EMGM 2024

Wednesday 3rd April

0930 - 1020

DNA methylation-based predictors of lifestyle and disease
Riccardo Marioni
University of Edinburgh, UK

Blood-based DNA methylation patterns can track differences in lifestyle behaviours and stratify risk of incident disease outcomes. Using epigenome-wide data from the Generation Scotland cohort (n>18,000 with >800,000 CpGs profiled), I will discuss some of the discuss challenges we have faced in our prediction modelling. This includes: 1) feature pre-selection prior to training; 2) features with non-linear effects; and 3) external testing in cohorts of non-European ancestries. I will present findings from upcoming pre-prints on smoking, alcohol consumption, six measures of metabolic health and incident type 2 diabetes.

1020 - 1040

Fine-mapping the results from genome-wide association studies of primary biliary cholangitis using SuSiE and h2-D2
Heather J. Cordell, Aida Gjoka
Population Health Sciences Institute, Newcastle University, UK

The main goal of fine-mapping is the identification of relevant genetic variants that have a causal effect on some trait of interest, such as the presence of a disease. From a statistical point of view, fine-mapping can be seen as a variable selection problem. Fine-mapping methods are often challenging to apply because of the presence of linkage disequilibrium (LD), that is, regions of the genome where the variants interrogated have high correlation. Several methods have been proposed to address this issue. Here we explore the Sum of Single Effects (“SuSiE”) method, applied to real data (summary statistics) from a genome-wide meta-analysis of the autoimmune liver disease primary biliary cholangitis (PBC). Fine-mapping in this data set was previously performed using the FINEMAP program; we compare these previous results with those obtained from SuSiE, which provides a more convenient and principled way of generating “credible sets”, i.e. set of predictors that are correlated with the response variable. This allows us to appropriately acknowledge the uncertainty when selecting the causal effects for the trait. We focus on the results from “SuSiE-RSS”, which fits the SuSiE model to summary statistics, such as z-scores, along with a correlation matrix. We also compare the SuSiE results to those obtained using an even more recently developed method, h2-D2, which uses the same inputs. As a proof of principle, we further apply SuSiE and h2-D2 to simulated data generated using the HAPGEN2 software.

1040 - 1100

Light-speed whole genome association testing and prediction via Approximate

Message Passing

Al Depope, Marco Mondelli, Matthew R. Robinson
Institute of Science and Technology Austria, AT

Efficient utilization of large-scale biobank data is crucial for inferring the genetic basis of disease and predicting health outcomes from the DNA. Yet we lack accurate statistical models to estimate the effect of each locus, conditional on all other genetic variants, controlling for both local and long-range linkage disequilibrium. Additionally, we lack algorithms which scale to data where health records are linked to whole genome sequence information. To address these issues, we develop a new algorithmic paradigm, based on Approximate Message Passing (AMP), specifically tailored for both genomic prediction and association testing. Our gVAMP (genomic Vector AMP) approach requires less than a day to jointly estimate the effects of 8.4 million imputed genetic variants in over 400,000 UK Biobank participants, and it provides an association testing framework capable of directly fine-mapping each genetic variant, or gene burden score, conditional on all other measured DNA variation genome-wide. Across 13 traits, we find 8,222 genome-wide significant autosomal associations that are localised to the single-locus level, conditional on all other imputed loci. Extending the model to jointly estimate the effects of rare variant gene burden scores from sequencing data and imputed X chromosome variants, conditional on all other genes and all 8.8 million variants, we find 60 genes where rare coding mutations significantly influence phenotype, and 76 associations localised to the single-locus level on chromosome X, for five traits. In comparison to existing state-of-the-art methods, both in simulations and in application to the UK Biobank, gVAMP yields out-of-sample prediction accuracy comparable to individual-level Bayesian methods, outperforms summary statistic Bayesian methods, and outperforms REGENIE for standard association testing, in a fraction of the compute time. This truly large-scale development of the AMP framework establishes the foundations for a far wider range of statistical analyses for hundreds of millions of variables measured on millions of people.

1130 - 1150

Target Trial Emulation in biobank data: estimating the effect of cholesterol-lowering therapy and PRS in the Estonian Biobank

Saskia Kuusk, Lili Milani, Krista Fischer
Institute of Genomics, University of Tartu, EE

For estimating the effects of a treatment on an outcome, Randomized Control Trials (RCTs) are acknowledged to be the gold standard. Population-based biobanks, however, offer a different, more cost- and time-effective approach to use observational data from Electronic Health Records. Although such data is not collected for research purposes and lacks random treatment assignment, it is still possible to obtain the same estimates as one would get from an RCT. We will demonstrate the first steps of such Target Trial Emulation (TTE) method on Estonian Biobank data to estimate the effect of cholesterol-lowering therapy and PRS on cardiovascular disease. This includes specifying the trial components as they would appear in a hypothesised RCT and how to emulate them in our observational data. By conducting a TTE analysis, we have identified the protective effect of cholesterol-lowering therapy on cardiovascular diseases, as well as validated the effect of PRS.

However, during the initial stages of this analysis, we recognized several common pitfalls inherent in TTE analyses. Among those encountered are classical immortal-time-bias and the omission of key risk factors.   Addressing these is crucial to accurately estimating the true causal effect of interest. To tackle the issue of immortal-time bias, additional measures are required. We aim to mitigate this bias by employing sequential TTE.

