward to tell which part of the algorithm or data contribute to the prediction power, which information is important for understanding the reproducibility of the algorithm and making clinical decisions. In this work, we performed stratified classifications and randomization at each step and produced an estimation of the source of the prediction power for different data type and prediction models. The findings exemplified a way to identify the source of explanation power in complex prediction models and omics data and shed light on the importance of thorough investigation of data architectures in clinical applications.

46 | Association of D7S2420 Marker with ARNSHL Non-Syndromic Deafness in Five Iranian Ethnic Groups

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

Autosomal Recessive Non-Syndromic Hearing Loss (ARNSHL) is the most common birth defect and it is a very heterogeneous trait and could be caused due to both genetic and environmental factors. *SLC26A4* gene is the most important risk factor for ARNSHL. In databases, several potential Short Tandem Repeat (STR) markers related to this region have been introduced. The purpose of this study was to examine the characteristics and informativeness of D7S2420 CA repeat (STR) marker in *SLC26A4* gene region in five ethnic groups of the Iranian population.

A total of 165 individuals from five different ethnic groups including Fars, Azari, Turkmen, Gilaki, and Arabs were genotyped using Polymerase Chain Reaction (PCR) followed by Polyacrylamide Gel Electrophoresis (PAGE) and fluorescent capillary electrophoresis. In this study, the results were analyzed by GeneMarker HID Human STR Identity software, GenePop program, and Microsatellite Tools software.

The allelic frequency revealed the presence of 10 alleles for D7S2420 marker in the Iranian population. An allele located at 137 base pairs in the D7S2420 locus was the most frequency with an allele frequency of 41.52%. The heterozygosity in all the ethnic groups was above 70%. The Turkmen ethnic group had the highest heterozygosity.

The analysis of polymorphism information content (PIC) demonstrated that D7S2420 marker was a highly informative marker in Iranian population (PIC value above 0.7). These findings indicate D7S2420 as a highly informative marker in diagnosis of *SLC26A4* based ARNSHL by Linkage analysis.

47 | Tobacco Smoking, Genes Involved in the Metabolism of Xenobiotics and Breast Cancer Risk

Takiy Berrandou[1], Emilie Cordina-Duverger[1], Claire Mulot[2], Patrick Arveux[3], Pierre Kerbrat[4], Pierre Laurent-Puig[2], Therese Truong[1], Pascal Guénel[1]

[1] Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France; [2] University Paris Descartes, Inserm UMR 5775 EPIGENETEC, Paris, France; [3] Centre Georges-François Leclerc, Côte d'Or Breast Cancer Registry, Dijon, France; [4] Centre Eugene Marquis, Rennes, France

Tobacco smoking and Xenobiotics Metabolism Pathway (XMP) genes involved in the metabolism of carcinogen compounds contained in tobacco smoke are suspected to play a role in Breast Cancer (BC) risk.

We have studied interactions between active or passive smoking and XMP genes in BC among 1125 cases of BC and 1172 population controls with lifelong data on smoking habits and valid genotyping data obtained from a dedicated chip.

The associations of BC with active/passive smoking and with 585 SNPs in 68 XMP genes, as well as interactions between genes and exposure to tobacco smoke were explored using the Adaptive Rank Truncated Product (ARTP) method. This approach allows investigating the role of genes (seen as a set of SNPs) and of the pathway (seen as a set of genes) in disease occurrence, and to gain in statistical power as compared to a SNP by SNP approach.

We reported an increased risk of BC among current smokers in postmenopausal women (OR = 1.46 [1.00-2.14]) and among never-smokers exposed to secondhand smoke for long duration (>20 years) (OR = 1.45 [1.01- 2.09]. Genetic variation in genes of the XMP pathway was significantly associated with premenopausal BC risk ($P_{ARTP} = 0.008$). We also observed significant interactions between XMP and smoking status / passive smoking duration, particularly via the CYP3A43, COMT, AKR1Cs and UGT1As genes.

The hypothesis that active and passive smoking increase BC risk is reinforced by our findings of an interaction with XMP genes. Further studies are needed to clarify the role of these genes in breast cancer.

48 | A Unified Partial Likelihood Approach for X-Chromosome Association on Time to Event Outcomes

Meiling Hao[1], Wei Xu[1,2]

[1] Department of Biostatistics, Princess Margaret Cancer Centre, Toronto, Canada; [2] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

For the expression of X-chromosome, it is undergoing three possible biological processes, namely X-chromosome Inactivation (XCI), escape of the X-chromosome inactivation (XCI-E) and skewed X-chromosome inactivation (XCI-S). To analyze the X-linked genetic association for survival data with the actual process totally unknown, we propose a unified approach of maximizing the partial likelihood over all of the potential biological processes. The proposed method can be used to infer the true biological process and derive unbiased estimates for the genetic association parameters. A partial likelihood ratio test that has been proved asymptotic chi-square distribution can be used to assess the genetic markers and adjusted for potential confounders. Furthermore, if the X-chromosome expression is under the XCI-S process, we can infer the correct skewed direction and magnitude of inactivation, which occupies significant meanings for the genetic mechanism. The developed methodology fills the gap of statistical methodology to explore X-linked genetic prognostic and predictive effect on time to event outcomes. Finite sample performance of this novel method is examined via extensive simulation studies. We implement the novel method on a randomized clinical trial study on colorectal cancer patients. The genetic association of X-chromosome wide genetic markers and overall survival has been evaluated.

49 | Incorporating Genetic Networks into Case-Control Association Studies with High-Dimensional DNA Methylation Data

Kipoong Kim[1], Hokeun Sun[1]

[1] Department of Statistics, Pusan National University, Busan, Korea

In human genetic association studies with high-dimensional microarray data, it has been well known that statistical methods utilizing prior biological network knowledge such as genetic pathways and signaling pathways can outperform other methods that ignore genetic network structures. In recent epigenetic research on case-control association studies, relatively many statistical methods have been proposed to identify cancer-related CpG sites and the corresponding genes from high-dimensional DNA methylation data. However, most of existing methods are not able to utilize genetic networks although methylation levels among linked genes in the networks tend to be highly correlated with each other. In this article, we propose a new approach that combines independent component analysis with network-based regularization to identify outcome-related genes for analysis of high-dimensional DNA methylation data. The proposed approach first captures gene-level signals from multiple CpG sites using independent component analysis and then regularizes them to perform gene selection according to given biological networks. In simulation studies, we demonstrated that the proposed approach overwhelms other statistical methods that do not utilize genetic network information in terms of true positive selection. We also applied it to the 450K DNA methylation array data of the four breast invasive carcinoma cancer subtypes from The Cancer Genome Atlas (TCGA) project.

50 | Genome-Wide Interaction Study of Smoking Behavior and Non-Small Cell Lung Cancer Risk in Caucasian Population

Yafang Li[1], Xiangjun Xiao[1], Younghun Han[1], Olga Y. Gorlova[1], David Christiani[2], Mattias Johansson[3], James D. McKay[3], Paul Brennan[3], Rayjean J. Hung[4], Christopher I. Amos[1]

[1] Biomedical Data Science Department, Dartmouth College, Hanover, New Hampshire, United States of America; [2] School of Public Health, Harvard University, Boston, Massachusetts, United States of America; [3] Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon, France; [4] The Lunenfeld-Tanenbaum Research Institute, Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Non-Small Cell Lung Cancer (NSCLC) contributes about 85% of lung cancer and both environmental and genetic risk factors are involved in lung cancer carcinogenesis. In this report, we conducted a genome-wide interaction analysis between SNPs and smoking status (never vs ever smokers) in European decent population. We adopted a two-step analysis strategy in the discovery stage: we first conducted a case-only interaction analysis to assess the relationship between SNPs and smoking behavior using 13,336 NSCLC cases; candidate SNPs with p-value<0.001 were further analyzed using a standard case-control interaction analysis including another 13970 controls. The significant SNPs with p-value$<3 \times 10^{-5}$ from the case-control analysis in discovery stage were further validated using an independent replication dataset comprising 5377 controls and 3054 NSCLC cases. We further stratified the analysis by histology subtypes. Two novel SNPs rs6441286 and rs17723637 were identified and the interaction odds ratio and meta-analysis p-value at these two SNPs were 1.24 with 6.96×10^{-7} and 1.37 with 3.49×10^{-7}, respectively. SNP rs4751674 was identified in squamous cell lung carcinoma subtype with an odds ratio of 0.58 and p-value of 8.12×10^{-7}. This study is by far the largest genome-wide SNP-smoking interaction analysis in lung cancer study. The three identified novel SNPs provide potential candidate biomarkers for lung cancer risk screening and intervention. The results from our study reinforce that gene-smoking interactions play important roles in the etiology of lung cancer and account for part of the missing heritability of this disease.

51 | Association Score Testing for Rare Variants and Binary Traits in Family Data with Common Controls

Mohamad Saad[1,2,3], Ellen M Wijsman[1,2]

[1] Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [2] Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, United States of America; [3] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Genome-wide association studies (GWAS) have been the predominant approach used to localize trait loci, mainly the common ones in the population. The multiple rare variants – common disease hypothesis was thought of as an important cause of the missing heritability. The advances of sequencing techniques and the decreasing of its costs coupled with methodological advances in the context of association with rare variants have made the study of rare variants at a genome-wide scale feasible. The resurgence of family-based association designs due to their advantage in studying rare variants has stimulated more method development, mainly based on linear mixed models. Other score tests can have more advantages over the linear mixed models, but they were mainly proposed for single marker association tests. In this paper, we extend several score tests $\left(X^2_{corrected}, W_{QLS}, M_{QLS}, SKAT\right)$ to the multiple variant association framework. We evaluate and compare its statistical performances along with the linear mixed model as baseline. Moreover, we show that three tests can be cast as the difference between allele frequencies of SNPs estimated in each of the group of affected and unaffected subjects. We show that these tests are very flexible since they can be based on related, unrelated, or both related and unrelated subjects. They also make feasible a design that only sequences a subset of affected subjects (related or unrelated) and uses for comparison publicly-available allele frequencies estimated in a group of healthy subjects. Finally, we show the great impact of linkage disequilibrium on the performance of all these tests.

52 | Ontogeny Related Changes in the Pediatric Liver Transcriptome

Brooke L. Fridley[1,3], Richard Meier[1], Charlie Bi[2], Roger Gaedigk[2], Shui Qing Ye[2], Dan Heruth[2], J. Steven Leeder[2]

[1] Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas, United States of America; [2] Division of Clinical Pharmacology and Therapeutic Innovation, Children's Mercy Hospital, Kansas City, Missouri, United States of America; [3] Department of

Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, United States of America

A major challenge in implementing personalized medicine in pediatrics is identifying the appropriate drug dosages for children. The majority of drug dosing studies are based on adult populations, with modification of the dosing for children based on size and weight. This rudimentary approach for drug dosing children is limited, as biologically a child can differ from an adult in far more than just size and weight. Specifically, understanding the ontogeny of childhood liver development is critical in dosing drugs that are metabolized through the liver, as the rate of metabolism determines the duration and intensity of a drug's pharmacologic action.

Therefore, we set out to determine pharmacogenes that change over childhood development, followed by a secondary agnostic analysis, assessing changes transcriptome-wide. We found evidence for transcripts in Very Important Pharmacogenes" (VIPs) *F5, ACE* and *SLC22A1* showing increased expression over the development period. The analysis of genome-wide changes detected transcripts in the following genes with significant changes in mRNA expression: *ADCY1, PTPRD, CNDP1*, *DCAF12L1* and *HIP1*. Gene set analysis determine ontogeny-related transcriptomic changes in the renin-angiotensin pathway, with lower expression of the pathway in general, observed in liver samples from younger subjects. Considering that this pathway plays a central role in blood pressure and plasma sodium concentration and our observation that *ACE* and *PTPRD* expression is increasing over the spectrum of childhood development, this finding could potentially impact the dosing of an entire class of drugs known as ACE-inhibitors in pediatric patients.

53 | Tissue-Specific Sexual Dimorphism in Autosomal Gene Expression

Irfahan Kassam[1], Yang Wu[1], Peter M. Visscher[1,2], Allan F. McRae[1,2]

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

Males and females differ in virtually all phenotypic traits. Genes located on the sex chromosomes are a well-known source of sexual dimorphism, however, autosomal gene expression also shows substantial differences across the sexes. This sexual dimorphism has been observed across different tissues and is often tissue-dependent. A recent study demonstrated that, on average, males and females share the same common genetic control of autosomal gene expression. Taken together, this suggests that tissue-specific sexual dimorphism is due to sex differences in the regulatory genome. To help us understand the biological mechanisms responsible for Tissue-Specific Sexual Dimorphism (TSSD) in human gene expression, we use data from the Genotype-Tissue Expression Project (GTEx) in $n = 449$ individuals across 39 tissues and test 32,432 autosomal genes that are common to at least two tissue-types for equality of effect size between the sexes across tissue-types. We identify 90 autosomal TSSD genes that show heterogeneity in effect size across tissue-types (uncorrected $p{<}2.30 \times 10^{-4}$, false-discovery rate of 5%), with the majority showing discordant direction of sexual dimorphism across tissues. A 100kb window around each of the TSSD genes shows enrichment for androgen and estrogen receptor sites, often in conjunction with annotated tissue-specific enhancers. TSSD genes also show a significant overlap with GWAS associations. This indicates that these TSSD genes may have functional consequences that contribute to downstream phenotypic differences observed across the sexes.

54 | Genome-wide Meta-analyses of Stratified Depression in Generation Scotland and UK Biobank

Lynsey S. Hall[1,2], Mark J. Adams[1], Aleix Arnau-Soler[3], Toni-Kim Clarke[1], Yanni Zeng[1,4], David J. Porteous[3], Ian J. Deary[5,6], Pippa A. Thomson[3,6], Chris S. Haley[4], Andrew M. McIntosh[1,6]

[1] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh United Kingdom; [2] Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; [3] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [4] Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [5] Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom; [6] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom

Few replicable genetic associations for Major Depressive Disorder (MDD) have been identified. However recent studies of depression have identified common risk variants by using either a broader phenotype definition in very large samples or by reducing the phenotypic and ancestral heterogeneity of MDD cases. Here, a range of genetic analyses were applied to data from two large British cohorts, Generation Scotland and UK Biobank, to ascertain whether it is more informative to maximize the sample size by using data from all available cases and controls, or to use a refined subset of the data - stratifying by MDD recurrence or sex. Meta-analysis of GWAS data in males from these two studies yielded one genome-wide significant locus on 3p22.3. Three associated genes within this region (*CRTAP*, *GLB1*, and *TMPPE*) were significantly associated in subsequent gene-based tests. Meta-analyzed MDD, recurrent MDD and female MDD were each genetically correlated with six of 200 health-correlated traits, namely neuroticism, depressive symptoms, subjective wellbeing, MDD, a psychiatric cross-disorder phenotype and Bipolar Disorder. Meta-analyzed male MDD showed no statistically significant correlations with these traits after correction for multiple testing. Whilst stratified GWAS analysis revealed a genome-wide significant locus for male MDD, the lack of independent replication, the equivalent SNP-based heritability estimates and the consistent pattern of genetic correlation with other health-related traits suggests that phenotypic stratification in currently available sample sizes is currently weakly justified. Based upon existing studies and our findings, the strategy of maximizing sample sizes is likely to provide the greater gain.

55 | Exploring the Genetic Architecture of nsCL/P

Laurence J. Howe[1], Gibran Hemani[1], George Davey Smith[1], Beate St. Pourcain[1,2], Sarah Lewis[1]

[1] Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, Bristol, United Kingdom; [2] Max Planck Institute for Psycholinguistics, Wundtlaan, Nijmegen, The Netherlands

Non-syndromic cleft lip/palate (nsCL/P) is a multifactorial complex disease with genetic and maternal environmental risk factors. Currently, there is a lack of knowledge about the genetic architecture of nsCL/P. European nsCL/P parent-offspring trios along with matched UK Biobank controls were used to explore the genetic architecture of nsCL/P using family based and unrelated case-control based methods. The Polygenic Transmission Disequilibrium Test (PTDT), which detects over-transmission of polygenic risk captured by markers on a genotyping chip from unaffected parents to affected offspring, was used to detect polygenicity. SNP heritability estimates were generated using Genome-wide Complex Trait Analysis (GCTA), Additive Variance Explained Method of Estimation (AVENGEME) and LD score regression.

The results of the PTDT suggest that nsCL/P has a polygenic component: affected nsCL/P offspring had 0.32 (95% C.I. 0.25, 0.39) standard deviations higher nsCL/P polygenic risk score than their unaffected parents ($p = 3.5 \times 10^{-18}$). LD score regression $h^2 = 0.33$ (95% C.I. 0.14, 0.51) and AVENGEME $h^2 = 0.20$ (95% C.I. 0.18, 0.22) generated comparable SNP estimates. However, GCTA estimates of SNP heritability ($h^2 > 0.47$) were likely inflated by batch differences between cases and controls. We conclude that nsCL/P has a polygenic component and it is likely that common SNPs explain between 20–35% of the variation in risk in Europeans.

56 | Gene-Based and Gene Set Enrichment Analyses Based on a Case-Control Genome-Wide Association Study of Multiplex Schizophrenia in Taiwan

Annemarie Lee Woolston[1], Po-Chang Hsiao[2], Chih-Min Liu[3,4], Hai-Gwo Hwu[1,3], Li-Ching Chang[5], Chien-Hsiun Chen[5], Jer-Yuarn Wu[5], Ming T. Tsuang[6,7,8], Wei J. Chen[1,2,3,4]

[1] Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan; [2] Genetic Epidemiology Core, Center of Genomic Medicine, National Taiwan University, Taipei, Taiwan; [3] Department of Psychiatry, College of Medicine and National Taiwan University Hospital, National Taiwan University, Taipei, Taiwan; [4] Graduate Institute of Brain and Mind Sciences, College of Medicine, National Taiwan University Hospital, National Taiwan University, Taipei, Taiwan; [5] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan; [6] Department of Psychiatry, Center for Behavioral Genomics and Institute for Genomic Medicine, University of California, San Diego, California, United States of America; [7] Department of Psychiatry, University of California, San Diego, California, United States of America; [8] Harvard Institute of Psychiatric Epidemiology and Genetics, Boston, Massachusetts, United States of America

Schizophrenia is a mental disorder resulted from complex genetic components underpinning the pathogenesis. The highly polygenic architecture, including susceptibility and modifier genes, results in genetic heterogeneity of schizophrenia. Inconsistent results between genome-wide association studies (GWAS) have reported possible associations between hundreds of genetic loci with variants and susceptibility of schizophrenia. This issue has been a major concern and revealed the importance of identifying confounding subtypes to increase the statistical power, such as patients with family history. This study aimed to identify genetic loci associated with schizophrenia in Taiwan. The associations of SNPs, genes, gene sets, and pathways with susceptibility for schizophrenia have also been investigated. A case-control GWAS was performed between 185 probands from multiplex families affected by schizophrenia and 925 community-based normal controls. Three strategies including 1) individual SNP association tests; 2) gene-based analysis; and 3) gene set enrichment analysis were used to interpret the resulting GWAS dataset. A total of 4 SNPs passed genome-wide significance, 48 SNPs reached suggestive threshold, *PECAM1* and *PIK3CG* as top genes, 7 gene ontology (GO) terms with *PECAM1* and 2 KEGG pathways with *PIK3CG* were identified. We have identified associations of SNPs, genes, gene sets, and pathways with susceptibility for schizophrenia. The genetic components with different dimensions may help further illuminate the pathogenesis of schizophrenia.

57 | Parent-of-Origin-Environment Interactions in Case-Parent Triads With or Without Independent Controls

Miriam Gjerdevik[1,2], Øystein A. Haaland[1,2], Julia Romanowska[1,3], Rolv T. Lie[1,2], Astanand Jugessur[1], Håkon K. Gjessing[1,2]

[1] Department of Global Public Health and Primary Care, University of Bergen, Norway; [2] Norwegian Institute of Public Health, Oslo, Norway; [3] Computational Biology Unit, University of Bergen, Bergen, Norway.

With access to case-parent triad data, it is often possible to deduce parental origin of the child's alleles. We can then estimate a parent-of-origin (PoO) effect, which we define as the ratio of relative risks associated with the alleles transmitted from the mother and the father, respectively. A probable cause of PoO effects is genomic imprinting; through DNA methylation in the germline, the expression of alleles might be silenced depending on their parental origin. Because environmental exposures may affect methylation patterns, one should extend methods for gene-environment interactions to search for differential effects of PoO across environmental strata (i.e., PoO × E interactions).

There has been little focus on PoO × E effects in the literature, and general methods for assessing such effects are needed. We have therefore developed a new and extensive framework to analyze PoO × E interactions. Our test enables a full GWAS screen of complete or incomplete case-parent triads with or without independent control triads. We analyze triads in each exposure stratum separately, applying maximum likelihood estimation in a log-linear model. PoO × E interactions are then tested using a Wald-based post-test of parameters across strata. Our framework includes a complete setup for power calculation, both through using direct power simulations and through using calculations based on the asymptotic variance-covariance structure. The approach is implemented in the R software package Haplin.

To illustrate our models, we analyzed data from a previously published GWAS on orofacial clefts to assess whether smoking during the periconceptional period modifies PoO effects on cleft palate only.

58 | Gender Differences in Brain-Derived Neurotrophic Factor (BDNF) Val66Met Variants and Stressful Life Events on Psychological and Metabolic Phenotypes

Rong Jiang[1], Michael A. Babyak[1], Beverly H. Brummett[1], Elizabeth R. Hauser[2], Brett C. Haberstick[3], Andrew Smolen[3], Ilene C. Siegler[1], Kathleen Mullan Harris[4,5], Redford B. Williams[1]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina, United States of America; [2]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, North Carolina, United States of America; [3]Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, Colorado, United States of America; [4]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina United States of America; [5]Department of Sociology, University of North Carolina, Chapel Hill, North Carolina, United States of America

Gender differences in the effect of *BDNF* Val66Met genotype on HPA axis response to a mental stress protocol have been found in our previous study, in which the Val/Val carriers had higher cortisol levels than Met carriers in women but not in men. Due to the extensive evidence of association between HPA axis and psychological and metabolic phenotypes, we hypothesized gender differences exist in the effect of Val66Met genotype by stress on the psychological and metabolic measurements among Caucasians ($n = 3207$) at Wave IV survey of the National Longitudinal Study of Adolescent to Adult Health. Stress was measured by an additive index (final score range 0–35) of stressful life events (SLE) assessed during the in-person interview, including only events of sudden onset within the 12 months preceding the interview. Significant three-way interactions of gender × Val66Met genotype × stress were observed in models predicting depressive symptoms (CES-D, $p = 0.029$), neuroticism (NEU, $P = 0.009$), systolic blood pressure (SBP, $p = 0.039$), diastolic blood pressure (DBP, $p = 0.024$), and heart rate (HR, $p = 0.056$). Follow-up analyses stratified by the gender showed that the two-way interaction of Val66Met genotype × stress was significantly associated with NEU ($p = 0.01$), SBP ($p = 0.039$), DBP ($p = 0.004$), and CES-D ($p = 0.0975$) only in women but not in men. These findings support our hypothesis that significant gender differences exist in the effect of Val/Val genotype by stress on psychological and metabolic phenotypes. These findings, if confirmed, may impact further research requiring consideration of gender along with the genotype and stress in clinical intervention studies.

59 | The effect of Brain-Derived Neurotrophic Factor (BDNF) Val66Met Variants on the Path from Psychosocial Stress to Depression, Body Mass Index and Pre-Clinical Atherosclerosis

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

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

Our previous work has shown that a functional *BDNF* SNP Val66Met (rs6265) was associated with endophenotypes related to higher risk of developing coronary heart disease (CHD). Women with the Val/Val genotype had larger blood cortisol response than Met women to a mental stress protocol. In two independent samples, the association of psychosocial stress with depressive symptom levels was greater in magnitude among Val/Val carriers than among Met carriers. Given that both stress and the Val66Met genotype have been associated with depressive symptoms and CHD endophenotypes such as body mass index (BMI) and pre-clinical atherosclerosis, we hypothesized that the association between psychosocial stress and atherosclerosis (indexed by common carotid intima-media wall thickness, CCIMT) would be mediated by depressive symptoms and BMI, and these mediating associations would differ depending on Val66Met genotype. Structure equation models with paths from psychosocial stress → depression → BMI→ CCIMT were compared across Val/Val and Met genotypes, in 2449 Caucasians from the Multi-Ethnic

Study of Atherosclerosis. Although the global test of stress × genotype interaction on the path was marginally significant ($p = 0.053$), the mediated effect of the whole hypothesized path was statistically significant in the Val/Val group ($p = 0.027$, $\beta = 0.072$, 95% CI = 0.008-0.136), but not in the Met group ($p = 0.846$). Depression and BMI together acted as the mediators of the association between psychosocial stress and CCIMT, but only in Val/Val group. The findings tentatively support our hypothesis that Val66Met genotype has an effect on the path from psychosocial stress to CCIMT and have the important clinical implication for preventive interventions.

60 | Comparison of Bayesian Network Methods for Estimating the Direction of Causality in Biological Data

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

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

Bayesian networks can be used to identify possible causal relationships in increasingly available biological data sets such as genome-wide association studies of gene expression and/or methylation in addition to disease phenotype. Many alternative causal inference methods, such as Mendelian randomisation, are not well-suited to large complex data sets, whereas Bayesian network analyses allow network space searches to identify perhaps previously overlooked causal relationships. The network with the best score, based on the likelihood of observing the data under that configuration, can be regarded as the best network. However, a directed connection between variables may only indicate correlation, rather than a true causal relationship. To address this issue, bootstrapping can be used to obtain an "average" network providing posterior probability estimates of the direction of causality between variables. We consider a newly proposed method to estimate these probabilities without the use of bootstrapping by making use of the network scores of all possible networks, although this method is only feasible for small networks. We utilise several small network models to simulate data and assess the performance of the different approaches using our own software package: BayesNetty. Bayesian networks provide a useful tool in understanding complex biological mechanisms and their further study and the development of user-friendly software implementations is important.

61 | Predicting Treatment Response in Rheumatoid Arthritis Using SNP Data

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*

There are several methods that allow genome-enabled prediction in situations where there are many more predictor variables than response variables. Here, we explore prediction of treatment response in Rheumatoid Arthritis patients from SNP data using penalised and non-penalised regression approaches. Many penalised approaches (such as the least absolute shrinkage and selection operator (lasso)) shrink the coefficients towards zero, thus performing variable selection, while non-penalised approaches (such as Genome-wide Complex Trait Analysis (GCTA)) fit the effects of all the SNPs as random effects by use of a linear mixed model.

The approaches are applied to data from the MAximising Therapeutic Utility in Rheumatoid Arthritis (MATURA) consortium. The data comprises treatment response measures and SNP data from approximately one thousand patients and five million genotyped or imputed SNPs, with none of the SNPs reaching genome-wide significance in a univariate regression analysis. We use 10-fold cross-validation to assess predictive performance, with nested 10-fold cross-validation used to tune the parameters when required. The results illustrate generally poor predictive ability as assessed by examining the correlation coefficient and the calibration slope. In order to investigate these results, we analyse several different real and simulated data sets. We find that penalised regression requires strong effects in order to achieve good prediction, while GCTA benefits from pre-filtering of SNPs.