1150 - 1210

Modelling genetic effects on gene expression dynamics in B cells

Daniela Zanotti (1,2), Alessandro Raveane (1), Matiss Ozols (3), Marc Jan Bonder (4), Francesca Ieva (2,5), Nicole Soranzo (1,3,6), Nicola Pirastu (1), Blagoje Soskic (1)

(1) Genomic Centre, Human Technopole, Milan, IT, (2) MOX, Department of Mathematics, Politecnico di Milano, Milan, IT, (3) Department of Human Genetics, Wellcome Sanger Institute, Genome Campus, Hinxton, UK., (4) Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, NL, (5) Health Data Science Center, Human Technopole, Milan, IT, (6) British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

During activation and differentiation B cells go through significant changes in gene expression. Previous studies have shown that expression quantitative trait loci (eQTLs) can have dynamic effects in B cells undergoing differentiation from naive to memory phenotype. Some of these studies cluster cells according to their position along pseudotime trajectories and are dependent upon the binning choice, while other methods that do not rely on clustering are too computationally intensive to be applied to large-scale settings. We aimed to address both these limitations by modelling gene expression as a continuous function of pseudotime, and investigating allelic effects on gene expression dynamics in B cells transitioning through differentiation states. Using single-cell RNA sequencing data of B cells from the OneK1K cohort, we first estimated pseudotime trajectories and then modelled the gene expression of 9,155 genes along them using positive constraint cubic splines. Finally, we applied Functional Principal Component Analysis (FPCA) to both the generated curves and their derivatives and used the resulting scores for cis-eQTL analysis. We detected 2,022 and 236 genes with significant dynamic cis-eQTLs using the FPCA scores of the original curves and the derivatives, respectively. In the first case, many functions showed differences in height between genotypes, while the derivatives capture effective differences in the functions’ shape. The identified genes were enriched for specific immune cell signatures, including distinctions between naïve and memory B cells, and responses to vaccination. Our approach demonstrated high efficacy in detecting genetic loci which regulate dynamic gene expression in B cells. Moreover, we were able to detect both dynamic and static eQTLs, without relying on prior assumptions such as clustering while retaining low computational costs. This suggests that our method could be efficiently used in large-scale single-cell cohorts.

1210 - 1230

Discovery of shared epigenetic pathways across human phenotypes

Ilse Krätschmer (1), Hannah Smith (2), Daniel M. McCartney (2), Elena Bernabeu (2), Mahdi Mahmoudi (1), Sarah E. Harris (2), Janie Corley (2), Simon R. Cox (2), Riccardo E. Marioni (2), Matthew R. Robinson (1)
(1) Institute of Science and Technology Austria, AT, (2) University of Edinburgh, UK

We present a Bayesian method to learn shared and outcome-specific effects for multiple traits in multi-omics data. The method determines the unique contribution of individual loci, genes, or molecular pathways, to variation in one or more traits, conditional on all other measured ”omics” data genome-wide. Simulations show the model accurately finds shared and distinct associations between omics-data and multiple traits and estimates omics-specific (co)variances, allowing for sparsity and correlations within the data. We will show results for 12 outcome traits in Generation Scotland, where novel shared epigenetic pathways among cholesterol metabolism, osteoarthritis, blood pressure and asthma were found. Localising epigenetic association to the single probe level, only 10 CpG probes with significant effects above the genome-wide background are found, due to the challenging highly correlated observational omics data.

1400 - 1450

Evaluating the 4 R's of Mendelian randomization studies: reporting, reproducibility, robustness and reasonableness

Rebecca Richmond
University of Bristol, UK

Mendelian randomization (MR) is a statistical method which uses genetic variants as an instrument variable that, under specific assumptions, can be used to investigate the causal relationship between an exposure and outcome. As the availability of large and deeply phenotypes population studies has increased, so too has the application of MR. This is particularly true for two-sample MR, in which summary-level data from genome-wide association studies are leveraged and analysis performed in a largely automated fashion. In previous work, we found a large proportion of the Mendelian randomization literature to be inadequately reported. The STROBE-MR checklist was subsequently introduced as a guideline to improve reporting of Mendelian randomization studies. As part of this talk, I will discuss the motivation for developing the STROBE-MR guidelines as well as whether it has had the anticipated effect of increasing reporting quality in MR studies. I will also ask the question whether higher reporting quality in turn improves the strict replication of MR findings and robustness of inferences (particularly with respect to addressing the MR assumptions). Finally, I will discuss an important threat, which is that the increasing ease of performing MR has in turn facilitated some mindless studies with unreasonable exposures (aka 'noodles'). I will discuss how we might navigate and confront this challenge in a new era of Mendelian randomization.

1450 - 1510

Exploring and accounting for genetically driven effect heterogeneity in Mendelian Randomization

Annika Jaitner(1), Krasimira Tsaneva-Atanasova(2,3), Rachel M. Freathy(1), Jack Bowden(1,4)

(1) Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, UK, (2) Department of Mathematics and Statistics, Faculty of Environment, Science and Economy, University of Exeter, Exeter, UK, (3) EPSRC Hub for Quantitative Modelling in Healthcare University of Exeter, Exeter, UK, (4) Novo Nordisk Genetics Centre of Excellence, Oxford, UK

Mendelian randomization (MR) uses genetic variants as instrumental variables to estimate the causal effect of a modifiable health exposure or drug target on a downstream outcome. A crucial assumption for accurate estimation of the average causal effect in a population using MR is Homogeneity. This means that the causal effect an individual experiences is not affected by the value of their genetic instrument. In contrast, pharmacogenetics seeks to actively uncover and exploit genetically driven effect heterogeneity for precision medicine. Observational data can be used to quantify the extent of genetically driven treatment effect heterogeneity, but the analysis can be compromised by strong confounding by indication and off-target genetic effects on the outcome of interest that are independent of any gene-drug interaction. A recently proposed method of pharmacogenetic causal inference using observational data – Triangulation Within a Study (TWIST) - defined the assumptions required to estimate the difference in treatment effect estimates between those with and without a pharmacogenetic variants, as a measure of genetically driven effect heterogeneity. We explore the integration of homogeneity-violating and homogeneity-respecting instruments and propose two new methods to porperly characterise the average causal effects and gentically driven effect heterogeneity in a standard epidemiological MR study setting. Through Monte-Carlo simulation, we verify the required assumptions for unbiased estimation, as well as assessing the accruacy, precision, power and coverage of our methods. Using data from the ALSPAC study, we apply our methods to estimate the causal effect of smoking before and during pregnancy on offspring birth weight in mothers whose genetics mean they find it (relatively) easier or harder to quit. SNP rs1051730 on chromosome 15 is used as a homogeneity-violating instrument due, to its association with smoking cessation in pregnancy but not smoking initiation. We generate a genetic risk for smoking initiation to act as our homogeneity-respecting instrument. Our analysis did not reveal a meaningful amount of effect heterogeneity, due to a low statistical power. Nevertheless, the results obtained support a negative causal effect of smoking before and during pregnancy with birth weight. In conclusion, we believe our framework is a useful methodological extension to investigate genetically driven heterogeneity in MR studies.

1510 - 1530

Dissecting ancestry-aware molecular causal paths of type 2 diabetes

Ana Luiza Arruda1,2,*, Ozvan Bocher1,*, Satoshi Yoshiji3,4,*, Xianyong Yin5,6, Davis Cammann7, Chi Zhao8, Henry J. Taylor9,10,11, Jingchun Chen7, Ravi Mandla3,19, Alicia Huerta-Chagoya3, Ta-Yu Yang4, Alexis C. Wood12, Kimberly M. Lorenz13,14,15, Fumihiko Matsuda4, Jason Flannick3,17,18, Josep M. Mercader3,19,20, Cassandra Spracklen8, James B. Meigs3,21,22, Jerome I. Rotter23, Marijana Vujkovic13,15,16,Benjamin F. Voight13,14,15,24, Andrew P. Morris25, Eleftheria Zeggini1,26

(1) Institute of Translational Genomics, Helmholtz Munich, Neuherberg, 85764, DE, (2) Technical University of Munich (TUM), School of Medicine and Health, Graduate

School of Experimental Medicine, Munich, DE, (3) Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, (4) Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, JP, (5) Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, CI, (6) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA, (7) Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, USA, (8) Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, MA, USA, (9) Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA, (10) British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, (11) Heart and Lung Research Institute, University of Cambridge, Cambridge, UK, (12) USDA/ARS Children's Nutrition Center, Baylor College of Medicine, Houston, TX, USA, (13) Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA, (14) Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, (15) Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, (16) Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, (17) Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA , (18) Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA, (19) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, (20) Harvard Medical School, Boston, MA, USA, (21) Department of Medicine, Harvard Medical School, Boston, MA, USA, (22) Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA, (23) Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA, (24) Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, (25) Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, M13 9PT, UK, (26) TUM school of medicine and health, Technical University Munich and Klinikum Rechts der Isar, Munich, DE

Multiple molecular mechanisms are involved in the pathogenesis of type 2 diabetes (T2D), with potentially different effects across ancestries. Recent large-scale efforts by the Type 2 Diabetes Global Genomics Initiative (T2DGGI) have broadly described the genetic architecture of T2D, data that can be subsequently used to pinpoint causal molecular mechanisms leading to T2D in an ancestry-aware manner. In this work, we sought to explore the causal effects of gene expression and proteins on T2D across ancestries by leveraging data from the latest T2DGGI multi-ancestry genome-wide association studies (GWASs). We conducted two-sample Mendelian randomization (MR) using cis-expression and protein quantitative trait loci (eQTL/pQTL) data from various datasets of four different major ancestries. To corroborate our findings, we performed statistical colocalization using PWCoCo. We then performed meta-analyses across ancestries for molecular traits showing evidence of causality in at least one ancestry (5% FDR-adjusted MR p-value and PWCoCo posterior probability ≥ 0.5).   We found causal evidence for changes in the genetically regulated expression of 267 genes and 15 proteins with T2D risk in at

least one ancestry. When meta-analyzing the results across ancestries, we found, for eQTLs, evidence of causality between 148 genes and T2D, including PABPC4, a high-confidence effector gene for T2D with evidence of ancestry-correlated heterogeneity (I2 = 85%). For pQTLs, NELL1, ANXA7, PCSK1, and CTRB2 were found to be causal to T2D risk in the cross-ancestry meta-analysis, with CTRB2 showing evidence of ancestral heterogeneity (I2 = 85%). Our findings highlight the power of large-scale GWASs and multi-omics MR to identify causal pathways involved in T2D risk. Our results also show the existence of ancestry-correlated heterogeneity, which emphasizes the need for expanding investigations into non-European ancestry populations to better understand T2D etiology.