62 | Genome-wide Meta-analysis of 24 Age-related Phenotypes Leveraging Longitudinal Follow-up and Genetic Heterogeneity Identified Abundant Associations in a Modest Sample of 26,371 Individuals

Alexander M. Kulminski[1], Yury Loika[1], Jian Huang[1], Jiayi Wang[1], Konstantin G. Arbeev[1], Olivia Bagley[1], Matt Duan[1], Anatoliy Yashin[1], Irina Culminskaya[1]

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

Genome-wide association studies (GWAS) are traditionally based on principles of medical genetics. This strategy is well adapted for Mendelian disorders. Genetics of phenotypes that leave human organisms vulnerable to diseases in late life (called age-related phenotypes) is, however, more complex. The fundamental complicating factor is the elusive role of evolution in fixing molecular mechanism of these phenotypes. This complexity implies a special type of an inherent genetic heterogeneity reflecting sensitivity of genetic associations with age-related phenotypes to the life course of individuals in different environments. Here we follow a two-stage genome-wide approach that leverages this heterogeneity and longitudinal follow-up. This approach is

demonstrated by examining genetic predisposition to 24 age-related phenotypes (16 biomarkers, 7 diseases, and death) in a modest sample (N = 26,371) from five studies (ARIC, FHS, MESA, CHS, and CARDIA) from the Candidate Gene Association Resource. In Stage 1, we performed the traditional univariate GWAS for each of 24 phenotypes enhanced by information from longitudinal follow-up in each study separately. In Stage 2, we used four meta-tests to combine statistics from Stage 1 across cohorts and phenotypes leveraging different types of genetic heterogeneity. Excluding proxy SNPs, individual-phenotype meta-analysis replicated 162 SNPs (81 loci) and identified 64 novel SNPs (59 loci) associated with different phenotypes at genome-wide level. Pleiotropic meta-analysis identified 171 novel SNPs (150 loci) in this modest sample (excluding the Major Histocompatibility Complex locus). Our findings demonstrate benefits of more comprehensive approaches than the currently prevailing ones to gain insights into the genetics of age-related phenotypes.

63 | Differences in Correlations of miRNA and Methylation with Target mRNA in Histologic Subtypes of Cervical Cancer

Prabhakar Chalise[1]

[1] *University of Kansas Medical Center, Kansas City, Kansas, United States of America*

Differences in molecular profiles and their interrelations between and within two major histologic subtypes, adenocarcinoma, and squamous cell carcinoma, have been found in many cancers. In order to examine the differences in regulatory impact of miRNAs and epigenetic effect of methylation on gene expression between the histologies, I propose a stepwise analysis approach. This analysis approach was applied to the cancer genome atlas (TCGA) studies on cervical cancer involving 31 cases of adenocarcinoma and 144 cases of squamous cell carcinoma. First, the DE analyses for miRNA, methylation and their target mRNA between the histologies were carried out followed by multiple testing adjustments using Benzamini and Hochberg's method. Second, the correlation analyses between the selected miRNA and target mRNAs, and methylation and mRNAs were carried out within each histologic subtypes separately. The statistically significant anti-correlations were compared further between the two histologies. Some anti-correlations were found statistically significant only in one type of histology but not in other. For example, a few miRNAs such as *miR-205*, *miR-200*, *miR-224* and *miR-193b* were found to have statistically significant anti-correlation with *ERBB3*, *JAG1*, *EYA4* and *ALDH1A2* respectively in squamous ($p = 2.03 \times 10^{-2}$, 1.34×10^{-3}, 1.08×10^{-4}, 1.00×10^{-6}) but not in adenocarcinoma ($p = 0.16, 0.20, 0.38, 0.16$). Similarly, the methylation probes near these genes were found significantly anti-correlated with corresponding genes in squamous histology only. Overall our study provides new insights into the differences in gene silencing mechanisms in the studies of cervical cancer.

64 | Height Associated Variants Demonstrate Assortative Mating in Human Populations

Xiaoyin Li[1], Susan Redline[2], Xiang Zhang[3], Scott Williams[1], Xiaofeng Zhu[1]

[1] *Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States of America;* [2] *Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, United States of America;* [3] *College of Information Sciences and Technology, Pennsylvania State University, University Park, Pennsylvania, United States of America*

Understanding human mating patterns, which can affect population genetic structure, is important for correctly modeling populations and performing genetic association studies. Prior studies of assortative mating in humans focused on trait similarity among spouses and relatives via phenotypic correlations. Limited research has quantified the genetic consequences of assortative mating. The degree to which the non-random mating influences genetic architecture remains unclear. Here, we studied genes that associate with human height to assess the degree of height-related assortative mating in European-American and African-American populations. We compared the inbreeding coefficient estimated using known height-associated variants with that using frequency matched sets of random variants. We observed a significantly higher inbreeding coefficient estimated from height-associated variants than from frequency matched random variants (p-value<0.05), demonstrating assortative mating in both populations at molecular level.

65 | Variant of the microRNA (*MIR6723*) is associated with Body Mass Index in a large sample of African Ancestry Populations

Guanjie Chen[1], Ayo Doumatey[1], Jie Zhou[1], Adebowale Adeyemo[1], Charles Rotimi[1]

[1] *CRGGH/NHGRI/NIH, Bethesda, Maryland, United States of America*

Obesity is a major risk factor for cardiac, pulmonary, metabolic, osteoarticular diseases and some forms of cancer. Despite considerable progress in the understanding of the pathogenesis of excess weight gain, the pathobiology of obesity at the individual and family levels remains largely unclear. Here, we investigated genomic contributions to obesity (measured by body mass index - BMI) in African ancestry populations.

A total of 8,599 individuals from five large cohorts of African Americans were included in this analysis. The cohorts are (Atherosclerosis Risk in Communities study (ARIC), Howard University Family Study (HUFS), Jackson Heart Study (JHS), and Multi-Ethnic Study of Atherosclerosis study (MESA), and The Cleveland Family Study (CFS). Imputation was performed on the GWAS genotype data available on these 8,599 individuals using the Sanger imputation server with a reference that has a total of 4,956 samples; this reference panel contains all of the African and non-African populations from 1000 Genomes phase 3, and ∼2000 whole-genome sequence data from Uganda (Baganda, Banyarwanda, Barundi and others) and ∼100 samples from each of a set of populations from Ethiopia (Gumuz, Wolayta, Amhara, Oromo, Somali), Egypt, Namibia (Nama/Khoesan) and South Africa (Zulu). The generalized linear mixed model association test (GMMAT) was used to perform association score test with cryptic relatedness random effect and included covariates of sex, age, type 2 diabetes status, and first two principal components.

The estimated heritability for obesity in these studies was high (70%). Variant rs567334121 (*MIR6723*, “A” allele, MAF = 0.062) was associated with BMI (p-value 1.18×10^{-13} (β = 1.46). This finding was replicated in 4,634 West Africans enrolled in the African American Diabetes Mellitus (AADM) study (“A” allele, MAF = 0.067, p-value = 0.046, $\beta = 0.43$). Meta-analysis displayed same direction of association with BMI (p-value 8.05×10^{-13} (Z score = 7.16)).

We identified a novel microRNA variant (rs567334121; MIR6723) that is associated with BMI. MicroRNAs are known to be involved in post-transcription regulation of gene expression by affecting the stability and translation of mRNA. The predicted targets for MIR6723 include *IL17A* and *TAOK3*. *IL17A* plays a vital role during acute inflammation, which is associated with obesity. It amplifies the inflammatory loop by inducing IL-6 secretion in adipocytes and neutrophils and affects the balance between inflammatory Th1 and regulatory T cell function with reciprocal inhibitory actions. In a genome-wide methylation analysis, *TAOK3* was founded to be associated with obesity, and a 1% increase in methylation of *TAOK3*decreases the odds of being obese by a factor of 0.91.

66 | A Novel Method to Detect Exon Usage Switches at Different Stages of Brain Development

Marie Forest[1,2], Alain Bateman[1,2,3], Anita A. Thambirajah[2,4], Celia M.T. Greenwood[1,2,3,5,6], Claudia L. Kleinman[1,2,3]

[1] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [2] Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Canada; [3] Department of Human Genetics, McGill University, Montreal, Canada; [4] Douglas Mental Health University Institute, McGill University, Montreal, Canada; [5] Department of Oncology, McGill University, Montreal, Canada, [6] Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada

Recent advances in sequencing technologies have facilitated the generation of comprehensive developmental transcriptomic and epigenomic datasets. In particular, the BrainSpan Atlas (https://www.brainspan.org/) is composed of more than 500 RNA-seq samples from sixteen human cortical and subcortical structures throughout development: from 8 weeks of gestation until 40 years of age. Here, we present a novel approach to identify variable isoform usage within genes in a development-dependent manner, based on functional data analysis (FDA) combined with clustering of exon-level trajectories across time.

To establish developmental trajectories, we aligned precursor fetal brain regions to their adult counterparts using anatomical and brain development information as necessary. For most genes, we estimated the gene expression trajectories of sixteen brain regions, using FDA for each exon in each brain region based on a smoothed combination of B-splines of order 5. We propose a clustering of these curves in three steps. (1) A distance between a pair of exons of a gene is defined by measuring the distance between the corresponding curves within each brain region and averaging these, followed by (2) hierarchical clustering of this distance matrix. (3) The optimal number of clusters is determined using the average silhouette and cophenetic measures of internal cluster validity. Finally, simulated data and random forest models build a predictor of whether the clusters of a gene are real.

This analytical approach may assist in the discovery of gene isoforms that are differentially expressed during critical transitions in the developing brain.

68 | Identifying and Utilising Genetic Variants Associated with Morning Plasma Cortisol: a CORtisol NETwork (CORNET) Analysis

Andrew A. Crawford[1,2], Brian R. Walker[1] on behalf of the CORNET Consortium

[1] BHF Centre for Cardiovascular Science, Queen’s Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom; [2] MRC Integrated Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

The latest genome-wide association meta-analysis conducted by the CORNET consortium investigated morning plasma cortisol in 25,314 individuals from 17 European cohort studies. Genetic variation in the SERPINA6/A1 locus on chromosome 14 reached a genome-wide level of significance (top SNP rs12589136, $p = 3.2 \times 10^{-19}$). Linkage disequilibrium score regression analyses estimated a SNP-heritability of 4% (0.038, se 0.018). The same method was used to estimate the genetic correlation between morning plasma cortisol and physical and mental health traits from GWAS consortia. Genetic correlations were identified between morning plasma cortisol and obesity, schizophrenia, total cholesterol, chronotype, body mass index, body fat and LDL cholesterol (absolute r_g effect sizes between 0.23 and 0.34, $p<0.05$). Pathway analyses were performed using MAGMA and suggested over-representation of genes involved in lipid metabolism and circadian pathways. Two sample Mendelian randomisation analyses were performed using three independent SNPs in the SERPINA6/A1 locus as an instrument for morning plasma cortisol and publicly available summary GWAS data. These analyses suggested that a 1SD higher morning plasma cortisol is causally associated with a 35% increase in type 2 diabetes (Inverse variant weighted method, lnOR 0.30, se 0.13, $p = 0.02$, DIAGRAM Consortium, cases = 26488; controls = 83964). They also suggest a 1SD higher BMI is causally associated with a 0.09SD reduction in morning plasma cortisol (beta -0.09, se 0.04, $p = 0.02$, GIANT 2015, $n = 339{,}224$). The evidence suggests that morning plasma cortisol shares genetic overlap with several cardiometabolic traits. We also provide evidence that elevated morning plasma cortisol is on the causal pathway to type 2 diabetes.

69 | Understanding the Causal Effects of Iron Metabolism on Chronic Disease Outcomes Using Mendelian Randomization

Anna Ramond[1], Daniel Freitag[2], James Staley[3], Adam Butterworth[1], Stephen Burgess[1,4], Emanuele Di Angelantonio[1]

[1]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Cardiovascular Data Sciences, Bayer Pharmaceuticals, Cologne Area, Germany; [3]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; [4]MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

Iron metabolism has been implicated in the pathogenesis of various common chronic diseases including cardiovascular disease, various forms of cancer and neurodegenerative diseases. While the observational evidence is indicative of an effect of iron metabolism in the pathology of several of these diseases, the potential causal effect remains unclear.

We have used an instrumental variable analysis approach, also known as Mendelian Randomization (MR), to investigate the causal effects of four iron biomarkers: serum iron, serum ferritin, transferrin and transferrin saturation (TS), on common chronic diseases, in particular, coronary heart disease (CHD), ischaemic stroke, colorectal cancer, lung cancer, Alzheimer's disease (AD) and Parkinson's disease (PD). We selected 7 single nucleotide polymorphisms associated with these four iron biomarkers at genome-wide significance level ($p<5.10^{-8}$) for use as instrumental variables. We used summary statistics from large genome-wide association studies of CHD, ischaemic stroke, colorectal cancer, lung cancer, PD, and AD, to investigate the effects of four iron biomarkers on these disease outcomes using MR.

Results of the MR analysis showed evidence of an increased risk of colorectal cancer with increased levels of ferritin and TS, whereas there was suggestion of a decreased risk of Parkinson's disease and lung cancer with increasing levels of serum iron and TS. There was no evidence of an association of CHD, ischaemic stroke or AD with any of the iron biomarkers.

Future work will focus on understanding the underpinning etiological mechanisms behind the effects of iron markers on these disease outcomes.

70 | Comparison of Whole-Genome Sequencing Data to Imputation Data for Cases with Venous Thromboembolism from the GENEVA Study

Brandon J. Coombes[1], Mariza de Andrade[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America

The advent of whole genome sequencing (WGS) allows a comprehensive view of the human genome. Although the cost of WGS continues to become cheaper, genotyping arrays combined with imputation to the entire genome are argued to be a more cost-effective strategy for studies. However, it is unclear how well current imputation strategies perform in a large scale study. Recently, WGS was obtained on a set of 1010 cases previously genotyped and analyzed as part of the Gene Environment Association Studies (GENEVA) genome-wide association study of venous thromboembolism (VTE). VTE cases were consecutive Mayo Clinic outpatients with objectively-diagnosed deep vein thrombosis and/or pulmonary embolism residing in the upper Midwest and are largely a European American sample (58% female). These cases were originally genotyped using the Illumina 660W-Quad BeadChip. After quality control filtering, 561,423 SNPs were observed from this array. Imputation analyses were initially performed using IMPUTE version 2 to impute around 38 million variants. The imputation analysis has since been updated using the recently developed minimac3 software on the Michigan Imputation Server. A comparison between WGS and the genotype array will be made to evaluate the agreement between the variants called on both platforms. We will then evaluate the imputation accuracy of IMPUTE2 and minimac3 compared to WGS across each chromosome. Particular focus will be placed on the imputation accuracy within genomic regions important to VTE such as Factor V. We expect imputation accuracy will be non-uniform across the genome. With this research, we hope to shed light on the debate between WGS and imputation.

72 | Preliminary Results from Genome-wide Meta-analysis of Survival Time in Idiopathic Pulmonary Fibrosis

Richard J. Allen[1], Justin Oldham[2], José M.L. Salazar[3], Shwu Fan Ma[4], Rebecca Braybrooke[5], UK ILD Consortium, Ian Sayers[6], Ian P. Hall[6], Martin D. Tobin[1,7], Imre Noth[4], R. Gisli Jenkins[6], Carlos Flores[3,8,9], Louise V. Wain[1,7]

[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2]Department of Internal Medicine, University of California Davis, Davis, California, United States of America; [3]Instituto Tecnológico y de Energías Renovables (ITER, S.A.), Santa Cruz de Tenerife, Spain; [4]Section of Pulmonary and Critical Care Medicine, University of Chicago, Chicago, Illinois, United States of America; [5]Division of Epidemiology and Public Health, University of Nottingham, Nottingham, United Kingdom; [6]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [7]National Institute for Health Research, Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, United Kingdom; [8]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [9]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

Idiopathic pulmonary fibrosis (IPF) is a rare lung disease of unknown cause, with few effective treatments available and poor prognosis (median survival time of 3 years). Genome-wide studies have identified variants associated with susceptibility to IPF. Although the effects of those variants on survival time have been investigated, no study has investigated survival time genome-wide for IPF. We set out to identify novel signals of association with IPF survival time and to replicate previous reports that variants in *MUC5B* that are associated with increased susceptibility to IPF, are paradoxically associated with an increase in survival time.

We performed genome-wide analyses investigating survival time in a UK IPF study and a USA IPF study and meta-analysed the results. Survival analyses were performed on variants with minor allele frequency (MAF)>0.5% in both studies using a Cox Proportional Hazards model. Independent variants meeting meta-analysis $p<5 \times 10^{-6}$ and $p<0.05$ in each of the separate studies (with consistent direction of effect estimate between studies), are being investigated further in an independent replication dataset.

A total of 963 individuals and 7,730,466 variants were included in the final meta-analysis with maximum follow-up of 16.5 years. A total of 79 independent signals (11 with MAF>5%) reached the criteria described above. Consistent with previous reports, the allele in rs35705950 in *MUC5B* that is associated with increased susceptibility to IPF shows some association with increased survival time (HR = 0.73, 95% CI: [0.61, 0.87], $p = 5.08 \times 10^{-4}$).

This details the first analysis to investigate survival time genome-wide in IPF.

73 | Incorporating Interaction Effects in Trait Prediction Using MB-MDR

Damian Gola[1], Inke R. König[1]

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

Common complex traits are based on complex molecular mechanisms. These mechanisms involve multiple, possibly interacting genetic features and may also depend on interactions with environmental features. Importantly, interacting features may have only negligible main effects. However, most algorithms used in prediction such as penalized regression or random forests tend to favour features with clear main effects and thus do not exploit information from interaction effects. In the context of common complex diseases, this can, therefore, lead to lower prediction performance. To improve this, we propose a novel prediction algorithm that is based on the *Model-Based Multifactor Dimensionality Reduction* (MB-MDR) method. MB-MDR is a nonparametric method to identify interacting features in whole genome datasets. It ranks all possible feature combinations based on an association test statistic with the trait. In our algorithm, the k highest ranked feature combinations are used for trait prediction of new observations. Specifically, the estimated trait values conditioning on the respective genotype combination in each of the k feature combinations are aggregated, and the optimal value for k is determined by internal cross-validation. The proposed algorithm is compared with state of the art prediction algorithms in an extensive simulation study regarding prediction performance and runtime. The performance in real data is illustrated by application of the proposed algorithm to a coronary artery disease case/control dataset.

74 | The *COL5A3* and *MMP9* Genes Interact in Eczema Susceptibility

Patricia Margaritte-Jeannin[1,2], Marie-Claude Babron[1,2], Catherine Laprise[3], Chloé Sarnowski[1,2], Myriam Brossard[1,2], Miriam Moffatt[4], William O. Cookson[4], Emmanuelle Bouzigon[1,2], Florence Demenais[1,2], Marie-Hélène Dizier[1,2]

[1]Inserm, UMR-946, Genetic Variation and Human Diseases unit, F-75010, Paris, France; [2]Univ Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, F-75010, Paris, France; [3]Université du Québec, Chicoutimi, Canada; [4]National Heart Lung Institute, Imperial College, London, United Kingdom.

Genetic studies of eczema have identified many genes, which explain only 14% of the heritability of this trait. Missing heritability may be partly due to ignored Gene–Gene (G-G) interactions. Our aim was to detect new interacting genes involved in eczema.

The search for G-G interaction in eczema was conducted using a two-step approach, which included a first step of gene selection based on biological knowledge related to eczema, and a second-step of interaction analysis of the selected genes. Analyses were carried out at both SNP and gene levels in three asthma-ascertained family samples: the discovery dataset of 388 French EGEA (Epidemiological study on the Genetics and Environment of Asthma) families and the two replication datasets of 253 French-Canadian SLSJ (Saguenay-Lac-Saint-Jean) families and 207 UK MRCA (Medical Research Council) families.

One pair of SNPs, rs2287807 in *COL5A3* and rs17576 in *MMP9*, that was detected in EGEA at $P_{\text{int-SNP}} \leq 10^{-5}$ showed significant interaction after meta-analysis of interactive effects estimated in EGEA, SLSJ and MRCA ($P_{\text{int-SNP}} = 10^{-8}$ under the significant threshold of 10^{-7}). Gene-based analysis confirmed strong interaction between *COL5A3* and *MMP9* ($P_{int\text{-}gene} = 4 \times 10^{-8}$ under the significant threshold of 4×10^{-6}) by meta-analysis of the three datasets. When stratifying the data on asthma, this interaction remained in both groups of asthmatic and non-asthmatic subjects.

This study identified two new genes, *COL5A3* and *MMP9*, interacting on eczema susceptibility, independently from asthma. Further confirmation of this interaction as well as functional studies are needed to better understand the role of these genes in eczema.

75 | Type 2 Diabetes Mellitus and Pancreatic Cancer Risk. An Independent Etiological Relation?

Marta Rava[1], Esther Molina-Montes[1], Rianne Boenink[1,2], Paulina Gómez-Rubio[1], Núria Malats[1], on behalf of the PanGenEU, ISBlaC and EPICURO Studies Investigators.

[1] Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [2] Radboud University, Nijmegen, The Netherlands

Type 2 diabetes mellitus (T2DM) has been consistently associated with increased risk of pancreatic ductal adenocarcinoma (PDAC). Whether it is a causal association is still under debate. The aim of this study was to explore the causal link behind T2DM and PDAC through Mendelian randomization analysis (MRA).

1146 PDAC cases and 444 controls from the European PanGenEU study had epidemiological and genotype data. Control group was enriched with 1043 hospital-based controls from two Spanish bladder cancer studies (ISBlaC and EPICURO). Information about T2DM, smoking status and body mass index (BMI) was available. A GWAS-catalogue database review was performed to identify SNPs associated with T2DM. Those SNPs were tested for association with PDAC and T2DM in controls and used as genetic instrumental variable (IV) to test for causal association with MRA.

T2DM was associated with PDAC with an OR of 2.23 (95%CI 1.79, 2.79). There were two T2DM-related SNPs significantly associated with PDAC. However, these SNPs had pleiotropic effects and could not be used as IVs. Thirty-five independent SNPs, neither associated with smoking nor BMI, were considered for MRA. Estimates for causal associations varied upon SNPs selection procedures that were used to rule out pleiotropic effects. Estimates were also sensitive to covariate adjustment, especially regarding BMI, which may act as confounder or mediating factor in the association with PDAC. SNP-diabetes effect estimates from external sources will be further considered. These preliminary findings highlight the importance of a careful SNP and covariates selection for MRA.

These results underline the importance of suitable surrogates for ancestry estimation which reflect the actual composition and genetic heterogeneity of study individuals.

76 | Deciphering the Common Genetic Susceptibility to Pancreatic Cancer

Evangelina López de Maturana[1,2], Lola Alonso[1,2], Isabel Martín-Antoniano[1,3], Francisco X. Real[4,5], Núria Malats[1,2], on behalf of the PanGenEU, ISBlaC and EPICURO Studies Investigators

[1] Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [2] Centro de Investigación Biomédica en red Cáncer (CIBERONC), Madrid, Spain; [3] Instituto de Medicina Molecular Aplicada (IMMA), Facultad de Medicina, Universidad San Pablo CEU Madrid, Spain; [4] Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain; [5] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain

Pancreatic cancer (PC) frequency is increasing in Westernized countries. PC is a complex disease in which the genetic factors play a role. Knowledge of PC genetic basis is still incomplete as limited number of genome-wide association studies (GWAS) have been conducted.

Here, we present the first results of a GWAS conducted in European population to identify novel hits associated to PDAC risk. We used the resources of 1317 cases and 1616 controls from the PanGenEU, ISBlaC and SBC/EPICURO studies.

We prioritized suggestive associations ($p<1 \times 10^{-4}$) for further bioinformatics analysis, including evidence of functional impact, annotation of tagged genes and pathways, eQTLs in pancreas tissue, and regulatory chromatin marks. Most analyses were performed using DoriTool, a novel integrative pipeline developed in our lab. Moreover, we used data from assays conducted in pancreas tissue cataloged in the ENCODE project.

We identified 139 novel variants associated with PDAC. Most of them were in chromosomes 1, 6, 8 and 10. In particular, 8q24.21, a region previously associated with many cancers, harbored the largest number of the identified variants, followed by the 10q11.21 region. Variants in 7q34, 12q24.33 and 17q24.2 were potentially functional since either they were eQTLs in pancreatic tissue, they overlapped with regulatory elements, or they tagged genes annotated in pathways associated with pancreatic cancer.

Our work adds important new knowledge to understand the genetics underlying PDAC through a post-GWAS functional *in-silico* analysis, by integrating additional –omics data that may reveal potential causal regions.

Funding: ISCIII-FIS #PI1501573.

77 | A Simulation Study of Winner's Curse Bias and Bootstrap Bias Reduction in Genome-wide Analysis of Low-frequency Variants

Ruiyang Yi[1], Ji-Hyung Shin[1], Shelley B. Bull[1,2]

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

The Winner's Curse is a phenomenon of upward bias in the magnitude of effect estimates in genome-wide studies. Because this bias depends on underlying power, low-frequency variants are more prone to suffer from the winner's curse. Inference in binary trait studies can be further complicated by violation of large-sample assumptions and failure of standard logistic regression. Bootstrap resampling methods have been used to correct for winner's curse bias, but most methods evaluations have focused on common variants. Our simulation study design was based on the Wellcome Trust Case-Control Consortium Type 1 Diabetes dataset, with low-frequency variants divided into four sub-groups by minor allele frequency (MAF) (0.01-0.02, 0.02-0.03, 0.03-0.04, 0.04-0.05 consisting of 9204, 9397, 9422 and 8947 variants, respectively). We generated disease outcome under a binary trait design with disease prevalence of 40% and sample size of 4901. Disease risk depended on four bi-allelic variants, one from each MAF sub-group. We applied bootstrap resampling to standard and Firth penalized logistic regressions. For all four causal SNPs, we observe inflated naïve log OR estimates for both maximum likelihood (ML) and penalized maximum likelihood (PML) as expected. On average the ML and PML naïve estimates are similar, although the PML LR test P values are smaller. Differences between ML and PML become more dramatic with lower disease prevalence. Bootstrap resampling reduces both ML and PML estimates for true positives but tends to be more conservative for the latter. Overall, bootstrap shrinkage for low-frequency SNPs is greater for lower compared to higher MAF.

78 | A Systematic Review of Genetic Syndromes with Diabetes

Daniel Shi[1], Dalton Budhram[1], Yuvreet Kaur[1], David Meyre[1,2]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada; [2]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada

Syndromes with diabetes typically follow Mendelian patterns of inheritance and involve the co-presentation of clinical features, such as developmental delay, epilepsy, renal cysts, ataxia, deafness, or obesity. Previous reviews have identified particular syndromes associated with diabetes, but no systematic review has been conducted to date. We recently completed the first systematic review on obesity syndromes, and apply here the same strategy to provide an exhaustive catalog of syndromes associated with diabetes. Five databases (Embase, Medline, Ovid, OMIM, Orphanet) were searched using terms including "diabetes", "syndrome", and "genetic" to retrieve relevant literature on syndromic diabetes. Our literature search identified 2,916 papers. Titles and abstracts of all retrieved articles were screened independently by two reviewers (DS and DB) to identify articles for full-text review. A training session for the two reviewers was provided by a subject matter expert (DM) on 100 articles. After preliminary abstract review of 600 papers, 126 were relevant. The inter-rater agreement was very good (kappa = 0.837). Our initial analysis of these papers found 52 distinct diabetes syndromes. Further analysis will focus on the state of genetic elucidation of these syndromes. Organizational inconsistencies in the reporting of diabetes syndromes and the quality of evidence will also be assessed and used to provide recommendations for improvement. Our systematic review will provide the first comprehensive catalog for syndromes with diabetes and the state of their genetic elucidation. The description of these syndromes is expected to stimulate gene identification efforts and improve prediction and treatment of syndromic and more common forms of diabetes.

79 | Systematic Review and Meta-Analysis of the Association between *FTO* rs9939609 Obesity Polymorphism and Suicide in a Large-Scale Multiethnic Population

Dalton Budhram[1], Daniel Shi[1], Hudson Reddon[1], Zainab Samaan[1,2], David Meyre[1,3]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada; [2]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada; [3]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada

An inverse association between the obesity risk rs9939609 A variant in the *FTO* gene and completed suicide has been recently suggested in a modestly powered case-control study from Poland. The aim of this study is to investigate the association of *FTO* rs9939609 polymorphism with suicide ideation, attempted and completed suicide using an innovative systematic review / meta-analytic approach in multiethnic populations. It includes a systematic review of the literature in Medline, Embase, Web of Science, PsycINFO,

an in silico extraction from GWAS literature based on a collaborative model, complemented by data extraction from pre-existing GWAS or custom-arrays in consortia and single studies, and de novo genotyping of *FTO* in the McMaster DISCOVER suicide case-control study. We employ a global meta-analytic random-effects model to calculate summary beta estimates and SEs or ORs and 95% CIs for suicide ideation score, attempted and completed suicide status analyses, respectively. Our initial search identified only one association study between *FTO* SNP and suicide in literature. A second strategy retrieved GWAS literature examining suicide ideation and suicide statuses (n = 230 articles), and 30 studies were included for collaboration invitations. The inter-rater agreement was good (kappa = 0.654). In addition, we have extracted genetic and clinical information for about 17,450 participants from 7 dbGAP studies. We expect to complete this meta-analysis by the end of the 2017 summer. Overall, this large-scale meta-analysis will help to identify the shared genetic bases for obesity and suicide predispositions and will contribute to innovative prevention strategies for these disorders.

80 | Replication of Tuberculosis Susceptibility SNPs Found by GWAS Suggested the Need of Better Alternative Phenotypes

Nelson Tang[1], Amy Wang[1], C.C. Leung[2], K.C. Chang[2], Mamie Hui[3], C.Y. Chan[3]

[1]Department of Chemical Pathology, The Chinese University of Hong Kong, Shatin N.T., Hong Kong SAR; [2]Department of Health, Hong Kong SAR; [3]Department of Microbiology, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong SAR

Genetic susceptibility plays an important role in tuberculosis (TB) infection. Like many other diseases, several GWAS have been reported. However, replication of TB predisposition SNPs have been mostly unsuccessful. SNP rs4331426 in chromosome 18 was first reported in an African TB cohort. However, its effect was either not replicated or reversed in replication studies in other population. Here, we explored other tagSNPs in this loci in order to fully understand the landscape of genetic association with TB in this chromosome 18 locus.

1150 Chinese TB patient and 1280 population controls were genotyped for tagSNPs in this loci. In the whole group analysis, no SNP showed significant association. We then performed a subgroup analysis for young TB patients defined as between 20 to 40 years-of-age (389 patients vs 1280 controls). Rs1013483 was significant ($p = 0.05$). OR:1.28 (CI = 1 to 1.6). Another snp rs8091189 also showed a borderline p-value of 0.07.

The results suggested that there was a huge genetic heterogeneity in TB susceptibility across ethnic groups as these SNPs were not detected in the African GWAS cohort. Furthermore, early onset disease (young TB patients) may have a strong genetic predisposition and would be a better phenotype for genetic study.

81 | Improved Phasing and Imputation for Large-Scale Data

Brian L. Browning[1,2], Sharon R. Browning[2], Xiaowen Tian[2]

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

We present new methods for genotype phasing and genotype imputation that have improved accuracy and compute time for large-scale data. Both methods perform hidden Markov model calculations on haplotypes instead of diplotypes and utilize a continuously-evolving, rolling window of reference haplotypes.

We compared our new genotype phasing method to Eagle2 (v2.3.2), the current state of the art, on UK Biobank data (n = 150,000) for chromosomes 1, 5, 10, 15, and 20. Using current default settings, our phasing method reduced per-chromosome switch error rates and compute times by 26–37% and 8–27% respectively relative to Eagle2.

In order to assess the performance of our genotype imputation method with very large reference panels, we simulated 3 million UK European samples, exploiting recent advances in demographic modeling and simulation methods. The simulated data incorporate recurrent mutation and have a near-saturated marker density of 1 non-singleton single-nucleotide variant per 5 bases. Relative to Beagle 4.1, the current state of the art for this size reference panel, our new imputation method reduces compute time by an order of magnitude with no loss in accuracy when imputing 1000 target samples from 3M reference samples in binary reference (bref) format. Extrapolating to a 3000 Mb genome, we estimate that the compute time for genome-wide imputation from 3M reference samples is < 12 min per sample on a 12 CPU-core server.

Our new genotype phasing and imputation methods will be incorporated in Beagle version 5 (https://faculty.washington.edu/browning/beagle/beagle.html).

82 | Rare variants and parent-of-origin effects on whole blood gene expression assessed in large family pedigrees

Andrew A. Brown[1], Ana Viñuela[1], Angel Martinez-Perez[2], Andrey Ziyatdinov[2], Maria Sabater-Lleal[3], Anders Hamsten[3], Juan C. Souto[4], Alfonso Buil[1], Jose M. Soria[2], Emmanouil T. Dermitzakis[1]

[1]University of Geneva Medical School, Genetic Medicine & Development, Geneva, Switzerland; [2]Unit of Genomics of Complex Diseases, Biomedical

Research Institut Sant Pau (IIB-Sant Pau), Barcelona, Spain; [3]Karolinska Institute, Medicine, Stockholm, Sweden; [4]Sant Pau Hospital, Hematology, Barcelona

Studying genetic effects on gene expression in related individuals provides insights inaccessible when using unrelated individuals, such as heritability estimates, rare (in population) regulatory variants commonly observed in the pedigree, imprinting, and parent-of-origin effects. We report the GAIT2 study of 935 individuals from 35 pedigrees, with whole blood RNA-seq, blood cell count and extensive phenotypic information available.

We identified 11,297 eQTLs (expression Quantitative Trait Loci, FDR<0.05) using a variance components based cis association mapping. To see if these eQTLs are rare in the general population, we examined the minor allele frequency (MAF) of the lead variants in 1000 Genomes populations. Compared to eQTLs from the DGN study with unrelated individuals, we see an excess of variants with MAF<0.01 (9.6% eQTL vs 0%, median MAF is 0.11 vs 0.27). Finally we looked for parent-of-origin effects on expression, meaning the effect of a variant depends on whether it was maternally or paternally inherited (parent-of-origin in expression QTL, poeQTL). We found 12 significant poeQTLs (FDR<0.05). Six affect known imprinted genes, implying a cis-eQTL with effect masked on the imprinted haplotype. However, for four remaining genes, measures of allelic ratios from either our study or the GTEx data showed a combination of mono-allelic and bi-allelic expression, suggesting these genes may be partially imprinted. Both rare variants and parent-of-origin genetic effects have been shown to be relevant for human disease. Studies such as this allow a deeper understanding of their properties, showing that exclusively studying imprinted regions will miss many parent of origin effects.

83 | A Powerful Framework for Integrating eQTL and GWAS Summary Data

Zhiyuan Xu[1], Chong Wu[1], Peng Wei[2], Wei Pan[1]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Biostatistics, University of Texas MD Anderson Cancer Research Center, Houston, Texas, United States of America

Two new gene-based association analysis methods, called PrediXcan and TWAS for GWAS individual-level and summary data respectively, were recently proposed to integrate GWAS with (expression Quantitative Loci) eQTL data, alleviating two common problems in GWAS by boosting statistical power and facilitating biological interpretation of GWAS discoveries. Based on a novel reformulation of PrediXcan and TWAS, we propose a more powerful gene-based association test to integrate single set or multiple sets of eQTL data with GWAS individual-level data or summary statistics. The proposed test was applied to several GWAS datasets, including two lipid summary association datasets based on ∼100,000 and ∼189,000 samples respectively, and uncovered more known or novel trait-associated genes, showcasing much-improved performance of our proposed method. The software implementing the proposed method is freely available as an R package.

84 | Imaging-Wide Association Study: Integrating Imaging Endophenotypes in GWAS

Zhiyuan Xu[1], Chong Wu[1], Wei Pan[1], for the Alzheimer's Disease Neuroimaging Initiative

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

A new and powerful approach, called imaging-wide association study (IWAS), is proposed to integrate imaging endophenotypes with GWAS to boost statistical power and enhance biological interpretation for GWAS discoveries. IWAS extends the promising transcriptome-wide association study (TWAS) from using gene expression endophenotypes to using imaging and other endophenotypes with a much wider range of possible applications. As illustration, we use gray-matter volumes of several brain regions of interest (ROIs) drawn from the ADNI-1 structural MRI data as imaging endophenotypes, which are then applied to the individual-level GWAS data of ADNI-GO/2 and a large meta-analyzed GWAS summary statistics dataset (based on about 74000 individuals), uncovering some novel genes significantly associated with Alzheimer's disease (AD). We also compare the performance of IWAS with TWAS, showing much larger numbers of significant AD-associated genes discovered by IWAS, presumably due to the stronger link between brain atrophy and AD than that between gene expression of normal individuals and the risk for AD. The proposed IWAS is general and can be applied to other imaging endophenotypes, and GWAS individual-level or summary association data.

85 | Genetics of the Measure of Physiological Dysregulation: Insights from Longitudinal Data

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

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

Recently a novel implementation of the statistical (Mahalanobis) distance measure (D_M) was 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. The D_M

allows reducing high-dimensional biomarker space into a single estimate which summarizes information about various biomarker trajectories. We showed in applications to Framingham Heart Study (FHS) data that this measure is associated with increased risks of death, onset of chronic diseases and worse survival following onset of diseases as well as with other aging-related characteristics which can be indirectly evaluated from longitudinal trajectories of biomarkers. We performed genome-wide association studies of "static" (value at age at biospecimen collection) and "dynamic" (slopes at different age intervals) characteristics of $\mathrm{D_M}$ using data from the FHS Candidate Gene Association Resource. We found the strongest associations for slopes at ages 40–60 (top two SNPs: rs10789863, $p = 8.0\mathrm{E}{-7}$; rs1363730, $p = 7.9\mathrm{E}{-6}$) whereas no variants reached a significance level of 1E−5 in the analyses of the "static" $\mathrm{D_M}$ characteristics. We also performed high-resolution analyses of effects of polygenic risk scores (constructed from SNPs passing different *p*-value thresholds) on $\mathrm{D_M}$ and lifespan. The results of this study indicate that the dynamics of $\mathrm{D_M}$ can have a genetic component and that this dynamics is associated with different aging-related outcomes. Additional analyses in longitudinal studies with larger genotyping platforms are needed to confirm these findings.

86 | Genome-wide Trans-Ancestral Meta-Analysis Provides New Insights into Genetic Architecture of Gout

Wen-Hua Wei[1], Tony Merriman[2], on behalf of Asia-Pacific Gout Consortium

[1]Department of Women's and Children's Health, University of Otago, Dunedin, New Zealand; [2]Department of Biochemistry, University of Otago, Dunedin 9016, New Zealand

Gout is a complex inflammatory arthritis affecting a quarter of people with elevated serum urate levels. The prevalence of gout varies widely across ethnic groups: 6–8% of Māori and Pacific (Polynesian) people living in New Zealand develop gout, compared to 3% of Europeans. In the Asia-Pacific region 1 ∼ 2% of Asians and 0.4% of South Americans develop gout. Previous genome-wide association studies (GWASs) of gout have yielded relatively limited discoveries including major loci urate-regulating loci *ABCG2* and *SLC2A9*, and a few novel associations. A trans-ancestral meta-analysis of gout that covers most ancestral groups will be ideal to better dissect the genetic architecture of gout. We, therefore, developed the Asia-Pacific Gout Consortium (APGC) to facilitate such a project. The first stage of the study will be based on published data sets including 5,700 European, 1,255 Han Chinese (Li *et al. Nat Commun* 6:7041, 2015), 945 Japanese (Matsuo *et al. Ann Rheum Dis* 75:652–659, 2016) and 1,200 Polynesian individuals with gout. The second stage will use larger unpublished sample sets of European, Chinese and Japanese ancestry. We will follow the trans-ethnic meta-analysis protocol used by Liu *et al.* (*AJHG* 99:56-75, 2016): (1) fixed-effect meta-analysis within ethnic group/cluster using software GWAMA; (2) trans-ethnic meta-analysis of summary statistics from (1) using software MANTRA; (3) fine mapping signals identified at (2) using MANTRA. Here we report the results of the first stage.

87 | Polymorphisms in MicroRNA Processing Machinery Genes were Associated with Liver Cancer Risk in a Chinese Population

Xing Liu[1,2], Shen-Chih Chang[1], Binh Y. Goldstein[1], Lina Mu[3], Lin Cai[4], Na He[2], Bao-Guo Ding[5], Jin-Kou Zhao[6], Shun-Zhang Yu[2], Qing-Yi Lu[7], Zuo-Feng Zhang[1]

[1]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, California, United States of America; [2]Department of Epidemiology, Fudan University School of Public Health, Shanghai, China; [3]Department of Social and Preventive Medicine, State University of New York at Buffalo, Buffalo, New York, United States of America; [4]Department of Epidemiology, Fujian Medical University, Fujian, China; [5]Taixing City Center for Disease Control and Prevention, Jiangsu, China; [6]Jiangsu Province Center for Disease Control and Prevention, Jiangsu, China; [7]Center for Human Nutrition, School of Medicine, University of California, Los Angeles, California, United States of America

SNPs in microRNA-related genes may affect cancer risk and prognosis, and microRNA machinery genes are responsible for precise microRNA generations. We examined seven SNPs from microRNA machinery genes including *XPO5*, *RAN*, *DICER1*, *AGO2*, *GEMIN3* and *GEMIN4* using the data from a population-based case-control study with 204 cases and 415 controls in Taixing, China to explore the potential genetic susceptibility for liver cancer. Epidemiologic data on environmental risk factors including alcohol drinking and tobacco smoking were collected using questionnaire, and genotyping was performed using a customized Fluidigm Dynamic 96.96 Array$^{\mathrm{TM}}$ Assay (Fluidigm, South San Francisco, CA). Rs11077 (*XPO5* gene) C/C type showed a positive association with liver cancer in co-dominant model (cOR = 5.26, 95% CI: 1.01-27.39) compared to A/A type. After controlling for age, gender, level of education, family income per capita 10 years ago, BMI, HBsAg status, HCV-Ab status, family history of liver cancer, pack-year of smoking, daily alcohol intake and plasma aflatoxin-albumin adduct levels, *XPO5* rs11077 A/C and C/C type showed an aOR of 2.15 (95% CI: 1.19-3.90) for liver cancer in additive model, aOR of 2.28 (95% CI: 1.13-4.61) in dominant model compared to A/A type. These results may offer epidemiological evidence for the association between SNPs from miRNA machinery genes and liver cancer risk.

88 | Extension of a Phenotype Imputation Approach in Genome-wide Association Studies

Yuning Chen[1], Gina M. Peloso[1], Josée Dupuis[1]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America

Statistical power is a limitation of genome-wide association studies (GWAS). Sample size is a major component of statistical power that can be easily affected by missingness in phenotypic data and restrain the ability to detect associated SNPs with small effect sizes. While some phenotypes are hard to collect due to cost and loss of follow-up, correlated phenotypes that are easily collected and complete can be leveraged. PhenIMP is a phenotype imputation method incorporating family structure and correlation between multiple phenotypes. We investigate the performance of PhenIMP under several conditions, derive the exact non-centrality parameter (NCP) of the test statistic for association and propose a new approach to analyze the imputed and observed phenotype values in GWAS. Our simulation results show that inflated type-I error can occur under some conditions. We verify our exact NCP derivation by comparing the theoretical power with the simulated power. We also illustrate that our method of analyzing the imputed and observed phenotype values has higher power than the method proposed in PhenIMP in most practical scenarios. We then extend the method to gene-based test. Based on the NCP in GWAS, an analytical approach to compute power for gene-based test is proposed. Finally, we apply the method to high-density lipoprotein cholesterol, low-density lipoprotein cholesterol and intermediate density lipoprotein cholesterol in the Framingham Heart Study.

89 | *HOXB13* G84E Mutation and Prostate Cancer Risk: Kin-Cohort Analysis Using Data From the UK Genetic Prostate Cancer Study

Tommy Nyberg[1], Tokhir Dadaev[2], Koveela Govindasami[2], Andrew Lee[1], Malgorzata Leslie[1], Zsofia Kote-Jarai[2], Rosalind Eeles[2], Antonis C. Antoniou[1]

[1]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Oncogenetics Team, Division of Cancer Genetics and Epidemiology, The Institute of Cancer Research, London, United Kingdom

G84E missense mutations in *HOXB13* are associated with prostate cancer. However, a wide range of risk estimates has been reported. Based on case-control studies, reported OR range from 2 to 20, often with wide confidence intervals because mutations are rare in the population. To obtain more precise risk estimates, we used a kin-cohort study design and modified segregation analysis, using family data on 11,988 PCa index-cases (4509 consecutive cases, 870 and 6609 cases recruited based on family history and young age at diagnosis, respectively) enrolled in the UK Genetic Prostate Cancer Study, who had been genotyped for G84E. Among index-cases, 182 carried at least one copy of G84E. PCa incidence was assumed to follow a mixed Cox regression model of the form $\lambda\ (t) = \lambda_0\ (t) \times \exp(G + P)$, where G is a fixed effect which depends on G84E, $P \in N(0, \sigma_P^2)$ a residual polygenic random effect, and $\lambda_0(t)$ is the baseline incidence for non-carriers to age t. Using maximum likelihood, after adjusting for ascertainment, we estimated the frequency and RR (i.e. penetrance) for G84E under different genetic models, and σ_P. Preliminary results suggest that under the best fitting model, the data are consistent with a multiplicative model where each copy of G84E confers RR for PCa of 2.6 (95%CI 1.7-4.2), and a significant σ_P of 1.8 (95%CI 1.7-1.9), indicating that family history increases risk above that resulting from being a mutation carrier. Ongoing work will evaluate effect-modification of RR and/or σ_P by age, birth cohort, and mutation status, and estimate absolute risks for reference family structures.

90 | The Michigan Genomics Initiative: A Model Framework for Genetic Discovery Using Patient Electronic Health Records

Ellen M. Schmidt[1], Lars G. Fritsche[1,2], Seunggeun Lee[1], Peter VandeHaar[1], Chad M. Brummett[3], Sachin Kheterpal[3], Gonçalo R. Abecasis[1]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, United States of America; [2]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, NTNU, Norwegian University of Science and Technology, Trondheim, Norway; [3]Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

Our understanding of the genetics underlying complex phenotypes has made remarkable strides, with thousands of trait-associated variants cataloged from genome-wide association studies (GWAS). Still, our bandwidth to measure diverse phenotypes in large, high-powered sample sizes is limited. Electronic health records (EHR) collected during hospital and clinic encounters are a valuable resource of precisely defined outcomes. In addition, the surgical procedural period provides a unique opportunity to collect patient biospecimens and enriched health information. Here we interrogate 7.7M common variants and 1,448 billing code-defined disease states to establish relationships between the genome and a diverse set of EHR-based outcomes.

Patients undergoing surgery at the University of Michigan (UM) Hospital are invited to participate in the Michigan Genomics Initiative (MGI), which currently has over 42,000 recruits. Genotypes are measured from patient blood samples on a customized HumanCoreExome array and imputed using the Haplotype Reference Consortium panel. Phenotypes inferred from ICD (International Classification of Disease) codes are translated into broader 'PheWAS' (phenome-wide association study) groups, revealing an enrichment of cases with neoplasms including skin cancer as well as cardio-metabolic traits such as hypertension.

We replicate several well-known genetic associations including Factor 5 and thrombosis (p-value = 4×10^{-39}), *TCF7L2* and type 2 diabetes (p-value = 2×10^{-11}), and *PITX2* and atrial fibrillation (p-value = 3×10^{-9}). In addition, we examine the shared genetic effects among related and apparently unrelated traits, providing insight into relationships between phenotypes not previously studied as well as novel genetic associations. Results are presented in dynamic GWAS-to-PheWAS landscapes at https://pheweb.sph.umich.edu.

91 | The Stressed Pancreas Determines the Islet Cells to Codify the Glucagon-L Peptide 1 Receptor with Bariatric Surgery used as Functional Stressful Factor

J. Arturo Prada-Oliveira[1], David Almorza-Gomar[2], Joshua Falckenheiner-Soria[3], Alejandra Moreno- Arciniegas[4], Alonso Camacho-Ramírez[4], Gonzalo M. Pérez-Arana[1]

[1] Department of Human Anatomy and Embryology, Faculty of Medicine, Cádiz, Spain; [2] Department of Statistics and Operational Research, Cádiz, Spain; [3] Virgen de las Montañas Hospital, Villamartin, Cádiz, Spain; [4] Department of Surgery, Puerto Real Universitary Hospital, University of Cádiz, Cádiz, Spain

Bariatric surgery is broadly employed in obese patients, with excellent results over the metabolic syndrome. One special goal is the improvement of Diabetes mellitus type 2 (T2DM), which is usually associated with obesity. Many reports focused on the influence of several enterohormones on the functional pathologic basis of T2DM. These are the increase of peripheric resistance to insulin and the final falling of pancreatic β-cell. The GLP-1 incretin, secreted in the ileum, has been related as a main effector to the processes of pancreatic upgrading. We studied the changes in the GLP-1 secreting cells in the ileum and the changes in the expression of the receptor to GLP-1 in the islet pancreatic cells. This study undergoes surgeries employed in human clinic.

Forty euglycemic Wistar rats were randomised and included in the surgical groups. The surgical groups used two of the most broadly used techniques: 1) Sleeve Gastrectomy, Roux–en Y Gastric Bypass, and 2) Intestinal Resection of 50% of jejunum. We included two surgical controls (fasting and surgical).

The results showed a significant increase in the number of ileum L-cell -expressed as GLP-1+ cells/mm^2 ileum- in the RI50 and RYGB groups versus the controls. We reported a significant increase in the expression of GLP-1 receptors in the endocrine pancreas –expressed as GLP-1R+ cells/mm^2 islet-.

We concluded that the conformational changes in the digestive tube provoke a serious stress factor in the endocrine pancreas, which activated the codification of GLP-1 as an excitatory response to the increased secretion of ileum GLP-1.

92 | Comparison of Haplotype-Based Tests for Detecting Gene-Environment Interactions with Rare Variants

Charalampos Papachristou[1], Swati Biswas[2]

[1] Department of Mathematics, Rowan University, Glassboro, New Jersey, United States of America; [2] Department of Mathematical Sciences, University of Texas at Dallas, Texas, United States of America

Detecting gene-environment interactions (G × E) is critical in unraveling the etiology of complex diseases. This is a challenging problem especially if the genetic variants under study are rare. Haplotype-based tests have several advantages over the so-called collapsing tests as highlighted in recent literature. Thus, it is important and timely to compare haplotype-based tests for detecting G × E including the recent ones developed specifically for rare haplotypes. We consider five methods – Haplo.glm, Hapassoc, HapReg, Bayesian hierarchical Generalized Linear Model (BhGLM), and Logistic Bayesian LASSO (LBL). HapReg has two versions and LBL has three versions depending on whether gene-environment (G-E) independence assumption is made or not; we consider all versions in our comparison. We carry out extensive simulations with data generated under different association scenarios and varying levels of G-E dependence. We find that at nominal levels BhGLM is extremely conservative and HapReg is liberal. When the type I error rates are controlled to be similar for all methods, LBL is most powerful followed closely by BhGLM. However, LBL is the most computationally intensive method of the ones considered here. We also applied the methods to a real dataset on lung cancer, in particular, in region 15q25.1 as it has been suggested in the literature that it interacts with smoking to affect the lung cancer susceptibility as well as is associated with smoking behavior. Only LBL was able to detect a rare haplotype-smoking interaction in this region.

94 | Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets

Silke Szymczak[1]

[1] Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany

Machine learning methods such as random forests are promising approaches for prediction based on high dimensional omics data sets. They provide variable importance measures to rank predictors according to their predictive power. If building a prediction model is the main goal of a study, often a minimal set of variables with good prediction performance is selected. However, if interpretation is more important, approaches that aim to identify all relevant variables should be preferred.

We evaluated several variable selection procedures based on simulated data as well as publicly available experimental methylation and gene expression data. Our comparison included the Boruta algorithm, the vita method, recurrent relative variable importance (r2VIM), a permutation approach (PERM) and its parametric variant (Altmann) as well as recursive feature elimination (RFE).

In the simulation studies, Boruta was the most powerful approach, followed closely by the vita method. Both approaches demonstrated similar stability in variable selection, although vita was the most robust approach under a pure null model without any predictor variables related to the outcome. In the analysis of the different experimental data sets, vita demonstrated slightly better stability in variable selection and was less computationally intensive than Boruta.

In conclusion, we recommend the vita approach for the analysis of high-dimensional data sets. In case of more traditional low-dimensional data, vita cannot be applied, but Boruta is a good alternative.

95 | Genetic Variants Associated with Longitudinal Change of Fasting Glucose

Heejin Jin[1], Soo-Heon Kwak[2], Nam H. Cho[3], Sungho Won[1], Kyong Soo Park[4]

[1]Department of Public Health, Seoul National University, Seoul, South Korea; [2]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea; [3]Department of Preventive Medicine, Ajou University School of Medicine, Suwon, South Korea; [4]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, South Korea

Genome-wide association studies on type 2 diabetes mellitus (T2D) have identified at least 80 genetic risk loci. However, genetic risk factors for longitudinal deterioration of fasting glucose were not thoroughly evaluated. The aim of this study is to identify genetic variants associated with longitudinal change of fasting glucose over time. A total of 6,268 individuals from Ansung-Ansan (KARE) Cohort Study who did not have T2D at baseline examination were investigated and 2,250 European white participants of the Multi-Ethnic Study of Atherosclerosis (MESA) cohort were included with the same conditions as KARE cohort. Fasting glucose of each subject was measured every two years from 2001 to 2012 in KARE Cohort and from 2000 to 2007 in MESA Cohort. Affymetrix Genome-wide Human single nucleotide polymorphism (SNP) Array 5.0 was used for genotyping. IMPUTE2 was utilized for genotype imputation using 1,000 Genomes Project Phase 3 as reference. We considered linear mixed model analysis with two random effects, intercept and slope over time, for each subject adjusting for age and sex. We performed meta-analysis of the two studies using METAL. None of the variants showed genome-wide significant ($p<1.0 \times 10^{-8}$) association for SNP interaction with time due to limited power. However, there were two variants with suggestive evidence of association ($p<1.0 \times 10^{-6}$): rs13147108 ($p = 1.53 \times 10^{-7}$) near MIR1255B1, rs6947411 ($p = 4.47 \times 10^{-7}$) in CDHR3. Among the known loci for T2D, two variants (rs7178572 in HMG20A, rs7177055 near HMG20A) were nominally ($p<0.01$) associated with longitudinal change in fasting glucose.