Thursday 4rd April

0930 - 1020

LDAK-KVIK is a fast and powerful tool for performing mixed-model association analysis of quantitative and binary phenotypes

Doug Speed

Aarhus University, DK

Mixed-model association analysis (MMAA) is the preferred tool for performing a genome-wide association study, because it enables robust control of Type 1 error and increased statistical power to detect trait-associated loci. However, existing MMAA tools often suffer from long runtimes and high memory requirements. We present LDAK-KVIK, a novel MMAA tool for analyzing quantitative and binary phenotypes. Using simulated phenotypes, we show that LDAK-KVIK produces well-calibrated test statistics, both for homogeneous and heterogeneous datasets. LDAK-KVIK is computationally-efficient, requiring approximately 20 CPU hours and 10Gb memory to analyse genome-wide data for 350k individuals. These demands are similar to those of REGENIE, one of the most efficient existing MMAA tools, and about ten times less than those of BOLT-LMM, currently the most powerful MMAA tool. When applied to real phenotypes, we find that LDAK-KVIK has the highest power of all tools considered. For example, across 40 quantitative traits from UK Biobank data (average sample size 349k), LDAK-KVIK finds 17% more significant loci than classical linear regression, whereas BOLT-LMM and REGENIE find only 16% and 11% more, respectively. LDAK-KVIK can also perform gene-based tests; across the 40 quantitative UK Biobank traits, LDAK-KVIK finds 17% more significant genes than the leading existing tool.

1020 - 1040

Epigenome-wide association studies using approximate message passing

Jakub Bajzik (1), Al Depope (1), Riccardo E. Marioni (2), Marco Mondelli (1), Matthew R. Robinson (1)

(1) Institute of Science and Technology Austria, AT (2) University of Edinburgh, UK

Recent technological advances allow for a diverse range of genomic features to be measured within individuals or single cells, that are linked to electronic health information and patient outcomes. However, current state-of-the-art statistical methods for these data do not allow for reliable significance testing and become inefficient in large scales and high dimensions. Ideally, we wish to utilize multiple types of genomic data (sequence, methylation, and/or gene expression) to jointly estimate the effect of each variable on an outcome, conditional on all others, whilst controlling for potential confounding factors. Marco Chain Monte Carlo (MCMC) methods have become a standard method for this in epigenome-wide association studies, however, they are characteristically slow and association testing is difficult. Here, we propose a new method called gVAMPomi, based on the recently developed genomic Vector Approximate Message Passing (gVAMP) algorithm, which is an iterative algorithm allowing for joint inference of model parameters, while providing joint p-value testing utilizing the properties of state evolution (SE). We initialize gVAMPomi with an efficient MCMC algorithm and then estimate biological variable associations and conduct association testing. Using simulations, we show that gVAMPomi reaches MCMC performance in both out-of-sample prediction accuracy and association testing. Next, we apply gVAMPomi to the largest human methylation dataset generated to date, the Generation Scotland study, and find 92 CpG probes whose effects are significantly associated with traits, conditional on all other CpG probes, representing a significant increase over 37 CpG probes discovered by baseline MCMC approach. Overall, gVAMPomi scales to accommodate future large-scale omics data, allowing for the construction of more complex models and testing novel biological hypotheses.

1040 - 1100

Isolating the Genetics of Mania through Genomic Structural Equation Modelling

Giuseppe Pierpaolo Merola (1), Johan Zvrskovec (1, 2), Gerome Breen (1, 2)
(1) Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK, (2) National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre, South London and Maudsley Hospital, UK

Bipolar disorder, also known as manic-depression, affects 1-2% of the population and is a severe, often progressive disorder that features relapsing-remitting and cyclical episodes of mania and severe depression, frequently accompanied by psychosis. While there has been considerable success in understanding the genetics of depression and psychosis via successful studies of major depressive disorder (MDD) and schizophrenia, there has been little study of the specific genetics of mania. This is because genetic studies of bipolar disorder include genetic effects associated with all three of its primary phenotypic dimensions: mania, depression, and psychosis. Understanding the genetic effects specific to mania could contribute to the design and discovery of more targeted therapeutic interventions for bipolar disorder.   We first processed and QC'd GWAS summary statistics via the GenomicSEM (genomic Structural Equation Modelling) R package, using its implementation of the LD Score Regression (LDSR) algorithm to estimate genetic coariances, which we then used in a gSEM model based on the GWAS-by-subtraction approach. We extended the traditional GWAS-by-subtraction to allow for dual phenotype subtraction and constructed a mania latent trait by subtracting the genetic effects of MDD and schizophrenia from the genetic effects associated with bipolar disorder. The SEM analysis revealed significant and variable associations between bipolar disorder and Depression, Psychosis, and Mania, with standardized estimates/weightings of 0.486, 0.692, and 0.533 respectively. I will also present the “synthetic” GWAS of mania and its genetic correlations and alternative modelling approaches. These findings indicate strong relationships between bipolar disorder and these psychopathology dimensions.

1130 - 1210

Causal genetic effects for blood pressure management through life

Malgorzata Borczyk (1), Nick Machnik (2), Jacek Hajto (1), Michal Korostynski (1), Matthew Robinson (2)
(1) Maj Institute of Pharmacology Polish Academy of Sciences, PL, (2) Institute of Science and Technology Austria, AT