96 | Epistasis Detection for Human Complex Diseases in Structured Populations

Fentaw Abegaz[1], Kridsadakorn Chaichoompu[1], Van Steen Kristel[1,2]

[1]GIGA-R Medical Genomics – BIO3, University of Liège, Liège, Belgium; [2]WELBIO, University of Liège, Liège, Belgium

Epistasis, or interaction between genes, has been identified as a component of complex phenotypes in a number of studies. When explaining multifactorial trait variation in humans, gene-gene interactions should not be ignored and potentially complex population-dependent modes of inheritance need to be assumed. Even simple genetic diseases may be complex. For example, Mendelian disorders such as Hirschsprung's disease and cystic fibrosis are documented examples of epistasis where modifier genes have been identified to affect phenotypic differences. In general, epistasis studies help to identify novel drug targets and biomarkers relevant to the underlying mechanisms of disease that are not captured by single locus analysis. Moreover, there have been various studies to find pharmacogenetic evidence of epistatic interactions underlying drug resistance, for example, in malaria, epilepsy, and influenza.

Despite the fact that more and more researchers explore epistasis or genome-wide association interaction (GWAI) studies in an attempt to discover more of the hidden or missing heritability of complex traits, there are still several important challenges and considerations to bear in mind. This study aims to investigate the effect of population substructures and admixture on epistasis detection and to develop and apply remedial measures to confounding by shared genetic ancestry in epistasis analyses. Both real-life data and synthetic data will be used for illustration. Starting point is the versatile epistasis analysis tool Model-Based Multifactor Dimensionality Reduction (MB-MDR). It is non-parametric in that it does not make any assumptions about the epistasis inheritance model and has several advantages over classic regression-based detection tools.

97 | Characterizing Heritability, Pleiotropy and Functional Impact in Primary Biliary Cholangitis Using Pre-existing Genome-wide Data

Lynsey S. Hall[1], George F. Mells[2], Heather J. Cordell[1], UK-PBC Consortium

[1] Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom; [2] Academic Department of Medical Genetics, Cambridge University, Cambridge, United Kingdom

Primary Biliary Cholangitis (PBC) is a rare autoimmune cholestatic liver disease, often associated with other autoimmune conditions. To date, genetic studies of PBC have confirmed associations at HLA and identified 33 non-HLA risk loci. In the current study, a range of genetic analyses were applied to non-HLA results from the most recent PBC genome-wide meta-analysis to estimate SNP-based heritability (h^2_{SNP}), test pleiotropy between PBC and other health-related traits, and provide updated functional annotation.

The estimated $h^2_{SNP} = 0.34$, consistent with previous epidemiology-derived estimates. PBC was genetically correlated with systemic lupus erythematosus (SLE), inflammatory bowel disease, ulcerative colitis, mean putamen volume, Crohn's disease, intracranial volume, multiple sclerosis, schizophrenia, body mass index (BMI), Parkinson's disease, rheumatoid arthritis (RA) and bipolar disorder. Polygenic profiling of traits which had no sample overlap with PBC indicated a significant association between PBC and polygenic scores for BMI, RA, and SLE, with scores for SLE explaining 3% of the variance in PBC case/control status.

PBC-associated locus 11q23.3 was associated with increased expression of *TREH* in liver, pancreas and stomach tissues. PBC-associated locus 12q24.12 was associated with the liver metabolite kynurenine. Dysfunctional activity in the kynurenine pathway has previously shown association with SLE, schizophrenia and multiple sclerosis. Gene-based analyses confirmed previously identified genes and gene-sets, in addition to some novel findings.

Our study suggests that there is still much information to be extracted from currently existing datasets. Interrogating this in terms of pleiotropic effects and functional impact can aid the understanding of disease pathways and biology.

99 | Machine Learning Optimised for Personalized Medicine: Predicting Lifetime and Recurrent Depression in the Generation Scotland Cohort Study

Viktoria-Eleni Gountouna[1], Archie Campbell[1], David J. Porteous[1,2], Kristin K. Nicodemus[1,2]

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [2] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom

Major Depressive Disorder (MDD) is a common psychiatric disorder. Data-driven approaches to predict lifetime risk for MDD and single vs. recurrent MDD would be a step forward in realising the promises of personalised medicine.

We applied Machine Learning Algorithms (MLAs) to the Generation Scotland cohort study (*N*>21,000), including phenotypic and genotypic data. The cohort was divided into a training and held-out test set for both lifetime and recurrent MDD. In the training set, cross-validation was used to estimate optimal hyperparameters for multiple MLAs, including Boosting, Support Vector Machines, Neural Networks and Penalised Regression. Using optimal hyperparameters, we ran each MLA on the training set, then used the independent test data in this model to calculate Area Under the Curve (AUC) values to assess performance. We used the rank aggregation Markov Chain 4 (MC4) algorithm, originally designed for meta-Internet search rankings, to determine the smallest set of predictors with equal predictive value vs. using all predictors.

AUC values for lifetime MDD were between 0.81-0.84; for recurrent MDD AUC values were between 0.69-0.76; all significantly higher than chance. Optimal performance using MC4-ranked predictors was obtained using 20 variables for lifetime MDD (AUC = 0.84) compared to 155 variables, and 10 for recurrent MDD (AUC = 0.76) compared to 180.

The MC4 ranked subsets performed equally well to the full subset, although it is likely other subsets could perform similarly. These highly-predictive variables could be easily collected in clinic to assist in accurate diagnosis and preventative treatment for recurrent MDD.

100 | Computing Competing Risks Based on Family History in Genetic Disease with Variable Age at Onset

Alexandra Lefebvre[1,2], Gregory Nuel[3,4]

[1] UPSud, Paris-Saclay, Orsay, France; [2] Institut Curie, Paris, France; [3] LPMA, UMR CNRS 7599, Paris, France; [4] UPMC, Sorbonne universités, Paris, France

When considering a genetic disease with variable age at onset (ex: diabetes, familial amyloid neuropathy, cancers, etc.), computing the individual risk of the disease based on family history (FH) is of critical interest both for clinicians and patients. Such a risk is very challenging to compute because: 1) the genotype X of the individual of interest is in general unknown; 2) the posterior distribution P(X | FH,T>t) changes with t (T age at disease onset for the targeted individual); 3) the competing risk of death is not negligible.

In this work, we present a modelization of this problem using a Bayesian network mixed with (right-censored) survival outcomes where hazard rates only depend on the genotype of each individual. We explain how belief propagation can be used to obtain posterior distribution of genotypes given the FH, and how to obtain a time-dependent posterior hazard rate for any individual in the pedigree. Finally, we use this posterior hazard rate to compute individual risk, with or without the competing risk of death.

Our method is illustrated using the Claus-Easton model for breast cancer (BC). This model assumes an autosomal dominant genetic risk factors such as non-carriers (genotype 00) have a BC hazard rate h0(t) while carriers (genotypes 01,10 and 11) have a (much greater) hazard rate h1(t). Both hazard rates are assumed to be piecewise constant with known values (cuts at 20,30,…,80 years). The competing risk of death is derived from the national French registry.

101 | Replication of Epigenome-wide Associations Related to Body Mass Index Using the Infinium MethylationEPIC BeadChip on Repeated Samples

Sophia Harlid[1], Robin Myte[1], Bethany Van Guelpen[1]

[1]Department of Radiation Sciences, Umeå University, Umeå, Sweden

Previous studies of both blood and tissue have identified >200 CpG sites with differential DNA methylation related to obesity. Genes associated with such sites might help us identify pathways related to chronic diseases such as cancer. We performed an epigenome-wide study of body mass index (BMI) and DNA methylation using the Illumina MethylationEPIC BeadChip, which measures CpG methylation at >850 000 CpG sites. Study subjects were from the Västerbotten Intervention Programme (VIP), a population-based cohort with >100 000 participants. This study included 138 VIP participants, each with two blood samples and extensive anthropometric and questionnaire data available (repeated measures 10 years apart). Primarily, we attempted replication of 187 CpG sites previously reported to associate with BMI based on data from the older Illumina HumanMethylation450 BeadChip. Associations between DNA methylation and BMI were determined by fitting linear mixed models for each CpG site. Out of 187 previously reported sites, 178 were accessible on the EPIC array, of which 38 exhibited both significant *P* values (<0.05) and the same direction of association as previously reported and 119 CpG sites exhibited the same direction of association without reaching statistical significance. For sites not present on the older 450K array, none passed the threshold for genome-wide significance. However, three sites in the genes *CTBP2*, *MIR22HG* and *KCNAB2* were significant at *P* values $<10^{-5}$. *CTBP2* is especially interesting as it is suppressed by the obesity-related microRNA MiR-342-3p, which enhances adipogenesis. Although small, this study validates and extends the evidence base for obesity-related DNA methylation.

102 | Possible Association between Polygenic Risk for Psychiatric Disease and Deep Grey Matter Volume in Preterm Infants

Harriet E. Cullen[1], Saskia Selzam[2], Gareth Ball[1], Paul Aljabar[3], Michelle L. Krishnan[1], Serena J. Counsell[1], A. David Edwards[1]

[1]Centre for the Developing Brain, King's College London, London, United Kingdom; [2]Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom; [3]Biomedical Engineering Department, King's College London, London, United Kingdom

This work uses results from a meta-analysis by the Psychiatric Genomics Consortium (PGC) to investigate how genetic risk for psychiatric disorders correlates with deep grey matter (DGM) volume in preterm infants. Preterm birth is a major cause of neurodevelopmental impairment and is strongly associated with psychiatric disease. DGM structures are particularly vulnerable following preterm birth.

Our sample comprised 194 unrelated preterm infants with no marked cerebral pathology (104 males, 90 females), mean gestational age (GA) 29.7 weeks, mean postmenstrual age at scan 42.6 weeks. Deformation-based morphometry was used to estimate absolute local volumes of the thalamus, subthalamic nucleus, caudate nucleus and lentiform nucleus with registration to a 40-week neonatal template. Infant DNA, extracted from saliva, was genotyped for common variants, genome-wide. Genetic risk scores were computed using a meta-analysis by the PGC that identified risk loci for five major psychiatric disorders: autism spectrum disorder, attention deficit hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia. SNP effect sizes from this study were used to obtain individual polygenic risk scores (PRS) in our cohort.

PRS were corrected for the first ten principle components of the ancestry matrix to adjust for population stratification then tested for their association with DGM volumes (corrected for intracranial volume and GA). Our analysis showed a negative association between psychiatric PRS and both lentiform nucleus volume ($\beta = -0.21$, $p = 0.0008$) and subthalamic nucleus volume ($\beta = -0.18$, $p = 0.01$). Our data suggest that common psychiatric genetic risk may be associated with smaller DGM volumes in preterm infants.

103 | An Improved Polygenic Risk Score for Risk Prediction in Breast Cancer

Nasim Mavaddat[1], Kyriaki Michailidou[1,2], Peter Kraft[3], Montserrat Garcia-Closas[4], Jacques Simard[5], Douglas F. Easton[1,6] on behalf of the Breast Cancer Association Consortium

[1]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, United Kingdom; [2]Department of Electron Microscopy/Molecular Pathology, Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus; [3]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville,

Maryland, United States of America; [5]Genomics Center, Centre Hospitalier Universitaire de Québec Research Center and Laval University, Quebec City, Canada; [6]Department of Oncology, University of Cambridge, Cambridge, United Kingdom

Stratification of women according to the risk of developing breast cancer can improve screening and prevention by targeting those most likely to benefit. Polygenic risk scores (PRSs) summarizing the effect of common susceptibility variants facilitate such stratification, but the clinical utility of current PRSs is limited. Improved stratification may be obtained by increasing sample size, including sub-genome-wide significant SNPs in the PRS and optimising for prediction of subtype-specific disease.

We analysed data from 94,094 cases and 75,017 controls of European ancestry from 69 studies in the Breast Cancer Association Consortium. The dataset was divided into training (90%) and test (10%) sets. Samples were genotyped using two genome-wide arrays with additional SNPs inferred by imputation: ~7 million SNPs were analysed. SNPs were selected first by their association P value in the training set, and subsequently by step-wise forward regression within 1Mb regions across the genome.

The best PRS (268 SNPs) was obtained at a P value cut off of $<10^{-5}$. The OR per 1 SD was OR = 1.65(95%CI = 1.58-1.73) vs OR = 1.49(95%CI = 1.43-1.57) for a 77-SNP PRS derived previously, with area under the receiver operator curve = 0.64 vs 0.61 respectively. Compared to women in the middle quintiles for subtype specific PRSs, those in the highest quintiles of risk had 2.4 and 1.8 fold increased risks, and in the lowest quintile 0.5 and 0.7 fold decreased risks, of developing ER-positive and ER-negative disease respectively. We validated findings in an independent prospective data-set. Future work includes incorporating genomic features to prioritise SNPs for inclusion in the PRS.

104 | Association of Polygenic Risk Scores with the Risk of Chronic Lymphocytic Leukemia (CLL) and Monoclonal B-Cell Lymphocytosis (MBL)

Geffen Kleinstern[1], Nicola Camp[2], Lynn Goldin[3], Timothy G. Call[4], Neil E. Kay[4], J. Brice Weinberg[5], Celine M. Vachon[1], Curt Hanson[6], James R. Cerhan[1], Neil E. Caporaso[3], Susan L. Slager[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America; [2]Department of Internal Medicine, Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, United States of America; [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America; [4]Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, Minnesota, United States of America; [5]Duke University and V.A. Medical Centers, Durham, North Carolina, United States of America; [6]Mayo Clinic College of Medicine, Rochester, Minnesota, United States of America

We previously found that a polygenic risk score (PRS) of 34 CLL susceptibility SNPs was associated with CLL risk. However, the samples used to identify these SNPs were the same ones to show the association of PRS with CLL risk. Herein, we validate the PRS and CLL association in an independent sample. We further assess its effect with risk of MBL, a precursor state to CLL.

We utilized genotype data from 203 CLLs and 95 MBLs, from the Genetic Epidemiology of CLL Consortium, along with 1,267 MAYO controls. The PRS was based on 41 CLL susceptibility SNPs with the effect sizes obtained from the CLL GWAS of 4,478 cases and 13,213 controls. We categorized the PRS into quintiles using the cutoff points based on the PRS distribution within the GWAS controls. We estimate ORs and 95% CIs adjusted for age and sex.

We confirmed an association of the PRS with CLL risk (P<0.0001). Those in the top 20% of the risk distribution had a 3.73-fold (CI = 2.35-5.93) increased risk for CLL compared to the middle quintile. We also found a significant association of the PRS with MBL risk (p<0.0001) with those in the highest quintile having a 4.34-fold (CI = 2.21-8.50) increased risk for MBL compared to the middle quintile.

In conclusion, for the first time we validated the PRS association with CLL risk in an independent set of CLL cases and controls. We also provide evidence for a strong association with MBL risk. The PRS may provide a means to stratify risk of CLL and MBL.

105 | Characterization of Methods for Familial Aggregation of Traits in Large Pedigrees

Christian X. Weichenberger[1], Johannes Rainer[1], Francisco S. Domingues[1]

[1]Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy

Pedigree information obtained in population studies provides an asset that can be used as a first step in investigating the underlying genetics of complex diseases. During the last decades, several methods have been presented to detect aggregation of traits in families, most of them utilizing the kinship coefficient as a means to quantify the relationship between family members and as a measure to highlight individuals with a high degree of affected relatives.

In this work, we compare various previously published and novel methods that have been summarized recently in our Bioconductor R package, FamAgg. In particular, we have constructed a simulation environment in which we investigated these methods' ability to detect families that are affected with a trait that follows a dominant Mendelian inheritance pattern. In a setup with 415 families consisting of 16,719 individuals from up to four generations, we studied the methods' performance when varying parameters for trait penetrance, prevalence, number of affected families, and number of generations that are affected by the trait. This allows us to highlight tests that perform well under various simulated conditions and discuss resulting application scenarios.

106 | Estimation of Heritability of the Disease on the Binary Trait using Liability Threshold Model

Wonji Kim[1], Duck-Woo Kim[2], Sung Il Kang[2], Sukyoung Bang[2], Sang-A Lee[2,] Sungho Won[1,3,4]

[1] Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, South Korea; [2] Department of Surgery, Seoul National University Bundang Hospital, Seongnam, South Korea; [3] Department of Public Health Science, Seoul National University, Seoul, South Korea; [4] Institute of Health and Environment, Seoul National University, Seoul, South Korea

Numerous methods for estimating the heritability have been developed so far, but each inherently has their own potential bias. In particular, unlike quantitative traits, binary traits often require additional assumptions and the heritability can be estimated in limited circumstances. In this study, we developed a heritability estimation algorithm applicable to the general pedigree data based on the liability threshold model for binary traits. Liability is used for unobserved latent variable in EM-related algorithm. It also allows estimation of effects for the covariates affecting the location parameter of liability. In the simulation study, we investigated various situations where some prevalences, heritabilities and pedigree structures are compounded. We also applied the proposed method to the colorectal cancer data collected at the Seoul National University Hospital (South Korea).

107 | Comparison of Whole Genome Sequencing Data to Imputation Data for Cases with Venous Thromboembolism from the GENEVA Study

Brandon J. Coombes[1], Mariza de Andrade[1]

[1] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America

The advent of whole genome sequencing (WGS) allows a comprehensive view of the human genome. Although the cost of WGS continues to become cheaper, genotyping arrays combined with imputation to the entire genome are argued to be a more cost-effective strategy for studies. However, it is unclear how well current imputation strategies perform in a large scale study. Recently, WGS was obtained on a set of 1010 cases previously genotyped and analyzed as part of the Gene Environment Association Studies (GENEVA) genome-wide association study of venous thromboembolism (VTE). VTE cases were consecutive Mayo Clinic outpatients with objectively-diagnosed deep vein thrombosis and/or pulmonary embolism residing in the upper Midwest and are largely a European American sample (58% female). These cases were originally genotyped using the Illumina 660W-Quad BeadChip. After quality control filtering, 561,423 SNPs were observed from this array. Imputation analyses were initially performed using IMPUTE version 2 to impute around 38 million variants. The imputation analysis has since been updated using the recently developed minimac3 software on the Michigan Imputation Server. A comparison between WGS and the genotype array will be made to evaluate the agreement between the variants called on both platforms. We will then evaluate the imputation accuracy of IMPUTE2 and minimac3 compared to WGS across each chromosome. Particular focus will be placed on the imputation accuracy within genomic regions important to VTE such as Factor V. We expect imputation accuracy will be non-uniform across the genome. With this research, we hope to shed light on the debate between WGS and imputation.

108 | A Comparison between Genetics Papers Relating to Immune Disorders and Psychiatric Disorders

Mahmoud El-Haj[1], Scott Piao[1], Paul Rayson[1], Jo Knight[2]

[1] School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, United Kingdom; [2] Data Science Institute and Medical School, InfoLab21, Lancaster University, Lancaster, United Kingdom

The explosion of literature in the field of genetics makes it hard to keep apace of new knowledge. Techniques developed in Natural Language Processing and Corpus Linguistics can help. Previously such techniques have been used to perform tasks such as identifying gene-gene or gene-phenotype interactions. We hope to identify words that will provide new clues to disease aetiology.

We have performed two searches in PubMed to capture corpora of genetic literature. We have a corpus based on immune-related diseases (21,422 papers, 4,815,641 words) and one based on psychiatric diseases (15,151 papers, 2,817,417 words). The searches have been adapted to ensure appropriate literature coverage. For example whilst including immun* in the abstract picks up papers on many diseases such as psoriasis, the same approach using the term psych* is not as effective.

We use Wmatrix (https://ucrel.lancs.ac.uk/wmatrix/) to compare the corpora of literature. First, we investigate the comparative proportional representation of words. We separate the results into expected results (based on immunological/psychiatric terms defined by relevant book indexes) and novel findings.

Many subject specific words have a much higher proportional representation in one corpus (e.g. schizophrenia). Other less predictable words such as “risk” are also found to be more frequent in psychiatric literature.

The increased proportional representation suggests that language is used different despite both corpora describing genetic studies of a complex trait.

Further refinement of the searches and exploration of the results using techniques such as semantic tagging and dispersion estimation.

109 | Gene-Environment Interactions between 65 Newly Identified Breast Cancer Susceptibility Loci and Non-Genetic Risk Factors in Association with Breast Cancer Risk

Pooja Middha[1], Sara Lindström[2], Audrey Jung[1], Montserrat Garcia-Closas[3,4], Pascal Guénel[5], Peter Kraft[6], Jacques Simard[7], Douglas F. Easton[8], Roger L. Milne[9,10], Jenny Chang-Claude[1,11] for the Breast Cancer Association Consortium

[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; [2]Harvard School of Public Health, Department of Epidemiology, Boston, Massachusetts, United States of America; [3]National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Rockville, Maryland, United States of America; [4]The Institute of Cancer Research, Division of Genetics and Epidemiology, London, United Kingdom; [5]Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France; [6]Harvard School of Public Health, Department of Epidemiology, Department of Biostatistics, Boston, Massachusetts, United States of America; [7]Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Canada; [8]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care and Department of Oncology, University of Cambridge, Cambridge, United Kingdom; [9]Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia; [10]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia; [11]Genetic Tumour Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Recent genome-wide association studies (GWAS) conducted in studies participating in the Breast Cancer Association Consortium (BCAC) and Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium (DRIVE) identified 65 novel loci associated with overall breast cancer (BC) risk and 10 loci associated with estrogen receptor negative (ER-) BC risk. We investigated whether the identified SNP-BC associations are modified by non-genetic BC risk factors in women of European ancestry.

Data from the OncoArray project (44,341 BC cases and 51,333 controls from 37 studies) and from the iCOGS project (28,401 cases and 27,180 controls from 19 studies) were used. Multiplicative interactions were assessed between 65 SNPs and 14 BC risk factors using logistic regression and evaluated by the likelihood ratio test. All models were adjusted for age, study, and ancestry-informative principal components. Analyses were conducted for the two data sets separately and the results combined using fixed effects meta analysis.

Preliminary analyses show an interaction between rs67958007(10p14) and parity (yes *vs* no) after Bonferroni correction ($p<7.6 \times 10^{-4}$) in relation to overall BC risk (OR_{int} [oncoarray] = 1.18, 95% CI: 1.08,1.30; OR_{int} [icogs] = 1.09, 95% CI: 0.97, 1.23; OR[meta] = 1.15, 95% CI: 1.06, 1.23, p = 0.0003). The per-allele ORs of rs67958007 were 0.96 in nulliparous and 1.05 in parous women. Further analyses are being conducted for SNP-risk factor interactions according to ER status.

Overall the associated effects on breast cancer risk of recently discovered susceptibility loci are not strongly modified by environmental risk factors.

110 | Genome-wide Association Study of Internal Hematopoietic Cellular Traits

Parsa Akbari[1], John Danesh[1], David J. Roberts[2,3], Willem H. Ouwehand[4], Nicole Soranzo[5], Adam S. Butterworth[1], William J. Astle[1]

[1]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; [2]Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford, United Kingdom; [3]Department of Haematology, Churchill Hospital, Headington, United Kingdom; [4]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, United Kingdom; [5]Department of Human Genetics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton United Kingdom

Hematopoietic development includes the maturation and development of platelets, red blood cells, and white blood cells such as neutrophils, monocytes, and lymphocytes. Variation in the development and function of haematopoietic cells plays an important role in human diseases including Rheumatoid Arthritis and Chronic Heart Disease.

Previously, population-scale epidemiological studies of haematological traits have focused on clinical indices (such as cell count and volume) that are easily measured. GWAS of these traits in 170,000 participants identified ∼2,700 loci associated with red cell, white cell and platelet indices (Astle et al, *Cell*, 2016). However, these indices may not accurately reflect biological and clinically relevant processes, we assayed ∼39 internal haematopoietic cellular traits (e.g., cell granularity and morphology, or nucleic acid content) in ∼50,000 blood donors from the INTERVAL study and conducted GWAS using ∼30 million imputed variants. We identified 640 loci influencing at least one cellular trait, of which only roughly 50% were detected with the clinical traits in a four-fold larger GWAS. As well as detecting loci with strong links to haematological processes (e.g., LYST and DEFENSIN), we also identified loci for which the causal gene is not immediately obvious. Fine-mapping, integration with epigenomic data and functional experimentation will help to prioritise candidates at these loci.

Our results suggest that genes that regulate the internal cellular traits are frequently distinct from those regulating clinical indices. Identification of the mechanisms by which these loci exert their effects will help to shed light on the biological processes determining these cellular traits.

111 | Phenotypic and Genetic Analysis of Cognitive Performance in Major Depressive Disorder in the Generation Scotland: Scottish Family Health Study

Joeri J. Meijsen[1,2], Archie Campbell[1], Caroline Hayward[3], David J. Porteous[1,2], Ian J. Deary[2,4], Riccardo E. Marioni[1,2], Kristin K. Nicodemus[1,2]

[1] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [2] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom; [3] MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; [4] Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom

Lower performances in cognitive ability in individuals with Major Depressive Disorder (MDD) have been observed. Understanding cognitive performance in MDD could provide insight into the aetiology of MDD as a whole. Using a large, well characterised cohort (N = 7012), we tested for: differences in cognitive performance by MDD status and a gene (single SNP or polygenic score) by MDD interaction effect on cognitive performance. Linear regression was used to assess the association between cognitive performance and MDD status in a case vs. control, single episode (sMDD) vs. recurrent MDD (rMDD) and control vs. rMDD study design. Scores on verbal declarative memory, executive functioning, vocabulary, and processing speed were examined. Cognitive measures showing a significant difference between groups were subsequently analysed for genetic associations. Those with rMDD showed lower processing speed vs. controls and sMDD ($\beta = -2.44$, p-value $= 3.6 \times 10^{-4}$; $\beta = -2.86$, p-value $= 1.8 \times 10^{-03}$, respectively). We found significantly higher vocabulary scores in MDD cases vs. controls ($\beta = 0.79$, p-value $= 2.0 \times 10^{-6}$), and in rMDD vs. controls ($\beta = 0.95$, p-value $= 5.8 \times 10^{-5}$). Observed differences were not linked to significant single locus associations. Polygenic scores created from a processing speed meta-analysis Genome-wide Association Study explained 1% of variation in processing speed in the sMDD vs. rMDD study (p-value $= 1.7 \times 10^{-3}$) and 0.5% of variation in the control vs. rMDD study (p-value $= 1.6 \times 10^{-10}$). Individuals with rMDD showed lower processing speed and executive function although showing higher vocabulary performance. Within MDD, persons with rMDD show lower processing speed and executive function scores relative to individuals experiencing a single episode.

112 | Hierarchical Model Selection with Quantile Regularization

Kan Wang[1], David V. Conti[1]

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

In fine-mapping studies, model selection procedures are often used to identify the number of signals and the underlying SNPs driving them. The least absolute shrinkage and selection operator (LASSO) is one powerful approach that generates a parsimonious model by adopting a L1 penalty on the estimates. It is equivalent to placing a double exponential prior on the effect estimates. In previous work, we have incorporated prior biological information into the model selection process via either a normal prior on the means of the estimates or a probit prior on the probability of inclusion. Here, we develop a novel alternative that expand upon the regularized regression framework and incorporates a quantile loss (Q1) penalty. This is equivalent to placing a conditional quantile prior on the effect estimates. The flexibility of choosing different conditional quantiles can provide better characterization of the data if there exists considerable heteroscedasticity between the conditional distributions of the estimates for true and null variants. Our framework allows for either pre-specifying the quantile or treating it as a tuning parameter. Importantly, LASSO is a special case of our method in the presence of no functional information and the second stage quantile is fixed at the median. We demonstrate via simulation that Q1 regression generates more accurate estimates as compared to alternative regularized regression approaches, especially in the presence of informative predictor-level information. As an applied example, we implement Q1 regression on prostate cancer fine-mapping data for men of European ancestry to prioritize functional variants at multiple susceptibility loci.

113 | Tissue-Specific Trans-Ancestral Analysis of Genetically Regulated Expression with 15 Metabolic and Cardiovascular Traits Identifies Novel Loci

Heather M. Highland[1,2], Lauren E. Petty[1,3], Craig L. Hanis[1], Esteban J. Parra[4], ARIC investigators, Jennifer E. Below[1,3]

[1] Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2] Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America; [3] Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America; [4] Department of Anthropology, University of Toronto at Mississauga, Mississauga, Canada

Cardiometabolic diseases, including obesity, hypertension, and dyslipidemia are a leading cause of health care expenditures and death in the United States. Measures of cardiometabolic diseases are highly heritable; however, interpretation of genome-wide association results is difficult due to a lack of biological context. To better understand the genetic etiology of complex cardiometabolic traits, we predicted individual-level gene expression from common variants using PrediXcan and determined genes with differentially predicted expression for a series of phenotypes. PrediXcan aggregates evidence from functional variation by leveraging transcriptome data to collapse common expression quantitative trail loci (eQTLs) into tissue-specific imputed levels of gene expression. We tested the association of predicted genetically regulated gene expression (GREx) with 15 cardiometabolic traits that included blood lipid levels, body mass index, height, blood pressure, fasting glucose and insulin, RR interval, fibrinogen level, factor VII level, and white blood cell and platelet counts in 15,755 individuals across three ancestry groups, resulting in 29 novel gene-phenotype associations (p-value$<2.5 \times 10^{-6}$). Top associations were followed up in publicly available summary datasets using MetaXcan. Top findings include *ZNF441* expression in the pancreas associated with low-density lipoprotein cholesterol *TAF6L* expression in whole blood and A*POL5* expression in visceral adipose tissue associated with triglyceride levels. Predicted expression levels were compared to whole blood expression measured by RNA-seq from a subset of 175 European ancestry participants.

114 | Findings from a Longitudinal Metabolome-wide Association Study of Cognitive Decline in Healthy Adults with Increased Risk for Alzheimer's Disease

Burcu F. Darst[1,2], Matthew J. P. Rush[1,3,4], Paul D. Hutchins[1,3,4], Jason D. Russell[1,4,5], Rebecca L. Koscik[1,6], Sanjay Asthana[1,6-8], Sterling C. Johnson[1,6-8], Kirk J. Hogan[1,6,9], Joshua J. Coon[1,3-5,10], Corinne D. Engelman[1,2,6,8]

[1] University of Wisconsin, Madison, Wisconsin, United States of America; [2] Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [3] Department of Chemistry, University of Wisconsin, Madison, Wisconsin, United States of America; [4] Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin, United States of America; [5] Morgridge Institute for Research, Madison, Wisconsin, United States of America; [6] Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [7] Geriatric Research Education and Clinical Center, Wm. S. Middleton Memorial VA Hospital, Madison, Wisconsin, United States of America; [8] Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [9] Department of Anesthesiology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America; [10] Department of Biomolecular Chemistry, University of Wisconsin, Madison, Wisconsin, United States of America

A longitudinal multi-omics examination of biomarker trajectories prior to Alzheimer's disease (AD) diagnosis is crucial. Metabolomic profiles were quantified with mass spectrometry on longitudinal plasma samples from the Wisconsin Registry for Alzheimer's Prevention (WRAP), a cohort of asymptomatic participants at enrollment, enriched with an AD parental history. Analyses include 28 participants with cognitive decline and 55 age- and gender-matched cognitively stable participants, each with up to four visits. A metabolome-wide association study (MWAS) was performed using linear mixed effects to assess longitudinal metabolite changes. Findings were further tested by clinical diagnosis using the Wisconsin Alzheimer's Disease Research Center (W-ADRC), an independent cohort including 27 AD, 17 mild cognitive impairment (MCI), and 29 cognitively healthy participants with plasma metabolomics for three visits. Of 615 metabolites tested in the WRAP cohort, seven met a liberal significance threshold (p-value<0.01). Findings were unchanged by controlling for an *APOE* risk score. Three of these were unidentified metabolites. Four were phospholipids all present at higher levels in decliners during younger ages, but lower levels during older ages, compared to non-decliners. W-ADRC analyses separately comparing AD and MCI to controls showed a consistent relationship for one of these phospholipids. These preliminary results suggest that phospholipid trajectories may differ between those with cognitive decline, MCI, and AD versus controls, such that cognitively impaired individuals have higher levels earlier in life, but lower later in life than controls, independent of risk due to *APOE*. This relationship is being further explored by performing integrated metabolomic and genomic analyses.

115 | Development and Application of Methodology for the Analysis of Rare Genetic Variants with Time to Event Outcomes Using SurvivalGWAS_RV

Hamzah Syed[1], Andrea L. Jorgensen[1], Andrew P. Morris[1]

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

Methodology and software for the analysis of common variants within genome-wide association studies (GWAS) have been extensively developed and applied to a range of different outcomes, including "time to event" data. These approaches have identified many loci for a variety of complex traits and diseases, but which account for only a small proportion of the genetic variance. Rare genetic variants may account for some of the "missing heritability" of these traits. Rare variants are most often analysed within "functional units" using burden or dispersion tests, each with their own benefits and limitations dependent on the underlying genetic architecture of the trait. Software implementing these tests, such as EPACTS and GRANVIL, are well developed for binary and quantitative traits. However, the methodology has not been widely adapted for time to event phenotypes, where the outcome of interest could be time to disease remission or occurrence of an adverse drug reaction. To address this need, we have developed the SurvivalGWAS_RV software implemented using C# and run on Linux operating systems. SurvivalGWAS_RV is capable of handling the scale and complexity of whole-genome sequence data. SurvivalGWAS_RV currently supports analysis using the burden test (with optional Madsen-Browning weighting) within a Cox proportional hazards or Weibull regression model. In conclusion, we introduce a new console application analysis tool for rare genetic variants with time to event outcomes. SurvivalGWAS_RV will aid in the discovery of novel genes associated with patient response to treatment for a range of complex human diseases, ultimately allowing personalisation of therapeutic intervention.

116 | Re-Evaluation of SNP Heritability in Complex Human Traits

Doug Speed[1], David J. Balding[1,2]

[1] Genetics Institute, University College London, London, United Kingdom; [2] Centre for Systems Genomics and Schools of Biosciences and Mathematics & Statistics, University of Melbourne, Royal Parade, Melbourne, Australia

SNP-heritability, the proportion of phenotypic variance explained by SNPs, has been reported for many hundreds of traits. Its estimation requires strong prior assumptions about the distribution of heritability across the genome, but the assumptions in current use have not been thoroughly tested. By analyzing imputed data for a large number of human traits, we empirically derive a model that more accurately describes how heritability varies with minor allele frequency, linkage disequilibrium and genotype certainty (the “LDAK Model”). Across 19 traits, using the LDAK Model leads to estimates of common SNP-heritability on average 40% higher than those obtained from the widely-used softwares GCTA and LDSC, indicating that common SNPs explain substantially more heritability than previously thought. Many researchers are using GCTA and LDSC to divide SNP-heritability based according to functional categories; for example, it has been reported that SNPs within DNaseI hypersensitivity sites (DHS) are highly enriched for causal variants, on average explaining 15 times as much heritability as non-DHS SNPs. Using the LDAK Model, estimates of enrichment are far more modest; for example, we estimate DHS SNPs explain only 45% more than non-DHS SNPs.

117 | Customizing the LASSO with External Information

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

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

The Least Absolute Shrinkage and Selection Operator (LASSO) has become popular for fitting high-dimensional data, as it allows the number of variables p to exceed the number of subjects n, sets many of the coefficients to zero (yielding parsimonious models), and is computationally efficient. Choice of the shrinkage parameter is commonly done by cross-validation. We propose using a log-linear model to allow the amount of shrinkage to vary by an amount that depends upon external information, such as bioinformatic annotations or functional assays. To estimate the coefficients of the shrinkage model, we further propose using a marginal likelihood, integrating out the first-level regression coefficients. For normally distributed phenotypes, a closed-form expression is available when the first-level predictors are independent, but is impractical when they are dependent. Instead, we propose a normal approximation to the LASSO, which is easy to implement even for dependent variables. This also allows an approximate solution for binary or other phenotypes relying on an asymptotic normal approximation of the first-level likelihood. We show by simulation that the method shrinks truly null coefficients more than non-null ones across a broad range of parameter choices. We illustrate the approach with data on liver cancer and targeted metabolomics, using a normal model for exposure effects on metabolites and a logistic model for disease and provide estimates of mediation of exposure-disease associations through metabolites singly and in combination. These results are compared with our earlier MCMC implementation of a spike and slab model.

118 | Genome-wide Meta-Analysis of Parent-of-Origin Effects of Asthma in Four Cohorts

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

[1] Centre for Heart Lung Innovation, Faculty of Medicine, University of British Columbia, Vancouver, Canada; [2] Groningen Research Institute for Asthma and COPD (GRIAC), University Medical Center Groningen, University of Groningen, Groningen, 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, Basic Sciences Department, Saguenay, 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, an epigenetic phenomenon where the expression of genes depends on their parental origin, may be one factor explaining missing heritability. We aimed to identify candidate genomic regions for imprinting in asthma using GWAS data from four family-based studies (trios): a) Canadian Asthma Primary Prevention Study (CAPPS), b) Study of Asthma Genes and Environment (SAGE), c) Saguenay–Lac-Saint-Jean Québec

Familial Collection (SLSJ), and d) Dutch Asthma GWAS (DAG). We used a likelihood-based variant of the Transmission Disequilibrium Test as implemented in UNPHASED. Parent-of-origin effects (POE) were tested with parent's sex as a modifier in the analysis. Meta-analysis was conducted using the results of SLSJ (251 trios), DAG (316 trios), and joint CAPPS and SAGE analysis (148 trios), weighted by the number of informative transmissions for each study. Number of SNPs (suggestive significance $p \leq 10^{-5}$) for POE was: seven in SLSJ, four in DAG, and 13 in joint CAPPS and SAGE analysis, with no SNPs overlapping. Fifteen out of 24 SNPs were in or near long non-coding (lnc)RNA genes. LncRNAs are known to be involved in genomic imprinting. In joint CAPPS and SAGE analysis, we showed a POE at a known imprinted gene, *CTNNA3*. Meta-analysis yielded two SNPs with significant POE ($p \leq 10^{-5}$)in *LOC105373804* and *LINC00974|KRT9*. We will conduct further analyses based on multinomial modeling and haplotype estimation (using EMIM software) to confirm our results.

120 | Insights into Iron Metabolism: Discovery of New Genetic Loci Associated with Soluble Tansferrin Receptor

Anna Ramond[1], Praveen Surendran[1], Tao Jiang[1], Qi Guo[1], Joanna MM Howson[1], Adam Butterworth[1], Emanuele Di Angelantonio[1]

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

Iron is an essential element required for oxygen transport and oxidative metabolism. Perturbations in iron metabolism have consequences on human health, including effects on iron deficiency anaemia, oxidative stress, liver disease and metabolic syndrome. Despite the clinical importance of iron metabolism, much is still unknown about mechanisms of iron regulation. Soluble transferrin receptor (sTfR) is a relatively novel marker of iron status used for the diagnosis of iron deficiency anemia, however, little is known about the genetic regulators of this receptor.

We have conducted the largest genome-wide association study (GWAS) of sTfR to date (~40,000 participants) from the British-based INTERVAL study using the UK Biobank Axiom array. We have investigated over 10M imputed and genotyped autosomal variants which were tested for association with sTfR using sex-specific linear mixed models implemented in BOLT-LMM. Sex specific-models were meta-analysed using a fixed-effects meta-analysis.

We identified 17 independent loci associated with sTfR at genome-wide significance ($p<5.10^{-8}$), confirming 3 previously identified loci and discovering 14 new loci. These included regions previously associated with other markers of iron metabolism such as *HFE*, *TMPRSS6,* and *TFRC*, as well as new regions not previously known to be associated with iron markers. Our study provides new insight into the biology of iron metabolism.

Future work will focus on replication of these results in independent studies, and functional annotation of the identified regions to further our understanding of iron biology and its role in diseases.

121 | Assessing the Clinical Utility of Lung Cancer Polygenic Risk Model

Rayjean J. Hung[1,2], Yonathan Brhane[1], Nilanjan Chatterjee[3,4], David Christiniani[5], Neil Caporaso[4], Xifeng Wu[6], Maria Teresa Landi[4], Paul Brennan[7], Christopher I. Amos[8] on behalf of Transdisciplinary Research in Cancer of Lung (TRICL) team of the International Lung Cancer Consortium (ILCCO) OncoArray Group

[1] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada; [2] Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [3] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America; [4] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [5] Department of Environmental Health, Harvard TH Chan School of Public Health, and Massachusetts General Hospital/ Harvard Medical School, Boston, Massachusetts, United States of America; [6] Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America; [7] International Agency for Research on Cancer, Lyon, France; [8] Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, United States of America

Genome-wide association studies uncovered multiple lung cancer susceptibility genes, and consortium efforts greatly increased our ability to investigate the genetic architecture of histological subtypes. However, the clinical utility of these genomic discoveries remains unclear. We, therefore, constructed a risk prediction model with polygenic risk score (PRS) based on 18,316 lung cancer patients and 14,025 controls with European ancestry, via 10-fold cross-validation with elastic net penalized regression. Model calibration was assessed, and it is being validated with UK Biobank data (N = 152,249 participants). To evaluate its potential clinical utility, the PRS distribution was simulated in the National Lung Screening Trial (NLST, N = 50,772 participants). Absolute risk was estimated based on age-specific lung cancer incidence and all-cause mortality as competing risk. A PRS was constructed based on 263 independent lung cancer variants. The lung cancer ORs for individuals at the bottom 10% and top 10% of the PRS distribution were 0.58 (95%CI = 0.52-0.66, $p = 2.7 \times 10^{-20}$) and 1.73 (95%CI = 1.54-1.95, p = 1.38e-17) in the training set, and 0.64 (95%CI = 0.51-0.81, $p = 5.34 \times 10^{-4}$) and 1.74 (95%CI = 1.36-2.25, $p = 3.60 \times 10^{-6}$) in the testing set, versus those at 40 to 60% as the referent group. The area under receiver of characteristics curve (AUC) was 0.74 based on risk factors only, and 0.78 when adding PRS (net reclassification index p-value = 0.0002).

When simulating the PRS distribution in the NLST population, we estimated 49.5% of cases occurred in the top 20% of the individuals with highest lifetime cumulative risk. This study provides insights on how inclusion of well-established genomic information in the risk model can contribute to the risk stratification of the population.

122 | Effect of Bias and Misclassification on Gene-Environment Studies Conducted in Observational Cohort Settings: A Simulation Study

Amanda A. Seyerle[1,2], Colleen M. Sitlani[3], Kari E. North[2], Craig R. Lee[4], Eric A. Whitsel[2,5], Til Stürmer[2,6], Christy L. Avery[2,7]

[1]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, United States of America; [2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [3]Department of Medicine, University of Washington, Seattle, Washington, United States of America; [4]Division of Pharmacotherapy and Experimental Therapeutics, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [5]Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [6]Center for Pharmacoepidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America; [7]Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Gene-environment studies represent an increasingly popular avenue to characterize complex traits. Yet few investigations have evaluated the degree to which gene-environment interaction estimates are influenced by sources of bias that affect estimates of the association between the environmental exposure and outcome. Using pharmacogenomics as our model, we simulated three designs (cross-sectional, repeat cross-sectional, incident exposure), two referent groups (whole cohort, alternate exposure), and two scenarios (extreme or modest environmental effects) of $N = 120{,}000$ participants using 1 million iterations to enable comparisons of 12 studies of the electrocardiographically measured QT interval (QT). For each study, we simulated a causal SNP with minor allele frequency of 25% and index exposure frequency of 17%, using $\alpha = 5 \times 10^{-8}$. When plausible degrees of exposure misclassification (sensitivity = 97%, specificity = 79%) were introduced, substantial bias towards the null in the environment-SNP interaction estimate was observed across all 12 study settings (relative bias range: 25–51%); statistical power was substantially decreased (e.g. relative power decrease of 50–99% to detect a simulated 2ms interaction). In the presence of exposure misclassification, detection of interaction effects of similar magnitude to published QT main effect GWAS (e.g. 2ms) required a repeat cross-sectional design with at least 150,000 participants. Exposure misclassification poses a sizable threat to gene-environment studies, greatly reducing estimated effect sizes, statistical power, and therefore perceived clinical and public health impact. Efforts to reduce the effects of exposure misclassification in gene-environment studies are therefore warranted.

123 | Combining Genetic, Transcriptomic and Clinical Variables to Predict Melanoma-specific Survival

Ernest Mangantig[1], Mark M. Iles[1], Julia A. Newton-Bishop[1], D. Timothy Bishop[1], Jennifer H. Barrett[1]

[1]Leeds Institute of Cancer and Pathology, School of Medicine, University of Leeds, Leeds, United Kingdom

Melanoma-specific survival (MSS) varies greatly by disease stage at diagnosis, but it would be useful to improve prognostic accuracy. Tumour gene expression profiles have been previously found to predict MSS. We combined clinical data with transcriptomic data from primary tumours and genome-wide SNP data to try to improve prediction of MSS. The study was based on the Leeds melanoma cohort of over 2000 patients, with clinical data, long-term follow-up and genome-wide SNP data; gene expression data were available on over 20,000 genes in 700 of the patients. One-third of patients with full data were reserved as a test set. Using all remaining samples, Cox proportional hazards (CPH) regression was used to select the most important clinical predictors. Lasso-penalized CPH was used to select gene expression probes predictive of MSS using cross-validation to choose the appropriate penalty. SNPs predictive of MSS were first filtered for those with $p<0.01$ in univariate analysis and then selected using the same approach. In total, 5 clinical factors, 16 expression probes, and 13 SNPs were selected. Various approaches were used to combine the data, including creating clinical, transcriptomic and genetic risk scores, which were then analysed in the test set in a multivariable CPH. In all approaches, the clinical and transcriptomic factors were both highly significantly related to MSS, while the genetic factors were not. However the clinical and transcriptomic factors were highly correlated, and the transcriptomic factors did not significantly improve prediction over and above clinical predictors in any of the approaches considered.

124 | Novel Mixed Model Algorithm for Fast and Efficient Epigenome-wide Analysis of Complex Traits

Jeffrey R. O’Connell[1]

[1]Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, Maryland, United States of America

Epigenome-wide analysis (EWA) of complex traits is a high-dimensional phenotype analysis as methylation probes are outcomes and the trait an independent variable in the model. Treating the probe as the outcome enables adjusting the analysis for important technical artifacts such as chip and blood cell composition necessary to achieve robust results. In addition to fixed effects, a random effect to model known or cryptic relatedness and population stratification using a genetic relationship matrix is needed, thus requiring mixed model methods. Current approaches for mixed model EWA analyze a single probe at a time, which, for example, requires running 450,000 mixed models for the Illumina 450K chip. This approach is clearly extremely cumbersome, computationally inefficient, and intractable for large sample sizes, but is used because no mixed model algorithms currently exist to tackle the challenge of high-dimensional phenotype data.

We present an elegant and simple solution to the EWA mixed model problem that is thousands of times more efficient than current approaches and can scale to large sample sizes. Our algorithm combines grid search across the heritability parameter space, polynomial interpolation to maximize the likelihood and efficient matrix operations using all phenotypes simultaneously to provide near exact fixed and random effects estimates and p-values. We analyze 8000 subjects and 100,000 probes in under 30 minutes, while 100,000 single probe analyses require over 25 days. Our algorithm extends to efficient methylation quantitative trait locus (mQTL) analysis with SNPs and is implemented into our mixed model software MMAP supporting cluster and cloud computing.

125 | A Comparison of Univariate and Multivariate GWAS Methods for Analysis of Multiple Dichotomous Phenotypes

Yasmmyn D. Salinas[1], Andrew T. DeWan[1], Zuoheng Wang[2]

[1]Department of Chronic Disease Epidemiology, Yale School of Public Health, Yale University, New Haven, Connecticut, United States of America; [2]Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, Connecticut, United States of America

Analysis of multiple phenotypes in genome-wide association studies (GWASs) has the potential to enhance statistical power and allows for exploration of pleiotropy. Multi-trait analyses can be conducted using both univariate and multivariate methods. To select an analytic approach, it is important to understand the performance of available methods. However, comparative evaluations of multi-trait methods have primarily focused on the analysis of quantitative traits. Therefore, this study aimed to evaluate the performance of multivariate GWAS methods for analysis of dichotomous (case/control) phenotypes using simulated data. We focused on three methods implemented through *R* statistical packages—MultiPhen, generalized estimating equations (GEEs), and generalized linear mixed models (GLMMs)—and also compared them to the standard univariate GWAS. We simulated data ($N = 20{,}000$) for one bi-allelic SNP and two case/control phenotypes assuming a classical liability threshold model, and varied the number of traits associated with the SNP, degree of association, trait-specific prevalences, and cross-phenotype correlation. We generated 10,000 replicates and evaluated power using a genome-wide significance level of 5×10^{-8}. Our results show that, in the absence of pleiotropy, multivariate methods outperform the univariate when there are strong, positive cross-phenotype correlations, but that, in the presence of pleiotropy, the univariate approach tends to outperform multivariate methods when the cross-phenotype correlation is positive. GEEs outperformed MultiPhen and GLMMs across most scenarios. This suggests that to maximize GWAS discovery, the use of univariate and multivariate (GEE-based) approaches in parallel can be recommended. This study provides researchers with empirical guidelines for the application of these methods to real data.

126 | Co-inheritance of *APOL1* Risk Variants and Polymorphisms of the *BCL11A*, *G6PD*, *HBA1* and *HBA2* Genes in Sickle Cell Anemia Patients from Nigeria

Bamidele O. Tayo[1], Titilola S. Akingbola[2], Santosh L. Saraf[3], Binal N. Shah[3], Lewis L. Hsu[4], Akinlolu O. Ojo[5], David T. Burke[6], Richard S. Cooper[1], Victor R. Gordeuk[3]

[1]Department of Public Health Sciences, Loyola University Chicago Stritch School of Medicine, Maywood, Illinois, United States of America; [2]Department of Hematology, University of Ibadan, Ibadan, Nigeria; [3]Division of Hematology & Oncology, Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, United States of America; [4]Division of Hematology & Oncology, Department of Pediatrics, University of Illinois at Chicago, Chicago, Illinois, United States of America; [5]Department of Medicine, University of Arizona College of Medicine, Tucson, Arizona, United States of America; [6]Department of Human Genetics, Medical School, University of Michigan, Ann Arbor, Michigan, United States of America

Sickle cell anemia (SCA) is an autosomal recessive disease caused by the well-known GLU-VAL mutation in the *HBB* gene. In addition to chronic hemolytic anemia, SCA patients experience acute and chronic complications including painful vaso-occlusive crisis, acute chest syndrome, avascular necrosis, chronic kidney disease, stroke and pulmonary hypertension. Variations in the severity and frequency of complications have been observed from patient to patient and are known to be influenced by polymorphisms at other genetic loci. Alpha-thalassemia (from a combination of polymorphisms in *HBA1* and *HBA2* genes), *BCL11A* rs1427407, and *APOL1* G1 and G2 have been reported to influence hemolysis, fetal hemoglobin levels, and risk of kidney disease. The objective of the present study was to determine both the distribution and co-inheritance of these variants among SCA patients from Nigeria. We enrolled and genotyped 234 SCA patients, aged 11 years and older, for polymorphisms in

APOL1, *HBA1*, *HBA2*, *G6PD* and *BCL11A* genes by PCR. Among the patients, 56.8% have no alpha-thalassemia deletion (the protective) allele and 21.4% have *APOL1* risk defined as G1/G1, G2/G2 or G1/G2. Also, 50.0% and 52.0% of those with *APOL1* risk do not have the protective alpha-thalassemia and *BCL11A* rs1427407 T alleles that might serve to ameliorate the influence of the *APOL1* risk. These proportions are similar to reports in US and European SCA cohorts. This report provides further support for the use of genetic variants to risk-stratify SCA patients with the goal of personalizing therapy, especially in places with limited access to basic medical care.

127 | A Hierarchical Approach to Genetic Fine Mapping Incorporating Functional Annotation

Virginia A. Fisher[1], L. Adrienne Cupples[1], Ching-Ti Liu[1]

[1] Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America

Genome-wide association studies (GWAS) have identified thousands of loci associated with numerous traits. However, these studies commonly preclude the identification of functional or causal variants in the presence of linkage disequilibrium (LD) within associated loci. Functional annotation provides an independent source of data regarding the potential relevance of each SNP. We propose a hierarchical framework which incorporates functional annotations to prioritize SNPs in biologically relevant categories with evidence of trait association after adjusting for other SNPs in the locus. Specifically, we estimate annotation effects using stratified LD score regression to define variances of random SNP effects within the locus. This leverages association results from the entire genome to detect enrichment within relevant categories while allowing for trait-specific annotation effects. The proposed approach is applicable to both studies with individual-level data and those with only GWAS summary statistics.

We assess method performance in simulated samples of 10,000 individuals with trait heritability of 0.001 at the true causal variant across a range of LD and annotation scenarios. Our method ranks the true causal variant within the top 10 SNPs per locus more frequently than naïve GWAS ranking or penalized multiple regression. We also present results from the GIANT consortium meta-analysis of body mass index in 322,154 individuals of European ancestry. SNPs ranked highly by our method in the *FTO* and *TMEM18* loci have generalized in Asian populations, and show strong association with gene expression in adipose and brain tissue.

128 | Adverse Childhood Experiences Influence DNA Methylation Profile among African American Mothers and Children in the InterGEN Study

Jacquelyn Y. Taylor[1], Veronica Barcelona de Mendoza[1], Yunfeng Huang[2], Kevin Newhall[3], Qin Hui[2], Cindy A. Crusto[4], Yan V. Sun[2,5]

[1] Yale School of Nursing, Yale University, Orange, Connecticut, United States of America; [2] Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, Georgia, United States of America; [3] Vassar College, Poughkeepsie, New York, United States of America; [4] Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, United States of America; [5] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, United States of America

Negative life experiences during childhood, known as adverse child experiences (ACEs), have been associated with cardiovascular disease in adulthood. Epigenetic mechanisms such as including DNA methylation (DNAm), have been proposed as an explanation for this association. To examine the effect of ACEs on DNAm, we conducted a methylome-wide association study of saliva samples from African American mothers ($n = 150$) and children ($n = 148$) from the Intergenerational Impact of Genetic and Psychological Factors on Blood Pressure (InterGEN) study using the 850K EPIC BeadChip. We assessed ACEs exposure using the Trauma Events Screening Inventory-Parent Report Revised (TESI-PRR) for children. After adjustment for smoking, age, cell type heterogeneity, batch effect and multiple testing, we identified 66 and 102 DNAm sites significantly associated with TESI-PRR in mothers and children, respectively. We identified eighteen DNAm sites on 15 known genes including *IST1*, *LRRC75A-AS1,* and *BARX2*, which were methylome-wide significant in both mothers and children. The pathway analysis revealed that the chemokine signaling pathway was the most enriched among genes significantly associated with ACEs. For a subset of six ACEs-associated sites, we were able to estimate their heritability. The highest h^2 of 22% out of six sites indicated that ACEs-associated DNAm sites were mostly driven by nongenetic factors. Behavioral stressors can manifest on a biological level through DNA methylation and confer changes to the functionality of molecular pathways. In addition, such stress-related epigenetic modifications are shared by mothers and young children through commonly perceived adverse experiences.