Hypertension (HP) is recognized as the leading modifiable risk factor for the global burden of cardiovascular disease and disability. Genetic variants can be used to predict the efficacy and safety of antihypertensive medications, however age-specific and drug-specific genetics of HP management remain largely unknown. In population-scale biobanks, linkage of electronic health records (EHRs) and longer participant follow-up times mean that repeated measures data are becoming increasingly available. These data make it possible to identify loci associated with HP treatment with each drug class at different stages of life, but they are complex as individuals receive a variety of medical interventions throughout their lives. Here, we show how graphical inference can be used to select genetic variants that are most likely to have a direct causal effect on a trait as well as to investigate relationships between traits. We leveraged a dataset of over 2M blood pressure measurements and associated prescriptions available in the UK Biobank (UKB) EHRs. We then inferred a graph skeleton of the repeated measurements grouped into five age groups together with genomic variables. In an analysis of both imputed genotyping and whole exome sequencing data we discovered 1400 likely causal links between tested traits and genetic variants (predicted FDR < 1%). The associations were present predominantly in the youngest age group suggesting a strong genetic component of HP treatment early in life. Among the results are the discoveries of the age-specific involvement of APOE and LDLR missense and loss of function (LoF) variants in statin therapy and of LoF variants in multiple ion channel genes that are likely causal for early-onset cardiological disorders. Finally, we explore the value of the discovered drug-associated variants in explaining individual differences in drug responses.Here we show the first of its kind exploration of longitudinal measurements of blood pressure management in a graphical modeling approach. The results indicate an important contribution of genetics to various HP-management strategies that can be further explored eventually leading to personalized approaches in HP treatment.

1210 - 1230

Novel genetic variants identified for kidney function decline and insights into genome-wide association analyses using longitudinal trait trajectories

Simon Wiegrebe, Mathias Gorski, Janina M. Herold, Klaus J. Stark, Barbara Thorand, Christian Gieger, Johannes Schödel, Florian Hartig, Han Chen, Thomas W. Winkler, Helmut Küchenhoff and Iris M. Heid
Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany. Statistical Consulting Unit StaBLab, Department of Statistics, LMU

Munich, Munich, Germany. Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany. German Center for Diabetes Research (DZD), Partner München-Neuherberg, Neuherberg, Germany. Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany. Department of Nephrology and Hypertension, Universitätsklinikum Erlangen und Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. Theoretical Ecology, University of Regensburg, Regensburg, Germany. Human Genetics Center, Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA.

Accelerated decline of kidney function is a serious health burden as it can lead to kidney failure, increases early mortality risk, and has limited therapeutic options. Yet knowledge about genetics of kidney function decline, in fact of trait dynamics in general, is limited. This is due to scarce longitudinal genome-wide association studies (longGWAS) for trait dynamics and substantial uncertainty about how to analyze them. We used longitudinal UK Biobank data (m=1,520,756 datapoints for creatinine-based estimated glomerular filtration rate, eGFR; n=348,225, age 35-80 years, newly curated from study center or electronic health records) to systematically compare multiple statistical approaches to genetic association analyses for eGFR-decline: the difference model, using annualized differences between individuals' first and last eGFR assessment as outcomes in linear regression on SNPs; a two-stage approach using random slopes (RS), estimated from a linear mixed model (LMM), as an outcome in linear regression on SNPs; or one-stage approaches using LMMs directly to model genetic decline effects via SNP-by-age or SNP-by-time interactions. Hypothesizing that genetics of eGFR-decline was a subset of genetics of eGFR, we primarily focused on 595 independent variants known for association with eGFR from cross-sectional GWAS. In simulated data and empirical analyses of the 595 SNPs, we found (i) larger standard errors (SEs) for the difference model and the time model, due to reduced power from using only two assessments per person or estimating additional parameters, respectively; (ii) deflated SEs for models with inappropriate accounting for RS variability, corroborated by type I error inflation in simulations; (iii) shrinkage in effect size estimates of the two-stage approach due to RS regularization; (iv) smaller SEs and subsequently larger power for the only model capable of incorporating individuals with =1 eGFR assessments, a one-stage LMM modeling eGFR levels over age and eGFR-decline as SNP-by-age interaction (age model RI&RS 350K), due to the increased sample size. Altogether, the age model RI&RS 350K had the best statistical properties regarding type I error, power, and unbiased effect estimates; using the GMMAT/MAGEE software packages, we made this model viable for longGWAS. We identified 13 genetic variants with independent associations for eGFR-decline (7 with Pdecline<5x10-8; +6 from focused search with Pdecline<0.05/595); 7 had previously been identified by a GWAS meta-analysis using the difference model, while 6 were identified here for the first time for eGFR decline including a novel independent association in the known eGFR-decline locus near UMOD/PDILT. We additionally identified stable-effect variants, characterized by large main effects on eGFR-levels but no association with eGFR-decline. By contrasting decline-associated to stable-effect variants, we could link genetics of eGFR-decline to age-dependent genetics of eGFR: e.g., the impact of the prominent eGFR variant rs77924615 (near UMOD/PDILT) increased from a negligible effect size at the age of 40 (beta=0.22 mL/min/1.73m2 per allele; 95%-CI=[0.08, 0.37]) to a

substantial effect at the age of 70 (beta=-1.59 mL/min/1.73m2 per allele; 95%-CI=[-1.71, -1.48]). Altogether, our results provide a systematic comparison of statistical approaches for modeling genetic effects on trait dynamics, novel insights into the genetics of kidney aging, and guidance regarding longGWAS.