129 | Genome Partitioning and Dimension Reduction Strategies for Multi-SNP Association Analysis of Genome-sequencing Data Utilizing Linkage Disequilibrium Structure

Yun Joo Yoo[1,2], Sun Ah Kim[1], Shelley B. Bull[3,4]

[1] Department of Mathematics Education, Seoul National University, Seoul, South Korea; [2] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea; [3] Prosserman Centre for Health Research, The Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Canada; [4] Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

In genetic association analysis using high-density genome sequencing data based on multi-SNP regressions, researchers should determine analysis units that guarantee good efficiency and interpretable results and adopt a strategy to deal properly with multicollinearity. We propose comprehensive strategies to partition the genome into blocks for analysis and for treatment of multicollinearity. For genome partitioning, we develop the GPART algorithm to specify big LD blocks and partition genomes into analysis units considering LD blocks and gene regions. This algorithm produces analysis blocks of average size of 40∼70 SNPs when the maximum block size is set to 100 to 200. To treat multicollinearity, we reduce sequencing data dimension by replacing collinear variables with principal components. In this method, the collinear SNP variables producing a high variance inflation factor (VIF) over a threshold value (here, 10) are selected to form a group with the other dependent SNPs. The principal component (PC) with the highest variance is taken to replace the group of multiple SNPs. By the dimension reduction strategies, the number of SNP variables in each analysis unit is reduced by 6∼70%. Empirical evaluation of multi-SNP regression-based global tests shows that dimension reduction using VIF and PC can prevent computational problems and improve power.

130 | The French Exome (FREX) Project: A Population-based Panel of Exomes to Help Filter Out Common Local Variants

Emmanuelle Genin[1], Richard Redon[2], Jean-François Deleuze[3], Dominique Campion[4], Jean-Charles Lambert[5], Jean-François Dartigues[6] for the FREX Consortium

[1]Inserm UMR-1078, CHRU Brest, Université de Brest, Brest, France; [2]Inserm UMR-1087 / CNRS UMR 6291, l'institut du thorax, Université de Nantes, Nantes, France; [3]Centre National de Génotypage, CEA, Evry, France; [4]Inserm UMR-1079, Faculté de Médecine, Université de Rouen Normandie, Rouen, France; [5]Inserm UMR-1167, Institut Pasteur, Lille, France; [6]Inserm UMR-1219, Université de Bordeaux, Bordeaux, France

High-throughput sequencing technologies enable the characterization of all the genetic variations in the exomes of individuals and the discovery of novel variants involved in monogenic and complex diseases. The study of these two kinds of diseases requires different methods and raises specific problems with a major challenge of understanding the functional role of the identified variants to filter out the neutral ones. For this purpose, investigators often rely on the observed frequencies of variants in public databases. However, these public databases are not well representatives of the different ancestries in Europe and this is true in particular for France.

Indeed, comparing the exomes of 573 individuals sampled in 6 different regions of France to the frequency data available in GnomAD, we identified several variants with significant allele frequency differences. We also found that, compared to the different public panels, the "French exomes" allow a more efficient filtering of variants leading to a substantial reduction in the number of candidate variants retained for validation.

Allele frequency differences are also detected within France between the different geographic regions, pointing out, for example, to some variants that could have been under positive selection. At this fine geographical scale, we were able to detect differences in the rare variant burden of several genes that could have consequences for the setting-up of rare variant association studies.

Population-specific exome panels are clearly useful to avoid false positives findings due to population structure and to allow a better understanding of genetic diversity at fine geographic scales.

131 | Detection of Deletions or Excess Homozygosity Associated with Head and Neck Cancer in a Whole-Genome Case–Control Study

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

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

Squamous cell cancer of the head and neck, which includes cancers of the oral cavity, pharynx (excluding nasopharynx), or larynx, is the sixth most common malignancy worldwide. Deletion copy number variations of DNA sequences are abundant and ubiquitous in the human genome and represent a variant class that is often associated with disease. Here, we performed whole-genome studies of association of deletions or excess homozygosity with head and neck cancer using 733,202 SNPs. We used the method that is based on the extension of approach that we previously developed and is based on logistic regression frame work, permitting to adjust for population structure. We tested each contiguous SNP locus between the 1154 cases and 1542 controls in batch 1 and 1031 cases and 2965 controls in batch 2. Our method is designed to detect statistically significant evidence of homozygosity at individual SNPs for SNP-by-SNP analysis and to combine the information among neighboring SNPs for cluster analysis. We found no SNPs of p-value$<5 \times 10^{-8}$ and 23 top SNPs (with p-value$<10^{-4}$) in batch 1 and 23 SNPs of p-value$<5 \times 10^{-8}$ and 117 top SNPs in batch 2. We performed a meta-analysis for combining the two SNP-by-SNP analysis results for batch 1 and batch 2 data sets and identified 5 SNPs of p-value$<5 \times 10^{-8}$. In the cluster analysis, we identified 1.1-kb segment of neighboring top SNPs on chromosome 3p in batch 1 and 3 distinct clusters on chromosomes 6q, 17p, and 18q in batch 2.

132 | Construction of an Exome-wide Risk Score for Schizophrenia Based on Weighted Burden Tests

David Curtis[1]

[1] *UCL Genetics Institute, UCL; Centre for Psychiatry, Barts and the London School of Medicine and Dentistry, London, United Kingdom*

Polygenic risk scores derived from GWAS SNPs have been widely used. Whole exome sequence studies yield information on very large numbers of variants, some of which are extremely rare. This means that a polygenic score derived from all variants cannot be obtained using standard methods. Variants from a whole exome sequencing study of schizophrenia were annotated and categorised as to whether a small number of attributes were applicable to them (e.g. in UTR, SIFT damaging, LOF). In addition, a number of gene sets of possible interest were specified, such as near GWAS hits, implicated in X-linked intellectual disability. For each subject, a set of risk scores was obtained aggregated across variant attributes and gene sets. Then, an optimisation scheme was used to obtain good-fitting weights for each attribute and each gene set which would produce a total score for each subject which maximally distinguished cases from controls. A five-way cross-validation procedure was applied. The scheme produced exome-wide scores which differed significantly (p = 0.001) between cases and controls. Only a small subset of parameters needed to be retained to obtain optimal performance. The approach described allows the construction of an exome-wide risk score. It should be possible to additionally incorporate polygenic risk scores from common SNPs and effects of rare variants with major effect.

133 | A Genome-wide Association Study in Fibromuscular Dysplasia Indicates Sexual Dimorphism in its Genetic Etiology

Siying Huang[1], Pierre-François Plouin[1,2,3], Michel Azizi[2,3,4], Pilar Galan[5], Xavier Jeunemaitre[1,2,6,7], Nabila Bouatia-Naji[1,2]

[1] *INSERM UMR970, Paris Cardiovascular Research Centre, Paris France;* [2] *Paris Descartes University, Faculty of Medicine, Paris France;* [3] *Assistance Publique-Hôpitaux De Paris, Department of Hypertension, Hôpital Européen Georges Pompidou, Paris, France;* [4] *INSERM, Clinical Investigation Center, CIC1418, Hôpital Europe en Georges Pompidou, Paris, France;* [5] *Nutritional Epidemiology Research Group, Sorbonne-Paris-Cité, UMR University of Paris 13/INSERM U-557/INRA U-1125/CNAM, Bobigny, France;* [6] *Assistance Publique-Hôpitaux De Paris, Referral Center for Rare Vascular Diseases, Hôpital Européen Georges Pompidou, Paris, France;* [7] *Assistance Publique-Hôpitaux De Paris, Department of Genetics, Hôpital Européen Georges Pompidou, Paris, France*

Fibromuscular dysplasia (FMD) is a non-atherosclerotic vascular disease, characterized with fibroplasia in tunica media, and is mainly found in renal, carotid and coronary arteries. FMD is often underdiagnosed and can lead to arterial stenosis, occlusion, aneurysm or artery dissection. It is a risk factor for hypertension, stroke and coronary artery dissection, and occurs predominantly in premenopausal females (80% to 90%). The etiology of FMD is inexplicit, though recent studies suggested a complex genetic pattern of inheritance. This study aims to discover the associations between common genetic variants and FMD status.

We performed a genome-wide association study in 614 French FMD patients and 1557 controls with European ancestries. Genotyped and imputed markers with imputation quality R^2>0.8 and minor allele frequency > 5% were included in the association tests. All models were adjusted for first three principal components assuming additive effect. We performed sex-adjusted and sex-stratified analyses and re-evaluated the markers with association test $p<10^{-5}$ under the Bayesian framework.

Approximately 5 million markers were included in the association tests. Thirty SNPs were found below 1% false discovery rate threshold (q) in the sex-adjusted analyses (q range: 4.19×10^{-31} - 0.009, $\log_{10}$(BF) range: 3.67 - 35.73). Seventeen of these SNPs were also found significant in females and showed higher magnitudes in ORs compared to sex-adjusted and males-only results. Interestingly, the reverse pattern in ORs was observed in males-only analyses, despite failing to reach genome-wide significance. These results suggest the sexual dimorphism in the genetic etiologies of FMD.

134 | Bayesian Model Averaging to Derive Multi-SNP Mendelian Randomization Instruments from Meta-GWAS Summary Statistics

Apostolos Gkatzionis[1], Stephen Burgess[1,2], Paul J. Newcombe[1]

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

Mendelian randomization provides a framework for answering questions of causality through the use of genetic proxies of a trait of interest. Recently, there has been a growing trend of Mendelian randomisation analyses based on summary data from large consortia meta-GWAS (Genome-Wide Association Studies). These studies seek to leverage the power harnessed by meta-analysis of large numbers of individuals but are often complicated by the availability of summarised data only.

The JAM algorithm (Joint Analysis of Marginal summary statistics, Newcombe, Conti and Richardson, 2016) is a recently proposed algorithm for identifying genetic variants associated with a specific trait from summary data. The algorithm facilitates Bayesian model selection and model averaging via a Reversible Jump MCMC procedure. Advantages of JAM are that its implementation accounts for genetic correlations despite only requiring access to the marginal one-at-a-time summary statistics and scalability to build multi-SNP instruments averaged over large numbers of variants.

In this talk, we will outline and discuss how JAM can be used in the context of Mendelian randomization based on meta-GWAS summary data. The algorithm can help identify those genetic variants that are more strongly associated with the trait studied, and construct a correlation-adjusted multi-SNP allele score from summarised data. The use of Bayesian model averaging can better reflect model uncertainty associated with meta-GWAS results, compared to traditional Mendelian randomization methods which require a fixed set of variants. The performance of the approach will be illustrated in simulated data, and some possible extensions will be discussed.

135 | Estimating the Heritability of Gene-Environment Interactions

Vincent Laville[1], Yun Ju Sung[2], Mike Province[3], Jim Gauderman[4], Daniel Chasman[5], DC Rao[2], Hugues Aschard[1], on behalf of the CHARGE Gene-Lifestyle Interactions Working Group

[1] Département de Génomes et Génétique, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France; [2] Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, United States of America; [3] Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, United States of America; [4] Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America; [5] Division of Preventive Medicine, Brigham and Women's Hospital, Boston Massachusetts, United States of America

Several genome-wide gene-by-environment (G × E) interaction analyses have been published with the aim of identifying single genetic variants interacting with environmental exposure in human traits and diseases. However little has been done to estimate the overall contribution of G × E interactions to heritability. For mathematical convenience, contribution of interaction effects to phenotypic variance is commonly estimated using an orthogonalized model where G × E contribution corresponds to the phenotypic variance explained on top of marginal genetic effects. In this model the genetic additive variance (whose standardized value is the narrow sense heritability) is measured at the mean of E, while G × E heritability (h^2_X) captures changes in heritability that would occur with changes in the exposure variance. We propose an alternative strategy which consists of estimating the G × E heritability (h^2_{INT}) as the difference between the in-sample genetic additive variance and the expected genetic variance at a baseline value of the exposure (e.g. when the exposure is absent). We first demonstrate analytically that $h^2_X \leq h^2_{INT}$ and further developed an approach to approximate both parameters (h^2_X and h^2_{INT}) using only genome-wide G × E analysis summary statistics and the LD score regression framework. We then demonstrate the validity and performance of our estimators using both simulated and real summary data from gene-smoking interactions in blood pressure in ~70,000 individuals from the CHARGE Gene-Lifestyle Interactions Working Group. Finally, we highlight the value of our estimator h^2_{INT} in the context of clinical relevance and public health. Supported by a grant (HL118305) from the NHLBI.

136 | Interaction of Genetic Variants with Secondhand Smoke Exposure in Early Life on Time-to-Asthma Onset

Pierre-Emmanuel Sugier[1,2,3], Chloé Sarnowski[1,2], Raquel Granell[4], Debbie Jarvis[5,6], Markus J. Ege[7,8], Catherine Laprise[9], Erika von Mutius[7,8], Marie-Hélène Dizier[1,2], A. John Henderson4, Manolis Kogevinas[10,11,12,13], Florence Demenais[1,2], Emmanuelle Bouzigon[1,2]

[1] Inserm, UMR-946, Genetic Variation and Human Diseases Unit, Paris, France; [2] Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France; [3] Université Pierre et Marie Curie, Paris, France; [4] School of Social and Community Medicine, University of Bristol, United Kingdom; [5] Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College, London, United Kingdom; [6] MRC-PHE Centre for Environment & Health, London, United Kingdom; [7] Dr von Hauner Children's Hospital, Ludwig Maximilian University, Munich, Germany; [8] Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research, Munich, Germany; [9] Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Quebec City, Canada; [10] Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [11] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; [12] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain; [13] Universitat Pompeu Fabra, Barcelona, Spain

Asthma is a complex and heterogeneous disease that results from many genetic and environmental factors, in which age-of-onset plays an important role. Secondhand smoke exposure in early life (ETS) is a known risk factor for both childhood-onset and late-onset asthma. To identify genetic variants interacting with ETS exposure on asthma occurrence, we conducted a meta-analysis of five genome-wide interaction studies (GEWIS) of time-to-asthma onset (TAO) including both asthmatics and non-asthmatics (totaling 3,643 exposed (ETS^+) and 5,275 non-exposed (ETS^-) subjects of European ancestry) by using survival analysis techniques. Following a previous genome-wide analysis which examined the effect of individual SNPs in presence of interaction with ETS, the current study focused on the SNPxETS interaction test which allows detecting SNPs with a small (or no) marginal effect interacting with ETS. A pathway analysis based on the gene-set enrichment analysis (GSEA) approach, using the Gene Ontology (GO) database, was then applied to the GEWIS outcomes. We detected 33 SNPs belonging to 11 independent loci showing suggestive SNP × ETS interaction ($p<10^{-5}$). The most significant interaction signals belonged to three loci: 13q21 (*KLHL1*, $p = 9.8 \times 10^{-7}$), 16p13 (intergenic region, $p = 6.7 \times 10^{-7}$) and 19q13 (*ZNF761*, $P = 10^{-6}$). Twelve GO categories were enriched in genes interacting with ETS on TAO (FDR≤5%); the most significant GOs (FDR<1%) were related to defense response to bacteria, oxidative stress, and lipid metabolism. Further analysis investigating enrichment of loci interacting with ETS in cis-regulatory elements and co-localization with loci detected by epigenome-wide analysis of ETS in early life is underway.

139 | Genome-wide Association Study Combining UK Biobank and GASP Consortium Highlights Novel Loci Associated with Moderate-Severe Asthma

Nick Shrine[1], María Soler Artigas[1], Michael A. Portelli[2], GASP Consortium[2], Martin D. Tobin[1], Ian P. Hall (2), Christopher E. Brightling[3], Louise V. Wain[1], Ian Sayers[2]

[1] Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [2] Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom; [3] Institute for Lung Health, University of Leicester, Glenfield Hospital, Leicester, United Kingdom

The genetic architecture of asthma to date has been described by the discovery of around 20 loci from genome-wide association studies (GWAS), primarily with cases covering mild-to-moderate asthma. We hypothesised that moderate-to-severe asthma, which is currently difficult to treat, may have a specific genetic architecture, however, there have not been large GWAS of moderate-to-severe asthma.

Accordingly, we selected 5,135 European ancestry moderate-severe asthma cases (British Thoracic Society criteria 3 or above) and 25,675 controls free from lung disease, allergic rhinitis, and atopic dermatitis, from UK Biobank and the Genetics of Asthma Severity & Phenotypes (GASP) cohort (cases only). We tested 33,771,858 SNPs and indels genome-wide (imputation against combined UK10K and 1000 genomes phase 3 panels) for association with moderate-severe asthma.

We identified 23 independent signals associated with moderate-to-severe asthma ($p < 5 \times 10^{-8}$), including novels signals in or near *GATA3, RIC1, ZNF652, RPAP3, and MUC5AC,* highlighting regions that harbour variants that affect gene expression or genes that play a role in respiratory disease and immune response. Previously described asthma loci where replicated including signals in or near *D2HGDH, CD247, HLA-DQB1, HLA-DQA1, TSLP/WDR36S, IL1RL1/IL18R1, CLEC16A, GATA3, IL33, SMAD3, SLC22A5/IL13, C11orf30, ZBTB10, IKZF3-ORMDL3, and IKZF4.*

This largest GWAS of moderate-severe asthma to date and highlights novel loci that may provide new biological insights relevant to treatment of severe asthma.

140 | Small Posterior Fossa in Chiari Malformation Affected Families is Significantly Linked to 1q43-44 and 12q23

Anthony M. Musolf[1], Sze Chun Winson Ho[2], Kyle A. Long[1], Ping Zhuang[2,3], Haiming Sun[1,4], Bilal A. Moiz[1], Elena G. Mendelevich[5], Enver I. Bogdanov[5], Joan E. Bailey-Wilson[1], John D. Heiss[2]

[1] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2] Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, United States of America; [3] Neuro-Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; [4] Laboratory of Medical Genetics, Harbin Medical University, Harbin, China; [5] Department of Neurology & Rehab, Kazan State Medical University, Kazan, Tatarstan, Russia

The posterior fossa is a cranial cavity at the base of the skull. When it is small, it causes the cerebellum and brainstem to be pushed downward, resulting in a Chiari malformation. Chiari malformations cause neck pain, balance issues, decreased motor skills and headaches in those affected.

We have posterior fossa measurements and whole exome sequence data on individuals from 9 extended families from the United States and Russia that have a family history of Chiari malformations. We performed parametric linkage analyses using an autosomal dominant inheritance model with a disease allele frequency of 0.01 and penetrance for small posterior fossa of 0.8 for carriers and 0.1 for non-carriers. Single variant two-point linkage analysis and collapsed haplotype pattern (CHP) two-point linkage analysis were performed. CHP two-point linkage analysis used rare variants (MAF≤0.05) to create multi-allelic pseudo-markers that correspond to a gene or section of a gene. This gene-based test can improve power in the presence of allelic heterogeneity.

Our results found a genome-wide significant linkage on chromosome 1q43-44 (HLOD = 3.5) and 12q23 (HLOD = 3.3) in both sets of linkage analyses. Most interesting was that both signals were driven by a single (different) family. Both regions contain several linked exonic variants including rare variants located in good candidate genes. In conclusion, we have located two significantly linked regions for small posterior fossa that are driven by linked variants in 1 family each and are potentially causal. Further laboratory work is needed to confirm these candidate genes.

141 | Genome-wide Association Study of Susceptibility to Mild Malaria in Two Cohorts of Young Beninese Children

Jacqueline Milet[1], Pierre Luisi[2], Audrey Sabbagh[1], Ibrahim Sadissou[3], Paulin Sonon[3], Nadia Domingo[4], David Courtin[1], Achille Massougbodji[4], André Garcia[4,5], Hervé Perdry[6]

[1] UMR216 Mère et enfant face aux infections tropicales, Institut de Recherche pour le Développement, Faculté de Pharmacie, Université Paris Descartes, COMUE Sorbonne Paris, France; [2] Laboratorio de Genomica Biomedica y Evolucion, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; [3] Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil; [4] Centre d'etude et de recherche sur le paludisme associé à la grossesse et

l'enfance, Faculté des Sciences de la Santé, Cotonou, Bénin; [5] UMR216 Mère et enfant face aux infections tropicales, Institut de Recherche pour le Développement, Faculté des Sciences de la Santé et Institut des Sciences Biomédicales Appliquées, Cotonou, Bénin; [6] Université Paris-Saclay, Centre de recherche en Epidémiologie et Santé des Populations, Institut National de la Santé et de la Recherche Médicale, Villejuif, France

Malaria remains a major worldwide public health problem in spite of numerous prevention and control efforts in recent years with ∼ 429,000 deaths in 2015, mostly in sub-Saharan Africa. The key role of genetics factors in disease susceptibility and progression is admitted, but molecular basis of susceptibility/resistance to malaria has not been elucidated. In the last few years, several genome-wide association studies (GWAS) have been published on severe malaria. This approach, based on high density genotyping arrays, besides replicating known associations (HBB, ABO, G6PD), revealed new genes (ATP2B4, FREM3/GYP genes cluster).

Until now, there was no GWAS of non-severe malaria, mainly because of difficulties inherent to field studies for large cohorts. We present the result of a GWAS of mild malaria attacks, based on two cohorts of children closely followed-up during their first year of life in South Benin, and genotyped with the HumanOmni5 array. After quality control, 800 children were available for analysis.

Two different phenotypes were defined to assess the susceptibility of children to clinical malaria: the total number of attacks recorded in the whole follow-up (classic phenotype in association studies on mild malaria) and the recurrence of attacks. Using entomological, environmental, and behavioral informations, both phenotypes were adjusted on individual exposure to malaria vector and on major covariates, using respectively a negative binomial regression and a mixed-effect Cox model. This last model allowed to incorporate precisely the time-dependent exposure. The GWAS was performed on the adjusted phenotypes, with a mixed model for accounting population structure.

142 | Comparison for secondary phenotype analysis in ascertained family studies: applications to the social anxiety disorder study

Renaud Tissier[1], Anita Harrewijn[1], Jeanine Houwing-Duistermatt[2]

[1] Department of Developmental and Educational Psychology, Leiden University, Leiden, The Netherlands; [2] Department of Statistics, University of Leeds, Leeds, United Kingdom

This work is motivated by a study which investigates the relationship between EEG measurements and social anxiety disorder (SAD) and the heritability of these traits. EEG data are available for nine large families with at least two cases with SAD. The researchers used SOLAR for data analysis by assigning the two affected family members the proband status. However, the design is multiple case family and therefore SOLAR might not be appropriate for analysis. We have recently developed a method which jointly model the primary and secondary phenotype and obtains parameter estimates by maximizing the retrospective likelihood and provides unbiased estimates when modelling secondary phenotypes in multiple case families.

We will present results of extensive simulation studies to compare the performance of various statistical methods for ascertainment adjustments, i.e. a naïve approach, SOLAR (Sequential Oligogenic Linkage Analysis Routines) and our recently developed secondary phenotype approach. The methods are compared in terms of bias, efficiency and computation time for various scenarios for different types of ascertainment. Finally, we present the results of the analysis of EEG measures in families with SAD.

SOLAR appears to be more efficient in terms of computing time and can handle correctly one-case ascertainment. However, for the multiple-cases ascertainment and the use of the joint model under retrospective likelihood is needed. Results on the social anxiety disorder show that our approach is better identifying genetic variation than SOLAR and, therefore, is the proper method to estimate heritability and identify endophenotypes.

143 | A General Framework for Variable Selection in Linear Mixed Models with Applications to Genetic Studies with Structured Populations

Sahir R. Bhatnagar[1,2,5], Karim Oualkacha[3], Yi Yang[4], Celia M.T. Greenwood[1,2,5]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada; [2] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada; [3] Département de Mathématiques, Université de Québec À Montréal, Montreal, Canada; [4] Department of Mathematics and Statistics, McGill University, Montreal, Canada; [5] Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Canada

Complex traits are thought to be influenced by a combination of environmental factors and rare and common genetic variants. However, detection of such multivariate associations can be compromised by low statistical power and confounding by population structure. Linear mixed effect models (LMM) can account for correlations due to relatedness but are not applicable in high-dimensional (HD) settings where the number of predictors greatly exceeds the number of samples. False negatives can result from two-stage approaches, where the residuals estimated from a null model adjusted for the subjects' relationship structure are subsequently used as the response in a standard penalized regression model. To overcome these challenges, we develop a general penalized LMM framework that simultaneously selects and estimates variables for structured populations in one step. Our method can accommodate several sparsity-inducing penalties such as the lasso and elastic net, and also readily handles prior annotation information in the form of weights. Our algorithm is computationally efficient, scales to HD settings and we mathematically prove that it converges to a stationary point. Through simulations, we show that when there are several correlated causal variants with small effects, our method has better power over the two-stage approach. We apply our method to identify SNPs that predict blood pressure in 20 large Mexican American pedigrees from the Genetic Analysis Workshop 18 data. This approach can also be used to generate genetic risk scores that can be useful for risk stratification and clinical decision making. Our algorithms are available in an R package (https://github.com/sahirbhatnagar/penfam).

144 | Demystifying Causal Effect Heterogeneity of Composite Risk-Factors in Multi-Instrument Mendelian Randomisation Studies Using a Novel Bayesian Feature Selection Algorithm

Christopher N. Foley[1], Steven Burgess[1,2]

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

Mendelian Randomisation (MR) uses single nucleotide variants as ’naturally-randomised’ instrumental variables to assess a causal link between a risk-factor and an outcome (e.g. disease). MR has thrived over recent years with the expanding success of genome-wide association studies (GWAS). The wide-spread availability of summary association statistics from hundreds of GWAS, which have identified ∼32,000 trait associated loci, have allowed for MR analyses which aggregate effect estimates across hundreds of instruments and multiple risk-factors. However, a key assumption of MR is the exclusion restriction: a genetic instrument only acts on the outcome through the risk-factor. Violations of this, when a locus is associated with multiple traits (pleiotropy), is common in human genetics; leading to bias of causal effect estimates and effect heterogeneity in aggregated analyses.

We focus on the assessment of composite risk-factors (CRFs), i.e. risk-factors that are composed of several lower-level risk-factors (LLRFs), e.g. BMI is a CRF computed from two LLRFs: body weight and height. If a LLRF is associated with the outcome directly as well as being mediated through the CRF, the MR estimates suffer symptoms of pleiotropy. We take advantage of this detail, however, and by blending a Bayesian Model Averaging technique with a novel feature selection algorithm we: 1) identify the number of latent LLRFs 2) match each instrument to a LLRF group and either 3a) estimate a CRF causal effect or 3b) interpret results when all instruments are invalid. We illustrate this hypothesis-free approach using BMI and inflammation CRF examples.