1400 - 1420

Evolutionary history of disease alleles: the problem of admixture

Amke Caliebe (1), Daniel Kolbe (2), Nicolas A. da Silva (2), Janina Dose (2), Guillermo G. Torres (2), Stefan Schreiber (2), Ben Krause-Kyora (2), Almut Nebel (2)
(1) Institute of Medical Informatics and Statistics, Kiel University, University Hospital Schleswig-Holstein, DE (2) Institute of Clinical Molecular Biology, Kiel University, University Hospital Schleswig-Holstein, DE

Information about the place and time of origin of many disease alleles and their later dispersal are scarce. Two major demographic admixture events occurred in Europe since ~4000 BCE which pose serious mathematical challenges to the evolutionary analysis. We developed an admixture-informed statistical test to investigate whether the allele frequency distribution of the offspring population shows significant differences from the expected frequencies under admixture with no selection. We applied our method to the longevity and Alzheimer’s gene Apolipoprotein E (APOE) and to seven missense single nucleotide variants which have been functionally linked to inflammatory bowel disease (IBD). We derived alleles from up to 3521 published and self-generated aDNA samples from the past 12,000 years. Results The first demographic event occurred around ~4000 BCE and saw the emergence of the late farmers (LF) through the admixture of local western hunter-gatherers (WHG) with early farmers (EF). The second significant admixture event happened around 2700–2500 BCE when a large group of western steppe herders moved to Europe and mixed with the LF resulting in several offspring distributions. (i) APOE There were large differences in the APOE allele frequencies between WHG and EF. Our admixture informed allele test showed that the APOE frequencies of the LF were consistent with expected frequencies under admixture. Following the expansion of steppe herders into Europe, APOE frequencies of European populations became more uniformly distributed. Interestingly, the frequencies of the ε2 allele increased in all European common era populations and significantly differentiated from their admixture-informed expected frequencies. (ii) IBD For the seven missense variants the admixture-informed tests yielded heterogeneous results. The frequency of some variants did not differ significantly from the expectation indicating that these frequencies were likely the result of a mixture between the parental populations. However, for other variants we noted clear deviations between observed and expected frequencies for several populations suggesting that admixture alone could not explain the allele frequency differences between populations. It is thus likely that selection, rather than admixture, played a larger role for these variants. Implications The large APOE frequency differences between WHG and EF is possibly due to changes in diet/lifestyle. In contrast, the allele distributions in populations from ~4000 BCE onward can mainly be explained by admixture. The resulting allele frequencies strongly influence the predisposition for longevity today, likely as a consequence of past adaptations and demographic processes. The statistically significant differences in the frequencies of IBD variants between Neolithic and modern populations can be explained by the adoption of an agricultural lifestyle and behaviour and concomitant possible microbiome changes in the earliest farmers. Later admixture events and selection against pathogens largely influenced the genetic risk architecture of IBD in contemporary Europeans. A better understanding of the evolutionary history of disease variants is an important first step in translating genetic findings into preventive health care.

1420 - 1440

A General Framework for Population Genetic Inference via Generalised Linear Models

Lynette Caitlin Mikula, Burçin Yildirim, Claus Vogl
(1) University of St Andrews, UK (2) University of Veterinary Medicine Vienna, AT

Introduction: Development of efficient methods for population genetic inference has become increasingly important with the growing availability of both genome-wide allele frequency data and the computational resources required to process large data quantities. The challenge has now become reconciling the abundance of new empirical observations with older modelling approaches, and expanding on these models while ensuring that they remain well-founded in both mathematical and evolutionary theory. Crucial to this is an understanding of the properties of the estimators employed to infer population genetic forces. Scientific Question: It is often desirable for an estimator to be a function of a sufficient statistic, which captures all of the information in the data relevant to estimation of the parameter. A notable example is the number of segregating sites in the infinite sites model, which is a sufficient statistic for the overall scaled mutation rate under the assumptions of mutation-drift equilibrium and free recombination. We have previously shown that in equilibrium K-allele models with low scaled overall mutation rates, the site frequency spectrum (SFS) is comprised of a set of sufficient statistics for different mutation parameters. While all mutation parameters can be estimated via maximum-likelihood (ML), those that map directly to sufficient statistics are also minimum-variance unbiased estimators (MVUE, which are optimal unbiased estimators). Here, we wish to find similar estimators for population genetic forces such as biased gene conversion or different modes of selection. Results: Treating the SFS as the response variable, equilibrium K-allele models with low scaled overall mutation rates can be expressed as a Generalised Linear Model (GLM) of the Poisson family with a log-link function. Linear and quadratic forces shaping the spectra, due respectively for example to GC-biased gene conversion or directional selection and to overdominance or balancing selection, can be incorporated into this GLM as covariates. In equilibrium, ML estimators for the directional and quadratic forces can be estimated in addition to those for the scaled mutation parameters, and the estimators are at least asymptotically MVUE. Importantly. hypothesis tests can be constructed to test not only for the presence of these linear and quadratic forces in equilibrium, but also for deviation from drift-mutation-selection equilibrium, eg due to demography. Implications: This population genetic GLM framework is a powerful approach for querying the population genetic forces shaping the SFS of specific genomic regions. It allows for disentangling of confounded forces such as demography and selection. We demonstrate these principles on simulated data as well as on SFS from Drosophila.

1440 - 1500

Quality control of GWAS summary statistics

Florian Privé
Aarhus University, DK

Results from genome-wide association studies (i.e. GWAS summary statistics) have been extensively used in different applications such as estimating the genetic architecture of complex traits and diseases, identifying causal variants with fine-mapping, and predicting complex traits with polygenic scores. One reason behind the popularity of GWAS summary statistics is that they are widely available and shared, e.g. in the GWAS Catalog. However, these GWAS summary statistics come with varying degrees of quality, and from many different tools and studies. I'll present several different cases that could go wrong when using GWAS summary statistics. I'll then present two complementary quality control steps that can be performed to mitigate these issues. Most of these are still unknown to many people using GWAS summary statistics on a regular basis, which can cause results they derive to be biased or suboptimal.

1500 - 1520

NMR metabolomics data as a powerful predictor of lifespan: challenges of modeling and interpretation in the Estonian Biobank cohort

Mara Delesa-Velina, Krista Fischer
University of Tartu, EE

There has been a great interest in studying nuclear magnetic resonance (NMR) metabolomics data as predictor of overall mortality and proxy for biological (or metabolomic) age in biobanks data. As the total no of participants with the NMR data in the Estonian Biobank exceeds 200000, with the mean follow-up time of 15 years for the first 50000 participants and 4 years for the newest 150000, we are interested to develop a model for all-cause mortality based on NMR data for the Estonian Biobank. The participants have been recruited in two large waves in 2002-2010 and in 2018-2019, using somewhat different recruitment strategies. This resulted in some notable differences between the two subcohorts in the distribution of baseline socio-demographic variables as well as in the prevalence of chronic diseases and also in the overall risk for all-cause mortality and incident diseases. These differences pose a challenge to survival modeling: the two cohorts cannot be analyzed together. However, when analyzed separately, the two cohorts produce models that lead to very different predicted risk levels for the same individual (if used for out-of-sample predictions). Instead of either making a choice between the two cohorts or producing some average risk estimates based on the two, we propose to use the data differences for model validation. We develop the survival model based on the subset of the first wave NMR records and validate it in the second wave data. We show that although the absolute risk estimates differ in the two cohorts, the predictions based on the biological age difference are still similar. Thus we conclude that one should find alternatives to absolute risk estimates, as the risk, interpreted as a “fraction of individuals getting the disease” would always depend on the particular subset of a population, represented by the given biobank cohort.

1600 - 1650

Fine-mapping GWAS meta-analyses with migraine as an example

Matti Pirinen
University of Helsinki, FI

The most informative disease GWAS tend to be meta-analyses prvoiding only summary statistics for downstream analyses. This causes problems for statistical fine-mapping because accurate linkage disequilibrium information is often missing and not all variants are available in all component studies. I will first present results of a fine-mapping study of a migraine meta-analysis with 98,000 cases and highlight ways to assess the reliability of the results. Then I will discuss fine-mapping approaches to account for varying missingness across variants in meta-analysis summary statistics.

Selected poster presentations

Poster 1

Multi-trait genetic colocalization of urine protein levels provide insights into renal protein handling and related clinical outcomes

Oleg Borisov*, Stefan Haug, Nora Scherer, Sara Monteiro-Martins, Yong Li, Anna Köttgen
Institute of Genetic Epidemiology, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Germany

Genetic association studies of protein levels provide a valuable link between molecular processes and human phenotypes including diseases. We generated urine proteomics data and integrated it with multiomics datasets and phenome-wide association studies to elucidate genetic mechanisms of protein handling in the kidney and its connection to diseases. We performed genetic colocalization analyses („coloc“ package) to assess whether urine protein levels measured in the German Chronic Kidney Disease study (N>1000) share common genetic architecture with orthogonal omics and clinical datasets. We colocalized urine protein genetic associations with multi-tissue transcriptomics, plasma proteomics, and clinical outcome datasets (UK Biobank and FinnGen). We performed 3.5 million colocalization tests in more than 100 genetic loci and identified 9,000 pairs of traits with strong colocalization evidence (posterior probability H4>80%). The majority of loci (>90%) showed evidence of multi-trait colocalization (up to 1000 colocalizing traits per protein). One example was Prostate Stem Cell Antigen (PSCA) – where cis associations (P<1e-300) with urine PSCA levels colocalized with PSCA expression in kidney tissues and urogenital clinical manifestations, including bladder cancer and urinary tract infections. We report colocalization results between urine protein associations and phenome-wide association studies including molecular and clinical traits. The generated resource will facilitate a deeper understanding of genetic factors influencing protein handling by the kidney and will help to reveal circulating disease biomarkers which are also informative when quantified in urine. This work was supported by the German Research Foundation (DFG) project ID 431984000 (SFB 1453).

Poster 2

Genetic Overlap Between Attention-Deficit /Hyperactivity Disorder and Migraine

Pau Carabí Gassol (1,2,3), Natàlia Llonga (1,2), Uxue Zubizarreta Arruti (1,2,3), Valeria Macías (1,2), Silvia Alemany (1,2), Chiristian Fadeuilhe (1,2,3,5), Montse Corrales (1,2,3,5), Vanesa Richarte (1,2,3,5), Josep Antoni Ramos-Quiroga (1,2,3,4,5), Marta Ribasés (1,2,3,4), Judit Cabana-Domínguez (1,2,3,6), María Soler Artigas (1,2,3,4,6)

(1) Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, ES, (2) Department of Mental Health, Hospital Universitari Vall d'Hebron, Barcelona, ES, (3) Biomedical Network Research Centre on Mental Health (CIBERSAM), Madrid, ES, (4) Department of Genetics, Microbiology, and Statistics, Faculty of Biology, Universitat de Barcelona, Barcelona, ES, (5) Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Barcelona, ES, (6) These authors jointly supervised this work