145 | Evidence of Genetic Predisposition for Metabolically Healthy Obesity and Metabolically Obese Normal Weight

Lam O. Huang[1], Ruth J.F. Loos[2,3,4], Tuomas O. Kilpeläinen[1,5]

[1] Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [2] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [3] The Genetics of Obesity and Related Metabolic Traits Program, The Icahn School of Medicine at Mount Sinai, New York, New York, United States of America; [4] The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York United States of America; [5] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America

Obesity has evolved into a global pandemic, which constitutes a major threat to public health. The majority of obesity-related health care costs are due to cardiometabolic complications, such as insulin resistance, dyslipidemia, and hypertension, which are risk factors for type 2 diabetes and cardiovascular disease. However, many obese individuals, often called metabolically healthy obese (MHO), seem to be protected from these cardiometabolic complications. Conversely, there is a group of individuals who suffer from cardiometabolic complications despite being of normal weight; a condition termed metabolically obese normal weight (MONW). Recent large-scale genomic studies have provided evidence that a number of genetic variants show an association with increased adiposity but a favorable cardiometabolic profile, an indicator for the genetic basis of the MHO and MONW phenotypes. Many of these loci are located in or near effector genes that implicate pathways involved in adipogenesis, fat distribution, insulin signaling, and insulin resistance. We further carried out bivariate genome-wide meta-analyses on pairs of adiposity and cardiometabolic trait to confirm association pattern reminiscent of the MHO and MONW phenotypes. We are able to detect additional loci using this approach with improved power. The preliminary bivariate analysis on body fat percentage and high-density lipoprotein cholesterol (HDL-C) has identified more than 300 loci exhibiting the MHO feature. This will contribute to the understanding of the genetic aspects of the mechanisms that underpin MHO and MONW, which is crucial to define appropriate public health action points and to develop effective intervention measures.

146 | The SigMod Network Analysis Method Identifies Gene Modules for Cutaneous Melanoma and Nevus Count that Share Relevant Candidates

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

[1]INSERM, UMR-946, Genetic Variation and Human Diseases unit, Paris, France; [2]Université Paris Diderot, Sorbonne Paris Cité, Paris, France; [3]AP-HP, Hôpital Avicenne et Université Paris 13, Bobigny, France; [4]AP-HP, Hôpital Cochin et Université Paris Descartes, Paris, France; [5]McGill University, Montreal, Canada

The major risk factors for cutaneous melanoma (CM) include pigmentation phenotypes, number of melanocytic nevi (NMN) and a family history of melanoma. To date, genome-wide association studies (GWAS) have identified 20 loci for CM risk and four of these loci are also associated with NMN. However, many other loci remain to be discovered. To identify new genes that have a small marginal effect but may collectively influence CM and/or NMN, we conducted network-assisted analysis which integrates outcomes of GWAS and biological relationships between genes retrieved from the STRING database. We used SigMod, an efficient network analysis method we recently developed. SigMod was applied to GWAS data (1000 Genomes- imputed SNPs) from the French MELARISK (3,976 subjects) and NEVRISK (totaling 429 subjects with nevus count in 210 families ascertained through a proband with >50 nevi) collections. We identified a gene module of 228 genes with 564 interactions for CM and a gene module of 193 genes with 334 interactions for NMN. More than 98% of these genes were associated with CM or NMN at the nominal level. The two gene modules share 10 novel genes which represent relevant candidates for both melanoma and nevi. Notably, *GNA11* is known to be involved in uveal melanoma tumors, *MAPK10* is a member of the MAP kinase pathway which plays a major role in CM tumors while other genes (*ZFYVE16*, *TAL1, NCOA1*) are involved in regulatory mechanisms. Further functional characterization of these gene modules and replication of results are underway.

147 | LD Score Regression for Non-continuous Traits

Emily Slade[1], Jennifer Sinnott[2], Sander Canisius[3], Marjanka Schmidt[3], Sara Lindstrom[4,5], Peter Kraft[1,6]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America; [2]Department of Statistics, Ohio State University, Columbus, Ohio, United States of America; [3]Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, Netherlands; [4]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, United States of America; [5]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America; [6]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America

Linkage disequilibrium (LD) score regression has emerged as a popular tool for differentiating between confounding biases and polygenicity in genome-wide association studies (GWAS) and for estimating heritability due to measured and imputed variants. Although LD score regression was developed in the setting of linear regression with a continuous outcome, it is nonetheless applied to results when the outcome binary and the GWAS analysis is logistic regression. Further, there is increasing interest to apply LD score regression to GWAS results from survival analysis. When implementing LD score regression for a continuous trait, there is a straightforward link between the Wald chi-squared test statistics from linear regression and heritability. However, there is no direct link between the "heritability" from LD score regression applied to results from logistic regression or Cox proportional hazards regression and classical definitions of heritability (like heritability on the liability scale, as estimated from twin studies). Nonetheless, the "heritability" estimate from LD score regression is still informative in these settings: for logistic regression, it can be interpreted as the variance in the log odds ratio; for Cox proportional hazards regression, it can be interpreted as the variance in the log hazard ratio. These parameters have implications for risk prediction modeling. We derive these results mathematically, verify via simulation, and illustrate with applications to breast cancer incidence and survival. In summary, care must be taken when interpreting results from LD score regression when used with logistic or Cox proportional hazards models.

148 | Genome-wide Association Study Identifies Nine Novel Loci for Subclinical Atherosclerosis and Downstream Regulatory Effects in Tissues Affected by Atherosclerosis

Nora Franceschini[1], Claudia Giambartolomei[2] for the CHARGE and UCLEB Consortia

[1]Department of Epidemiology, University of North Carolina at Chapel Hill, North Carolina, United States of America; [2]Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America

Genome-wide association studies (GWAS) have identified four loci associated with carotid artery intima-media thickness (cIMT) and two with carotid plaque, which are measures of subclinical atherosclerosis associated with coronary heart disease and stroke. To further identify loci for subclinical atherosclerosis, we undertook GWAS meta-analyses using dense imputed 1000 Genomes data in up to 71,129 for cIMT and 41,145 individuals for carotid plaque and identified 9 new susceptibility loci at p-value$<5.0 \times 10^{-8}$. We integrated our GWAS results of cIMT/plaque with expression quantitative loci (eQTL) from six cardiovascular-relevant tissues of up to 600 patients with coronary heart disease. These analyses identified three loci (four genes: *ADAMTS9, LOXL4, CCDC71L/PRKAR2B*) with a posterior probability > 75% of sharing the same variant associated with cIMT or plaque genome-wide with eQTLs in tissues affected by atherosclerosis. *CCDC71L* had the highest posterior probability of

co-localization of carotid plaque with aorta eQTLs (93%), and cIMT with mammary artery eQTLs (95.6%). Our study provides insights into genes and tissue-specific regulatory mechanisms for atherosclerosis traits.

149 | Family-based Association Tests of Myopia Reveal a Potentially Hidden Association Signal Upstream of Two GABA Receptor Genes

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

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

Myopia is an eye condition in which distant objects appear out of focus. Within recent years, the incidence and prevalence of myopia have increased in most populations. We performed a family-based association test (FBAT) of myopia using Exome Chip genotyping (Illumina Human Exome v1.1 array plus 24,263 custom SNPs) in five family cohorts for a total of 1718 subjects in 261 multiplex myopia families.

Individuals in the families were defined as myopic if their average refractive error was ≤ -1 Diopter (D) and were considered unaffected if it was > 0.0 D. After quality control, there were ~127,000 polymorphic SNPs. FBAT analysis resulted in a significant association in a region upstream of two gamma-Aminobutryric Acid (GABA) receptor genes (GABRA6; GABRB2). GABA is a neurotransmitter that has been implicated in refractive development. The associated SNP, rs1373602, is not found in the Genotype-Tissue Expression (*GTEx*) project, but a nearby SNP, rs62381591, has been identified as an expression quantitative trait locus for the GABRA6 gene. As the significant variant is common, we wondered why the larger, population-based association studies of Myopia have not detected association in this region. We found that this variant is not in high linkage disequilibrium with any other variants in our dataset and is indicated as triallelic in the 1000 genomes dataset (although it was biallelic in our dataset). Potentially, this region harbors a missed signal that is tagged by a variant that is filtered out before GWAS analysis.

150 | Revisiting Broad-sense Heritability Estimation in a Population Isolate

Anthony F. Herzig[1,2], Teresa Nutile[3], Marina Ciullo[3,4], Hervé Perdry[5], Anne-Louise Leutenegger[1,2]

[1] Université Paris-Diderot, Sorbonne Paris Cité, UMR946, F-75010 Paris, France; [2] Inserm, UMR 946, Genetic variation and Human diseases, F-75010 Paris, France; [3] Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy; [4] IRCCS Neuromed, Pozzilli, Isernia, Italy; [5] Université Paris-Saclay and CESP, Villejuif, France

Contradictory results on the existence of a significant non-additive (dominant) genetic variance for certain complex traits were obtained in studies of either isolated or outbred populations. We propose to investigate these discrepancies by comparing methods for estimating genetic variance components in a population isolate.

Additive and dominant genetic effects are estimated using a linear mixed model in which the variance of the phenotypes is modelled using two matrices: the genetic relatedness matrix (GRM) of pairwise kinship coefficients and the dominance matrix (DM) of coefficients estimating the probability of two individuals sharing two alleles identical by descent. A population isolate would intuitively offer an ideal setting for estimating dominance variance due to the close relatedness between individuals.

Our study relies on phenotypes and genotypes from the known population isolate of Cilento in South Italy as well as on simulated data based on the structure of Cilento. We test the effects of the choice of method to estimate the GRM and DM matrices on heritability estimation. While in studies on unrelated individuals, the GRM and DM have to be estimated by the method of moments, in population isolate, one can alternatively perform this estimation using pedigree information either alone or in conjuncture with the observed genotypes. Furthermore, we propose to investigate the benefits of estimating heritability from next-generation sequencing data and of incorporating genotype uncertainty into the estimations.

151 | Exploring the Use of Fuzzy Clustering Approaches to Classify HbA1c Associated Signals into Glycaemic and/or Erythrocytic Pathways

Ji Chen[1] on behalf of the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators

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

Glycated haemoglobin (HbA1c) is widely used to diagnose Type 2 Diabetes. Previous studies have identified 18 loci affecting HbA1c through pathways related to glycaemia ("glycaemic loci") or erythrocyte biology ("erythrocytic loci"). A recent large trans-ethnic meta-analysis performed by MAGIC has identified 101 loci associated with HbA1c in Europeans. Our aim was to classify the European HbA1c signals as influencing HbA1c through glycaemic and/or erythrocytic pathways using a fuzzy clustering approach which has an advantage over traditional hierarchical clustering by allowing loci to belong to multiple clusters. To classify the loci, we used summary statistics from European genome-wide association studies (GWAS) of three glycaemic, twelve erythrocytic and four iron traits from which we generated allele variance normalized SNP effect sizes (genetic impacts) as inputs for the clustering. The 19 traits were clustered first and showed two clear clusters of glycaemic and erythrocytic traits with iron traits contained within the erythrocytic cluster. The loci were then clustered into three different clusters, which were categorised as glycaemic, erythrocytic or unknown by a membership coefficient weighted summation of impacts and verification using data available for HbA1c adjusted for fasting glucose. Fifty-three of 86 loci without missing lookup results were classified into 18 glycaemic, 17 erythrocytic and 18 unknown loci with membership coefficient > 0.9. The glycaemic locus *G6PC2*, erythrocytic locus *ERAL1,* and unknown locus *SYF2* are consistent with previous studies. One of the 33 loci belonging to multiple clusters, *ABO*, has previously been associated with erythrocytic traits and fasting glucose.

152 | On the Gain of Mega-imputation and Mega-analysis compared to Meta-imputation and Meta-analysis Exemplified on Genetics of Age-related Macular Degeneration

Mathias Gorski[1], Thomas Winkler[1], Iris Heid[1]

[1] *Department of Genetic Epidemiology, University Regensburg, Regensburg, Germany*

While most genome-wide association analyses utilize summary statistics from numerous studies, the collection of individual participant data (IPD) from multiple studies or large multi-site studies is an alternative approach that increases the options of statistical analyses. However, little is known as to how to impute and analyze large multi-study IPD: by study or overall. We exemplify pros and cons of by-study-imputation (by-study-phasing/by-study-imputing) versus overall-imputation (overall-phasing/overall-imputing) combined with different models of association analyses (model-I: ignoring study membership, model-II: study-specific background risk, model-III: study-specific accounting for other covariates, model-IV: study-specific modelling) on our large IPD on age-related macular degeneration (Fritsche et al, 2017, 16,144 cases, 17,832 controls, 439,350 genotyped variants).

We yielded 360,000 additional well imputed variants in overall-imputed data compared to by-study-imputed data (RSQ>0.8, 9,821,296 vs. 9,460,246), which are mostly rare (MAF<1%, 2,988,839 vs. 2,636,518). Applying the various models on by-study-imputed data, we found inflation of statistics and 40 additional genome-wide significant loci ($p<5 \times 10^{-8}$) applying model-I or II that vanished with model-III or IV. As all these 40 additional loci reside at the tails of chromosomes, these are considered false positives. We did not find such artefacts using overall-imputed data for any of the models, but slightly more significant loci compared to by-study-imputed data (e.g. 29 loci versus 27 loci using model-IV).

We followed these findings by extending our imputation approaches (overall-phasing/by-study-imputing and vice versa), type I error and analytical power computations. Our investigations highlight the importance of how to choose the statistical model and how this depends on the imputation approach.

153 | TRQUANT: A New Implementation of Tiled Regression for Quantitative Traits

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

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

Tiled regression is a method of fitting a regression model designed for the analysis of high-dimensional genomic data; the method selects a set of genetic predictors that jointly and independently contribute to trait variation. In this method, variable selection is applied in stages starting with small genome segments (tiles), and variables selected from different tiles are considered together for further selection at chromosome and genome-wide stages. This procedure was implemented in TRAP for different types of traits and variable selection methods. A new implementation of tiled regression, TRQUANT, focuses on quantitative traits with stepwise selection methods. New options are included for analyzing unrelated and family data, as well as changes intended to speed up analysis. One new feature is an option to handle missing values in a way that may use more of the available data. Stepwise regression typically requires that only observations complete on all potential predictors be considered throughout the process. An alternative is to require a common set of observations only for the two models being compared in one step, when adding/removing one predictor to/from the model, ignoring missing data patterns in unused potential predictors. Another change is in the handling of family relationships. Instead of assuming constant correlation within clusters, a flexible relationship matrix is used, along with a general least squares approximation for a linear mixed model, similar to that used in EMMAX and QTLRel, extended to stepwise selection. TRAP and TRQUANT are compared with respect to speed and fitted models.

154 | Han Chinese Families Show Significant Linkage for Myopia on 10q26 and Suggestive Linkage on 9q33

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

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

Myopia is caused by an overgrowth of the eye which causes light to focus in front of the retina, leading to blurry vision. We studied 34 Han Chinese families ascertained with an apparent dominant inheritance of myopia to search for linkage between genomic variants (Illumina Human Exome v1.1 array plus 24,263 custom SNPs) and the disease. Affection status was based on mean spherical equivalent in Diopters (D): affected ($\leq$-1D), unaffected ($\geq$0 D) or unknown ($<$0D, > -1 D). Three types of parametric linkage analyses were performed: single variant two-point, multipoint, and collapsed haplotype pattern variant linkage (CHP). CHP creates a multi-allelic pseudomarker that corresponds to a genomic region from multiple rare variants. This raises information content. Two-point linkage analysis is performed using CHP markers.

CHP linkage analysis identified a genome-wide significant heterogeneity LOD score (HLOD = 3.73) at 10q26.13, centered on *TACC2*. This pseudomarker consisted of several rare exonic SNPs from the *TACC2* gene. CHP analysis also found 6 more suggestive linkages in 10q24.2-26.2. Single variant two-point identified 34 suggestive loci on 10q24-26, while multipoint identified 8 suggestive loci in 10q26.11-13. Many of the suggestive markers in these analyses were found in *HTRA1*, a known age-related macular degeneration gene. Several other promising candidate genes, such as *BAG3* and *DOCK1*, are also present in this region. Multipoint analysis also identified a highly suggestive region at 9q33.1 (HLOD>3.2). This region includes *TLR4*, a gene known to interact with alpha-crystallin in the retina. Targeted sequencing for both regions is planned to elucidate the causal variants.

155 | Block-wise Descent Algorithms for Group Variable-Selection in Quantile Regression

Karim Oualkacha[1], Mohamed Ouhourane[1], Yi Yang[2], Celia M.T. Greenwood[3]

[1] Department of Mathematics, Université du Québec à Montréal, (UQAM), Montreal, Canada; [2] Department of Mathematics and Statistics, McGill University, Montreal, Canada; [3] Lady Davis Institute for Medical Research, Jewish General Hospital and McGill University, Departments of Oncology, Epidemiology, Biostatistics & Occupational Health, Human Genetics and Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Canada

The data generated from high-throughput experiments/approaches, such as '-omics' data, are often complex and high dimensional. The challenge for statisticians is to extract relevant biological insight from the massive volume of data that these approaches produce. Quantile regression (QR) models are attractive in several fields due to their capability to estimate the conditional quantile in different locations of the conditional distribution of the outcome/phenotype, which provides a more complete picture of the phenotype conditional distribution. QR can cover several areas of '-omics' data including genetic association studies, gene expression data, methylation data etc…

In this work, we consider the problem of selecting grouped variables in a high-dimensional linear quantile regression model. We introduce a Group variable-Selection framework for Quantile Regression (GSQR) with most relevant group-penalties: the group lasso penalty, the group non-convex penalties (MCP and SCAD), local approximations of the non-convex penalties and the sparse group lasso penalty. We propose a smooth block-wise descent algorithm combined with a maximization-minimization algorithm for updating each group variable simply and efficiently.

We illustrate the proposed methodology performance using a detailed simulation study in high dimensional settings. The numerical results show that our implementation reaches consistent sparse group solutions. In a context of QTL genetic association, we are currently analyzing real data from the Genetic Analysis Workshop 18 with individuals selected from 20 Mexican American families and a dense set of single-nucleotide polymorphisms in 959 individuals in these families, in an effort to identify genomic regions (groups) regulating complex QTL traits such as systolic blood pressure.

156 | Exome CNV Overlapping (ECO): an Integrative Copy Number Variation Caller for Exome Sequencing

Peng Zhang[1], Hua Ling[1], Elizabeth Pugh[1], Kim Doheny[1]

[1] Center for Inherited Disease Research (CIDR), Johns Hopkins Genomics, Institute of Genetic Medicine, The Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

Due to the uneven distribution of reads and the sparse nature of target regions for whole exome sequencing (WES), calling copy number variations (CNVs) has been a challenge and most of existing programs can only use read counts as inputs and calls often vary between programs. For example, the numbers of CNVs called were 174,183 (ExomeDepth), 2,670 (HMZDelFinder), and 38,952 (XHMM), respectively, for ~1600 WES samples from the Centers for Mendelian Genomics (CMG) project. As part of the validation process, we found that some confirmed causal CNVs were called by multiple programs while others were not. In addition, each program often requires different input files and its output format often varies with different breakpoints for the CNV calls, which makes it difficult to compare and summarize results across programs.

We present here a practical pipeline that integrates multiple CNV calling programs and generates one combined VCF-like report with merged calls and annotations. It incorporated three prevalent CNV calling programs (ExomDepth [Plagnol et al. 2012], CANOES [Backenroth et al. 2014], and CODEX [Jiang et al. 2015]) with the ability to incorporate results from two additional programs (XHMM [Fromer and Purcell 2014] and HMZDelFinder [Gambin et al. 2017]). In addition, our pipeline: 1) Generates read counts only once, either from BAM or CRAM; 2) Runs the three methods in parallel; 3) Merges calls by a user-defined overlap percentage and a size threshold; 4) Provides annotation such as gene names in the regions and call frequencies.

157 | A Comparison of Methods for Identification of Genetic Variants Related to Age-of-Onset of Cystic Fibrosis Related Diabetes

Hua Ling[1], Peng Zhang[1], Elizabeth W. Pugh[1], Melis Atalar[2], Scott M. Blackman[3]

[1] Center for Inherited Disease Research, Johns Hopkins University, Baltimore, Maryland, United States of America; [2] The McKusick-Nathan Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America; [3] Division of Pediatric Endocrinology, Johns Hopkins University, Baltimore, Maryland, United States of America

Cystic fibrosis (CF) is a monogenic disease that affects more than 80,000 people worldwide and causes life-limiting lung disease and pancreatic dysfunction. Diabetes (CF-related diabetes or CFRD) is the most common extrapulmonary complication of CF and affects >40% people with CF by adulthood with a broad range of age of onset. This variation in CFRD risk has been shown to be heritable, and Blackman et al. (2013) identified five loci associated with CFRD onset, analyzed as a survival trait while excluding related individuals in a total of 3,059 samples. To better account for relatedness in survival analyses, we investigate different strategies using a family subset of the above data, the CF Twin and Sibling Study, which includes 396 samples from 288 small families (siblings and half siblings) genotyped on the Illumina 610-Quad. Our preliminary analyses show using mixed-effect Cox models, either with family-specific random intercept or with correlated random intercept using a kinship coefficient matrix, yield slightly better control of inflation of type 1 error compared to Cox proportional hazard model including related individuals ($\lambda = 1.00$ and 1.096 for family-specific and correlated random intercept respectively vs 1.12 for Cox proportional hazard model with related individuals), but not compared to maximally unrelated subset analysis ($\lambda = 1.03$). Analyses of PC-adjusted Martingale residuals in linear mixed model with relatedness as random effects yield similar results ($\lambda = 1.096$). Further analyses will be performed to better understand the difference in results between models. Supported by CF Foundation.

158 | Adjusting Family Aggregation and Population Stratification via Genetic Relationship Matrix in Association Analysis of Genetic Variants and a Binary Trait

Biqi Wang[1], Seung Hoan Choi[2], James G. Wilson[3], Emelia J. Benjamin[4,5,6], Josée Dupuis[1], Kathryn L. Lunetta[1], and the NHLBI Trans-Omics for Precision Medicine Whole Genome Sequencing Program

[1] Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America; [2] The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America; [3] Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, United States of America; [4] National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, Framingham, Massachusetts, United States of America; [5] Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, United States of America; [6] Cardiology and Preventive Medicine Sections, Evans Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, United States of America

Logistic mixed effect models (logME) can be used to test for associations between binary traits and genetic variants while accounting for population structure and the relationships among individuals using a random effect defined by an empirical genetic relationship matrix (GRM). However, the optimal choice of variants to include when generating the GRM is not well understood. Here, we investigate the performance of logME models under a range of GRM choices using meta and joint analysis when the study sample includes samples from two distinct ancestries.

Genotypes from 4177 adults of European ancestry (Framingham Heart Study, FHS), and 3417 adults of African American ancestry (Jackson Heart study, JHS) were measured from the Trans-Omics for Precision Medicine program freeze4 whole genome deep sequencing data. We computed GRMs in FHS and JHS and in the combined FHS+JHS sample using 1) MAF≥0.1% and 2) MAF≥5% variants after linkage-disequilibrium pruning. A binary trait with 50% heritability and a difference in prevalence of 10% versus 5% in the two samples was simulated based on a mix of variants with large and small differences in frequency between the FHS and JHS samples. On average, power was higher for combined than for meta-analysis. The ≥5% MAF and ≥0.1% MAF GRMs produced equivalent type I error and power with meta-analysis, but for combined analysis, the power was lower for the ≥0.1% MAF GRM than for the ≥5% MAF GRM. We recommend combined analysis and GRMs based on common variants for logME association analyses with binary phenotypes.

159 | The Effect of Mating Asymmetry on Maternal Gene-Environment Testing in the Context of an Orofacial Cleft Study

Julie Hudson[1], Jean-François Lefebvre[2], Kelly M. Burkett[1], Marie-Hélène Roy-Gagnon[2]

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

Mating asymmetry (MA) may confound maternal genetic effects and cause spurious associations since it is often necessary to assume mating symmetry to test for these maternal genetic effects, such as when only case-parent trio data are available. It has been shown through simulations that, under MA, assuming mating symmetry in case-parent trios yields increased type I error, whereas including parents of controls can completely account for MA and control type I error. This study aims to investigate through simulations the effect of MA on maternal gene-environment (GE) association tests and the feasibility of controlling for potential MA in data without control parents.

Samples of case-parent trios and control-parent trios were simulated with various levels of maternal GE effect sizes and MA levels comparable to those observed in case-parent trio data from an orofacial cleft study. Multinomial modelling (EMIM software) followed by a test of heterogeneity was used to test for maternal GE effects. Even with low levels of MA, using only case-parent trios yielded inflated type I error rates of the maternal effect tests but not the GE interaction tests under the assumption that MA does not depend on the environment. Adding data on controls only (not their parents) controlled type I error rates but with loss of power compared to adding control parents. This study suggests that it may be possible to use unrelated controls or some external measure of MA to correct for MA and detect maternal GE effects in case trio data.

160 | An Evolutionary Framework for the Study of Gene Function and Disease

Patrick D. Evans[1], Nancy J. Cox[1], Eric R. Gamazon[1]

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

The development of mechanistic explanatory models of protein sequence evolution has broad implications for our understanding of cellular biology, population history, and disease process. Several molecular features have been proposed as potential causal factors for the rate of protein evolution. Here we analyze a variety of gene properties, including expression level, expression breadth, gene age, network interaction, and heritability, and quantify their contribution to protein sequence divergence. We find that many factors contribute independently and we present strong evidence for the importance of gene function. Most notably, co-expression network connectivity and gene expression heritability provide the best univariate models for protein evolutionary rate. Heritability, a measure of the influence of the regulatory genome, is a novel and, an important determinant of evolutionary rate. Furthermore, we use the GTEx human transcriptome resource to examine the forces that shape protein evolution in a multi-tissue framework; in particular, analysis of gene expression using this resource. We implement a self-organizing map (SOM) neural network to jointly analyze all examined potential factors and to account for nonlinear effects, utilizing it to develop a dispensability score for each gene. Methodologically, we demonstrate that genome-wide studies of disease may be enhanced by incorporating the determinants of evolutionary rate into the search for disease genes. Our study presents a comprehensive analysis of a range of factors that constrain molecular evolution and proposes a novel framework for the study of protein function and disease mechanism.

161 | Precision Weighted RNA-seq analyses of Molecular Abundance (RoMA) for Detecting Differential Gene Expression

Guoshuai Cai[1,2], Michael L. Whitfield[2,3]

[1] Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America; [2] Department of Molecular and Systems Biology; [3] Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, United States of America

In this study, we propose a novel method, RoMA, to accurately detect differential expression and unlock the integration with numerous upstream and downstream analyses on mRNA abundance in RNA-seq studies. Various methods have been proposed, each with its own limitations. Some naive normal-based tests have low testing power with the invalid normal distribution assumptions for RNA-seq read counts, whereas count-based methods lack a biologically meaningful interpretation and have limited capability for integration with other analyses packages for mRNA abundance. RoMA incorporates information from both mRNA abundance and raw counts by modeling RPKM (reads per kilobase per million), which represents the relative abundance of mRNA transcripts, and borrowing mean-variance dependency from CPM (counts per million) as a precision weight accounting for the variability in sequencing depth. Studies on simulated data and two real datasets showed that RoMA provides an accurate quantification of mRNA abundance and a value adjustment-tolerant DE analysis with high AUC, low FDR and a desirable type I error rate. This study provides a valid strategy for mRNA abundance modeling and data analysis integration for

RNA-seq studies, which will greatly facilitate the identification and interpretation of DE genes. The method is implemented in a user-friendly R package (*RoMA*).

162 | Family-based Rare Variant Association Study of Familial Myopia in Caucasian Families

Deyana D. Lewis[1], Claire L. Simpson[1,2], Anthony M. Musolf[1], Kyle Long[1,3], Laura Portas[1,4], Federico Murgia[1,4], Elise Ciner[5], Dwight Stambolian[6], Joan E. Bailey-Wilson[1]

[1] Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America; [2] Genetics, Genomics and Informatics, University of Tennessee Health Sciences Center, Memphis, Tennessee, United States of America; [3] University of Texas at El Paso, El Paso Texas, United States of America; [4] Institute of Population Genetics, CNR, Li Punti, Sassari, Italy; [5] Salus University, Elkins Park, Pennsylvania, United States of America, [6] Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Myopia is a common refractive error (RF) which affects at least a third of most populations. Individuals with high myopia are vulnerable to ocular complications later in life. Consequently, great efforts have been undertaken to identify and understand the mechanisms underlying the development and progression of myopia. Genome-wide association studies and linkage studies have identified loci influencing the risk of developing myopia but, few causal variants have been identified with the majority of them being common (minor allele frequency > 0.05). Therefore, this study aims to identify regions associated with rare variants that increase myopia risk using dense exome chip data from 75 myopic Caucasian families from the Penn Family Myopia Study. Myopia was defined based on mean spherical equivalent in Diopters (D): affected (≤ -1D), unaffected (≥ 0 D) or unknown (<0D, >-1 D).

We used the rare-variant transmission disequilibrium test (RV-TDT) to perform gene-based tests with rare variants in 25 Caucasian parent-child trios. After quality control, 27,121 SNPs were analyzed and RV-TDT analysis identified a suggestively associated locus on chromosome 2.q31.1for a non-synonymous coding SNP rs34564141 in *LRP2* after correcting for multiple testing ($p\text{-value} = 2.65 \times 10^{-3}$). *LRP2* has been implicated in a study of a rare form of severe myopia in patients who had a mutation encoding a receptor *LRP2* in the retina. Greatly enlarged eyes have been observed in mice lacking this *LRP2* gene. We are extending these analyses to include 350 individuals from additional families and will present these results.

163 | Machine Learning Based Methods for Identifying and Modeling Genetic Interactions

Emily R. Holzinger[1], Silke Szymcak[2], James D. Malley[3], Joan E. Bailey-Wilson[1]

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

Standard analysis methods for genome-wide association studies (GWAS) are not robust to complex disease models. These effects very likely contribute to the genetic etiology of complex human traits. Machine learning methods that are capable of identifying complex effects, such as Random Forests (RF), are an alternative analysis approach. We have developed an RF based variable selection method called relative recurrency variable importance metric (r2VIM) to address key hurdles in applying RF to genome-wide analyses of complex human phenotypes.

Previously, we have shown that r2VIM has improved false positive control and power over RF, especially when non-linear effects contribute to the phenotype. Here we assess the benefit of r2VIM for post-selection modeling and prediction. We do this using simulated GWAS data with a range of SNP counts (from 100 to 300,000) and different levels of linkage disequilibrium (LD). We compare the best models from an artificial neural network (ANN) analysis of the variables selected by r2VIM to the complete set of variables. Our results show that initial filtering with r2VIM improves the ability of ANNs to identify non-linear interactions by reducing noise in the data. We also assess the prediction accuracy of the r2VIM-based models compared to the non-filtered models.

In summary, we have developed a machine learning pipeline using r2VIM and ANNs to identify and model complex genetic effects in a computationally efficient manner. This has the potential to elucidate novel biological pathways which could improve both treatment and prediction of complex human diseases.

164 | Test Gene-Environment Interactions for Multiple Phenotypes Traits in Sequencing Association Studies

Jianjun Zhang[1,2], Shuanglin Zhang[3], Han Hao[1], Qiuying Sha[3], Xuexia Wang[1]

[1] University of North Texas, Texas, Denton, Texas, United States of America; [2] East China Normal University, Shanghai, People's Republic of China; [3] Michigan Technological University, Houghton, Michigan, United States of America

A set of correlated phenotype traits such as broader autism phenotype traits may share common genetic factors. Examination of the multiple traits can yield valuable insights about the disease etiology and increase power in detecting genetic variants. In this study, we develop novel approaches to test gene-environment interactions (G × E) for multiple traits in sequencing association studies. We first make the transformation of multiple traits by using either principle component analysis or standardization analysis. Then, we detect the effect of G × E for each transferred phenotype trait using novel proposed tests: TOW-GE and/or VW-TOW-GE. Both TOW-GE and VW-TOW-GE are robust to directions of effects of causal G × E. Finally, we employ the Fisher's combination test to combine the p-values of TOW-GE and/or VW-TOW-GE. Extensive simulation studies based on the Genetic Analysis Workshop 17 data show that the type I error rates of the proposed methods are under control. Compared to the existing method iSKAT, the power of applying TOW-GE is more powerful when there are only rare risk and protective variants; the power of applying VW-TOW-GE is more powerful when there are both rare and common risk and protective variants. Application to the ARRA autism sequencing consortium data demonstrates that our proposed methods are very powerful.

165 | 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, California, United States of America

Because of its simplicity, linear discriminant analysis (LDA) performs well in classification problems when the sample size is small. In recent years, LDA has been extended for classifying high-dimensional data ($p \gg n$) in different ways. Scout is a family of regression and classification methods for high dimensional settings that uses a shrunken estimate of the inverse covariance matrix of the features. We propose an extension of scout LDA that uses sparsity-inducing penalties to estimate both the inverse covariance matrix of the genomic features and the difference in means between the classes. Specifically, we use the graphical LASSO (GLASSO) to obtain a sparse estimate of the inverse covariance matrix along with soft-threshold operator to estimate the difference in means. Through simulation, we show that our proposed LDA approach yields better prediction than existing methods extending LDA to the high-dimensional setting. We apply our approach to a prostate cancer dataset with 187 subjects and high dimensional gene expression features to identify a gene expression signature predictive of clinical recurrence after radical prostatectomy in early stage prostate cancer patients.

166 | Inaccuracies in Kinship Estimation can be Minimized with Principal Components by PC Relate in Data Sets Containing Subjects with Diverse Ancestry

Elizabeth E. Blue[1], Jessica X. Chong[2], University of Washington Center for Mendelian Genomics

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

Kinship estimates have many uses in the field of genetic epidemiology, including the confirmation of sample identity, the prioritization of modes of inheritance for Mendelian gene discovery, and the correction of population stratification in genome-wide association testing. It has been established that population stratification within a sample can cause inflated kinship estimates. Tools such as KING robust, peddy, and PC Relate have been developed specifically to address this problem. However, the demonstrated efficacy of these tools relied on large samples of SNP markers and/or subjects ascertained from a single metapopulation. We compare the performance of these tools using exome data collected on a diverse set of subjects drawn from a national repository, including 46 duos/trios clustering with African, European, or Native American reference populations. We show that the distribution of kinship estimates among unrelated pairs is closer to zero for PC relate than either peddy or KING robust. The practical impact of this observation can be seen in the count of unrelated pairs with kinship estimates appropriate for second cousins: KING robust has 60 pairs, peddy has 88 pairs, and PC Relate has 11 pairs. Unlike the expectation under true cryptic relatedness, the list of subjects contributing to these pairs varies across programs. We anticipate this variation in the precision of kinship estimation would impact association testing methods which rely on a genetic relatedness matrix to correct for structure within a case-control data set, leading to biased test statistics.

167 | ukbrest: REST API for Easy and Efficient Access to UK Biobank Data

Milton D. Pividori[1,2], Hae K. Im[1,2]

[1] Section of Genetic Medicine, Department of Medicine, University of Chicago, Illinois, United States of America; [2] Center for Data Intensive Science, University of Chicago, Illinois, United States of America

The UK Biobank provides an unprecedented opportunity for researchers to obtain insights about genetic and non-genetic disease factors, with massive collections of genotypic and phenotypic data from the UK that has been and is being collected. Scientists have access to data as diverse as questionnaires and physical measures to multimodal imaging and electronic medical records. Although this tremendous amount of data represents a unique resource, its highly heterogeneous data types, different sources, and structure, in addition to its continuously increasing size, can be a challenge to conduct an effective data integration, thus posing obstacles for the analysis of the complex interplay of different determinants that cause diseases. We introduce here *ukbrest* (https://github.com/miltondp/ukbrest), a REST API implementation that allows researchers to efficiently access subsets of SNPs and phenotypic traits, leveraging a BGEN indexer and a SQL database. We have used this tool to efficiently distribute jobs performing phenome-wide gene level association with PrediXcan/MetaXcan to hundreds of phenotypes, leveraging a secure biomedical cloud called Bionimbus Protected Data Cloud (PDC), which operates at FISMA moderate as IaaS with an NIH Trusted Partner status for analyzing protected datasets. By making data access not only efficient and secure but also its integration much more streamlined, ukbrest also represents a step towards improved reproducible research of the UK Biobank, allowing all researchers with proper access to easily reproduce results from others and leverage on them for further investigations.

168 | A Novel Test Method for Joint Effect of Gene and Methylation Level in GWAS and EWAS Data

Gaokang Wang[1], Jianzhong Li[1], Hong Zhang[1], Jin Xu[1], Zhaogong Zhang[1,2]

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

To test whether differential methylation of cytosine-(phosphate)-guanine dinucleotides (CpGs) and SNP variants have joint effect with quantitative traits, we develop statistical methods. We propose two strategies and their combination for this purpose: the iterative regression strategy and the extreme values strategy. In the iterative regression strategy, we use iterative regression on residuals and a multi-marker association test to identify a group of significant variants. In the extreme values strategy, we use individuals with extreme trait values to select candidate genes and then test only these candidate genes. These two strategies are integrated into a hybrid approach through a weighting technology. We apply the proposed methods to analyze a simulation data set of GWAS and EWAS data. The results show that the hybrid approach is the most powerful approach. Using the hybrid approach, the average power to detect causal genes for quantitative traits is about 30%.

169 | A Regularized Hierarchical Regression Framework 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, California, United States of America

Incorporation of annotation information available on the structure, function, and regulation of the genome into high-dimensional regression models has the potential to improve the prediction of health-related outcomes based on genomic features. We propose a novel penalized hierarchical model that extends ridge regression to include annotation data in a second level regression model. A sparsity-inducing penalty on the high-dimensional annotation features allows the model to identify relevant annotations to the prediction task at hand. To efficiently fit the model, we have developed an algorithm that exploits the convexity of the model objective function to alternate between fitting each level of the hierarchy conditional on the current estimates of the other level. Through simulation, we show that when external information is truly informative, we can improve the predictive ability of models compared to standard approaches that do not include external information. Moreover, there is little to no penalty on prediction performance when the external data is non-informative. We apply our hierarchical model with Gene Ontology annotations to identify a gene expression signature predictive of clinical recurrence after radical prostatectomy in a study of early stage prostate cancer patients.

170 | Generalizing Genetic Risk Scores from Europeans to Hispanics/Latinos

Kelsey E. Grinde[1], Qibin Qi[2], Timothy A. Thornton[1], Simin Liu[3,4], Aladdin H. Shadyab[5], Kei Hang K. Chan[6], Alexander P. Reiner[7], Tamar Sofer[1]

[1] Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [2] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, United States of America; [3] Department of Epidemiology, Brown University School of Public Health, Providence, Rhode Island, United States of America; [4] Department of Medicine, Alpert School of Medicine, Brown University, Providence, Rhode Island, United States of America; [5] Division of Epidemiology, Department of Family Medicine and Public Health, University of California, San Diego School of Medicine, La Jolla, California, United States of America; [6] Laboratory of Molecular Epidemiology and Nutrition, Department of Epidemiology, Brown University, Providence, Rhode Island, United States of America; [7] Department of Epidemiology, University of Washington, Seattle, Washington, United States of America

Genetic risk scores (GRS) are summaries of multiple genetic variants constructed by, for example, taking the weighted sum of trait-increasing alleles with weights equal to the estimated effect sizes of the variants. GRS are widely used in studies involving disease risk or trait prediction and Mendelian randomization. As with genome-wide association studies (GWAS), most of the research on GRS in humans has been conducted in populations of European ancestry. However, optimal GRS may differ between Hispanic/Latino and European populations due to different linkage disequilibrium patterns, gene-environment interactions, and allelic heterogeneity. In this study, we investigated methods for constructing GRS in samples of Hispanics/Latinos that combined GWAS results from both medium-sized studies of Hispanics/Latinos and large studies of Europeans. Specifically, we studied multiple approaches for selection of both SNPs and weights in constructing GRS. We built GRS for a number of anthropometric and blood cell count traits using GWAS results from publically available European studies (tens of thousands of individuals) and the Hispanic Community Health Study/Study of Latinos (12,784 individuals), and evaluated the performance of these scores in a smaller, independent study of 3,582 Hispanic women from the Women's Health Initiative SNP Health Association Resource. We further evaluated the performance of the proposed methods in simulations of genetic loci exhibiting various trait-genetic architectures. We found that using a combination of Hispanic/Latino and European GWAS results to build GRS for Hispanics/Latinos can improve prediction error and correlation with the trait compared to GRS built using just European GWAS results.

171 | ExpressionLncr: A Pipeline for Leveraging Latent Gene Expression Data in lncRNA Studies

George Ellis[1], Loubna Akhabir[1], Carolyn Brown[2], the ASTOR Consortium, and Denise Daley[1,3]

[1] Centre for Heart Lung Innovation, University of British Columbia, Vancouver, Canada; [2] Department of Medical Genetics, University of British Columbia, Vancouver, Canada; [3] Department of Medicine University of British Columbia, Canada

Investigation of long non-coding RNAs (lncRNA) is active and growing area of research. There are databases that catalog lncRNA's (NONCODE or LNCipedia), but the the challenge is that only a small number of lncRNA's are validated (i.e. demonstrated to produce an RNA product), most are "putative" lncRNA's and there is little information on the genes that these lncRNA's regulate. Demonstrating that a putative lncRNA produces an RNA product is both time-consuming and expensive. Informatics tools that leverage existing databases and information sources to prioritize lncRNA's for further investigation and determining the genes regulated are needed. A wealth of functional genomics data (70 thousand experiments and 1.8 million samples) is available in the NCBI Gene Omibus (GEO) database, however, harnessing this data is challenging for those without informatics expertise.

We have created a program called ExpressionLncr which is a bioinformatics pipeline to investigate the functionality of lncRNAs and other chromosomal features by computing positional overlap between lncRNA databases and existing gene expression probe information in the NCBI GEO database. The software identifies "putative" lncRNAs from NONCODE or LNCipedia and combines this with expression information from the GEO database. The tool computes matches for positional overlap between Ensembl expression probes and lncRNAs. Summary results from GEO DataSets relevant to these overlapping features are used to calculate presence or absence of expression at each lncRNA which can be used in combination with hierarchal regression models may prioritize lncRNA for further investigation.

172 | Identification of Genetic Heterogeneity of Alzheimer's Disease Across Age

Min-Tzu Lo[1], Chun-Chieh Fan[1,2], Karolina Kauppi[1,3], Nilotpal Sanyal[1], Rahul S. Desikan[4], Lindsay A. Farrer[5], Jonathan L. Haines[6], Richard Mayeux[7], Margaret Pericak-Vance[8], Gerard D. Schellenberg[9], Chi-Hua Chen[1], Alzheimer's Disease Genetics Consortium

[1] Center for Multimodal Imaging and Genetics, Department of Radiology, University of California, San Diego, La Jolla, California, United States of America; [2] Department of Cognitive Science, University of California, San Diego, La Jolla, California, United States of America; [3] Department of Radiation Sciences, Umea University, Umea, Sweden; [4] Neuroradiology Section, Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, United States of America; [5] Departments of Medicine (Biomedical Genetics), Neurology, Ophthalmology, Epidemiology, and Biostatistics, Boston University, Boston, Massachusetts, United States of America; [6] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, United States of America; [7] Taub Institute on Alzheimer's Disease and the Aging Brain, Gertrude H. Sergievsky Center, and Department of Neurology, Columbia University, New York, New York, United States of America; [8] The John P. Hussman Institute for Human Genomics, and Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, Florida, United States of America; [9] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States of America

Age has been showed to modify *APOE* effects on Alzheimer's Disease (AD). We used the Genome-wide Complex Trait Analysis (GCTA) tool to estimate heritability in two age groups (age 60–79 and $\geq$ 80 years) using Alzheimer's Disease Genetics Consortium (ADGC) sample. Overall, heritability was estimated to be 18.1% after adjustment for age, sex, study cohorts and top 10 principal components. Heritability estimates for two age groups were 16.4% and 23.2% and their genetic correlation (0.64, p-value = 0.043) significantly differed from 1 suggesting genetic heterogeneity between younger and older AD cases. We performed genome-wide association tests for two age groups in ADGC phase 1 sample ($N = 16{,}159$). To verify discovery findings, we performed replication analyses in phase 2 sample ($N = 5{,}295$) and combined phase 1 and 2 samples ($N = 21{,}454$). Given p-value$<5 \times 10^{-8}$ and the same directions of effect sizes, we identified 15 and two LD-independent SNPs in age 60–79 and $\geq$ 80 years, respectively, according to discovery and replication analyses. The effects of SNPs close to *APOE* on chromosome 19 were robust across samples and contributed larger effects to age 60–79 years. In addition, the SNP on chromosome 2 close to AD susceptibility gene *BIN1* were significant in age 60–79 years but not in age $\geq$ 80 years. Our results suggested that further GWAS could be performed for each age group with genetically homogeneous AD cases and this analysis strategy might help to explore biological mechanisms for pharmaceutical development.

173 | Leveraging Large Affected Families and Publicly Available Data to Shrink Confidence Intervals of Trait Location

William C.L. Stewart[1]

[1]The Research Institute at Nationwide Children's Hospital & The Ohio State University, Columbus, Ohio, United States of America

For cosegregation studies involving hundreds of small families and high-throughput genotype data, it is difficult to improve upon the near-optimal estimator of trait location that averages location estimates over random subsamples of the dense marker data. However, for studies with a small number of large families there is still considerable room for improvement. In this setting, accurate estimation of the variance of the near-optimal estimator is particularly difficult due to correlations within the high-throughput genotype data. Here, I describe an importance sampling approach that accurately approximates the precision of the near-optimal estimate. My approach uses Monte Carlo simulation and (optionally) the genotypes of publicly available reference samples to account for the correlations in the dense marker data. I applied my proposed estimator of the variance to the dense marker data of four large families segregating a specific language impairment gene on chromosome 13. My 95% confidence interval for trait location is 25% shorter than the standard interval, and this degree of shrinkage agrees with predictions in the literature based on simulated data. As such, researchers with large affected families and high-throughput genotype data should now be able to significantly reduce their targeted resequencing costs, and greatly expedite the rate at which disease genes are found.

174 | Accounting for Cryptic Relatedness across Families in Family-based Association Testing

Ellen M. Wijsman[1,2], Elizabeth E. Blue[1], Tyler R. Day[1], Alejandro Q. Nato Jr.[1], Harkirat K. Sohi[1], Andrea R. Horimoto[1], Rafael Nafikov[1], Mohamad H. Saad[3], Timothy A. Thornton[2]

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, United States of America; [2]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America; [3]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Availability of modern whole-genome sequencing (WGS) provides unprecedented opportunities for investigation into inherited variation. Although studies of unrelated subjects through GWAS are common, family-based studies continue to be of interest. Family studies are particularly effective for identification of rare variants and are efficient since only a subset of individuals need to have directly-observed sequence data, with sequence data imputed into the remaining subjects via the pedigree information. Even if linkage analysis is also used to limit the analytical space, however, some association testing is still necessary. Approaches that adjust for relatedness exist, but most, if not all, rely on kinship estimates between pairs of individuals. We have been using association testing of WGS variants in samples with multiple families. Here the WGS has been augmented by family-based imputation into unsequenced subjects to increase the effective sample size. In initial association analyses, we obtained excessively inflated type-I errors, even when adjusting for within-pedigree relatedness and population structure via Principal Components (PCs) defined by subjects with observed genotype data. We evaluated the cause by using kinship matrices defined on subsets of the data: subjects with both WGS and GWAS SNP data, adding subjects with observed SNPs and imputed WGS, and finally, adding subjects with only imputed WGS. Where we have observed SNP data, cross-pedigree cryptic relatedness can be modelled, thus controlling the type-I error. Full control of the type-I error requires estimating the kinship matrix across pedigrees including, e.g., the imputed subjects, thus requiring methods that handle genotype dose files.

175 | Characterizing Disease and Genetic Risks in Familial Colorectal Cancer Type X Families

Yun-hee Choi[1], Agnieszka Krol[2], Laurent Briollais[2,3]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Canada; [2]Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Toronto, Canada; [3]Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Lynch Syndrome (LS) is the most common hereditary colorectal cancer (CRC) syndrome caused by mutations in DNA mismatch repair (MMR) genes whereas the Familial Colorectal Cancer Type X (FCCTX) is a familial form of cancer whose members suffer from multiple CRCs but with tumors that are DNA-mismatch-repair proficient. It is known that clinical and pathologic characteristics differ between LS and FCCTX families. Individuals with LS have a high risk of developing a first and a second primary CRC. However, these risks are not well characterized in FCCTX families. In this work, we compare the age-dependent absolute risk estimates of first and second CRCs in FCTTX and LS families recruited through the Colon Cancer Family Registry (CCFR). In the risk estimation, we account for the presence of competing events such as death or other cancers. We also estimate the familial correlation for each event in the FCCTX families and compare it to the familial correlation explained by MMR mutations in LS families using frailty models with a correlation structure given by the IBD matrix. Our goal is to assess whether the familial correlation in FCCTX families can fit the segregation pattern of a high-penetrant rare mutation (such as in LS families) or the co-segregation of multiple moderate-penetrant alleles.

176 | A Novel Bayesian Multiple Testing Approach for Region-Based Analysis of Next Generation Sequencing (NGS) Data

Jingxiong Xu[1,2], Wei Xu[1,3], Rayjean Hung[1,2], Geoffrey Liu[1,3], Laurent Briollais[1,2]

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada; [2]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada; [3]University of Health Network, Princess Margaret Hospital, Toronto, Canada

The discovery of rare genetic variants through Next Generation Sequencing (NGS) is becoming a very challenging issue in the human genetic field. We have recently proposed a novel region-based statistical test based on a Bayes Factor (BF) approach to assess evidence of association between a set of rare variants located on this region and a disease outcome. Gene-based inference can be performed using permutation testing. For genome-wide inference (i.e. multi-regions test), we introduce a Bayesian False Discovery Rate (BFDR) control procedure inspired by the work of Efron (2005). In this setting, it is critical to estimate the null distribution of the BF for the non-associated regions. We used an empirical estimate of the null distribution and studied its properties by simulation and analytically. In particular, we showed that the choice of priors can impact seriously the BFDR procedure. Our BF approach has been applied to a study of lung cancer from Toronto including 262 cases and 261 controls with whole exome sequencing data. We applied the BFDR control procedure to >13.7K genes and estimated the FDR for the top 5, 10 and 20 genes to be respectively, 2.9%, 7.5% and 12.8%. Our top genes included for example TERT and TLR6. In conclusion, the use of empirical Bayes priors along with a Bayesian control of FDR offer a comprehensive framework to make genome-wide statistical inference about the important chromosomal regions associated with the disease of interest in the context of NGS data.

177 | Addressing the Missing Data Issue in Multi-Phenotype Genome-Wide Association Studies

Mila D. Anasanti[1], Marika Kaakinen[1,2], 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

Multi-phenotype genome-wide association studies (MP-GWAS) play an important role in improving the power for locus discovery. However, joint analysis of multiple phenotypes increases the proportion of missingness, leading to inefficiency of the standardly implemented complete case (CC) analysis. We investigated the properties of conditional imputation, multiple imputation (MI), expectation-maximisation (EM), and EM+MI within the MP-GWAS framework, and compared them with the full data and CC analyses. We simulated genetic data for 5,000 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 chose common (minor allele frequency, MAF = 0.157), low-frequency (MAF = 0.0208) and rare (MAF = 0.0016) variants to be significantly ($P<5 \times 10^{-8}$) associated with the simulated phenotypes. We considered different proportions of missing data (1/2.5/5/10/20/30/40/50%) under the three mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The resulting betas and SEs from the analyses regressing each SNP on the linear combination of the phenotypes, after applying the selected missing data methods, were compared to the true values from the full data analysis. These analyses showed that the estimates from MI/EM/EM+MI were the least biased, followed by those from conditional imputation, even under the scenario of MNAR, although MI assumes at least MAR. The performance of CC analysis worsened with higher number of phenotypes, although other approaches were not influenced neither by the number of phenotypes nor by the differences in phenotype correlations or MAF. In summary, MI/EM/EM+MI are recommended over the commonly applied CC analysis.

178 | A Copula-Based Likelihood Approach for the Analysis of Secondary Phenotypes in Selected Samples

Fodé Tounkara[1], Geneviève Lefebvre[2], Celia Greenwood[3], Karim Oualkacha[2]

[1]Lunenfeld-Tanenbaum Research Institute, University of Toronto, Toronto, Canada; [2]Université du Québec à Montréal (UQAM), Montreal, Canada: [3]Lady Davis Institute for Medical Research, Jewish General Hospital and McGill University, Departments of Oncology, Epidemiology, Biostatistics & Occupational Health, Human Genetics and Ludmer Centre for Neuroinformatics and Mental Health, Montreal, Canada

Selected samples, where individuals are ascertained based either on their primary phenotypes (e.g. case-control or extreme-trait studies) or on their primary and secondary phenotypes (e.g. Multiple-traits studies), are often used for genome-wide association studies of secondary phenotypes.

However, if there is dependence between the primary and the secondary phenotypes, naive analysis methods relying on standard regressions may lead to false-positive association.

Here, we propose a copula-based approach to model the dependence between a primary trait and a continuous secondary phenotype. The prospective likelihood approach is used to correct for bias induced by the sampling mechanisms.

We present simulation study examining the performance of our proposed methods under different scenarios and with several different copulas. Results are compared with those obtained under a naive analysis method and existing multivariate normal-based approaches.

We demonstrate the effectiveness of our method by analyzing data from the ALSPAC study. We have ascertained individuals from this cohort based on high HDL levels, and we are examining genetic associations with three secondary phenotypes including LDL, ApoB and ApoA1 levels.

Numerical results show that when there is genetic association with the primary trait, then the proposed method controls type 1 error well in the presence of dependent primary-secondary phenotypes. The method is also robust to the misspecification of the copula model. We plan to extend our current method to allow detection of associations between secondary phenotypes and rare genetic variation with individuals in clustered sets such as families.

179 | Applications of Multidimensional Time Model for PDF to Model Permeability of Plasma Membrane and Transcription of Cytoplasmic DNA

Michael Fundator[1]

[1] *Division of Behavioral and Social Sciences and Education of the National Academies of Sciences, Engineering, and Medicine, Washington, District of Columbia, United States of America*

The differences between single and multi-cell analyses include chemical problems as well as statistical problems related to noise models for single-cell transcriptomics. Another challenging aspect of these differences is the problem of distinguishing genuine from technical stochastic allelic expression that is important in the decomposition of tissues into cell types. To address these challenges we propose a new method based on changes of Cumulative Distribution Function in relation to time change in sampling patterns. Multidimensional Time Model for Cumulative Probability Distribution Function can be reduced to a finite-dimensional time model, which can be characterized by Boolean algebra for operations over events and their probabilities. The infinite dimensional time model can be reduced to a finite number of dimensions of time model using an index set and by considering the fractal-dimensional time arising from alike supersymmetrical properties of probability. The new method is based on properties of Brownian motion.