Background: Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder that emerges in childhood and often persists into adulthood. Migraine is one of the most common neurologic disorders, with a high prevalence and morbidity. Previous studies have reported an association between ADHD and migraine and showed that individuals with headaches and ADHD have lower quality-of-life and higher incidence of other psychiatric disorders than healthy individuals. Since both ADHD and migraine are heritable, we aim to study their genetic overlap and potential causal relationship using Mendelian randomization. Methods: We used data from the largest genome-wide association meta-analyses to date of ADHD (38,691 cases and 186,843 controls) and migraine (102,084 cases and 771,257 controls) to: (i) estimate their genetic overlap (LDscore and MiXer), (ii) perform cross-trait analyses to identify shared genetic variants by combining two methods: pleioFDR and PolarMorphism, (iii) use the polygenic risk score (PRS) of migraine to study differences between individuals with and without migraine or headache in an independent in-house clinical sample of 930 deeply phenotyped individuals with ADHD and (iv) perform a bidirectional causality analysis using a two-sample Mendelian randomization approach. Results: We confirmed a robust positive genetic correlation between the two traits (rg=0.205; SE=0.032; P=2.38E-10). Besides, we found that ADHD is much more polygenic than migraine (7729 variants, SE=363 and 1731 variants, SE=91, respectively), and both traits share 951 variants, most of them with concordat direction of effect (84%), according to MiXer. Cross-trait analyses with pleioFDR and PolarMorphism identified 15 independent loci jointly associated with both ADHD and migraine,13 of which were novel for both traits. An intronic variant (rs4856605) in CADM2, which encodes a member of the synaptic cell adhesion molecule 1 (SynCAM) family, was among the most significant signals. PRS for migraine were associated with childhood headache (CH) (OR=1.22, P=0.008) in our ADHD clinical sample (n= 314 ADHD cases with CH and 459 ADHD

cases without CH). The comparison of these two groups revealed nominally significant differences in depression (Beck BDI-II dispersion test), anxiety (SATI and ZKPQ), neuroticism (ZKPQ) and cognitive functioning (FAST), having the individuals with CH more risk or severity. However, none of them exhibited association with the PRS of migraine.   MR analyses showed evidence of a positive causal effect of ADHD genetic liability on migraine (IVW B=0.077 and P=0.036) consistent across sensitivity analyses, but no evidence of causality in the opposite direction. Conclusions and further work: Our findings indicate that ADHD and migraine differ in their genetic architecture but share a common genetic background. Cross-sectional analyses allowed the identification of 13 new genome-wide significant variants associated to both traits. Furthermore, the genetic liability of migraine as a PRS was associated to childhood headache in ADHD individuals and the genetic liability of ADHD may have a causal effect on migraine. Further work includes pathway analyses to understand the biological mechanisms underlying this genetic overlap and the incorporation of pleiotropic variants to the migraine PRS aiming at improving its performance in ADHD.

Poster 3

Comparative analysis of genome-based computational models for prediction of drug-metabolising enzymes function

Jacek Hajto (1), Małgorzata Borczyk (1), Marcin Piechota (1), Gabriel Boyle (2), Douglas Fowler (2), Michał Korostyński (1)
(1) Laboratory of Pharmacogenomics, Department of Molecular Neuropharmacology, Maj Institute of Pharmacology, Polish Academy of Sciences, Krakow, PL, (2) Department of Genome Sciences, University of Washington, Seattle, USA

In precision medicine, it is important to understand the impact of gene variants on drug response, particularly within drug-metabolising enzymes (DME), which play a critical role in individual drug efficacy and safety. Historically, variation within these genes has been represented through haplotype-based star allele nomenclature, where each haplotype gets a consecutive number and is assigned a consequence of the enzymatic function. This approach is difficult to continue with an ever-growing pool of novel variants and lacks mechanisms to include rare mutations.   Therefore, robust computational methods for accurate variant consequence prediction are needed. Furthermore, as genes encoding DME show higher variability than most other protein-coding genes, existing variant predictors may not be suitable for this gene class. Here, we systematically compared several available computational methods, including variant prediction scores such as CADD, MutationAssessor, FATHMM-XF, and the ADME Prediction Framework (APF). We also propose two entirely novel approaches to improve the current predictors: a machine learning-based (PharmMLScore) and a knowledge-based model (PharmGScore).   The performance of all the approaches was evaluated against a curated database of pharmacogenetic haplotypes with known functional consequences (PharmVar). In this dataset, the PharmGScore approach had the best performance (AUC = 0.86) in comparing alleles assigned with no and normal function of the protein.   We further evaluated the predictors using additional approaches at different levels of complexity. Firstly, we analysed all possible missense variants of two major pharmacogenes, CYP2C9 and CYP2C19, from a large dataset of enzymatic activity and protein abundance levels obtained from a massively parallel variant characterization. Secondly, we leveraged the whole exome sequencing (WES) data from the UK Biobank, correlating computed scores for 200k patients with haplotypes identified by existing star allele callers. Contrary to the previous reports, we show that the tested computational approaches correctly identify haplotypes encoding nonfunctional copies of DME. In conclusion, the proposed computational methods, particularly the novel PharmGScore, can accurately describe genetic variants with numerical values, enabling their use in statistical models and estimating the functional impact of previously uncharacterized genetic variants. This approach to variant interpretation could advance the pharmacogenomics field and enhance existing clinical guidelines by including more of the relevant genomic information in models that predict individual drug responses.

Poster 4

Pleiotropic effects of pathway-partitioned asthma genetic risk scores in UK Biobank

Matthew J. Saward (1), Robert J. Hall (2), Liam G. Heaney (3), Ian Sayers (2), Katherine A. Fawcett (1)

(1) Department of Population Health Sciences, University of Leicester, Leicester, UK; (2) Centre for Respiratory Research, National Institute for Health Research Nottingham Biomedical Research Centre, School of Medicine, Biodiscovery Institute, University of Nottingham, Nottingham, UK; (3) Wellcome-Wolfson Centre for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK

INTRODUCTION: Asthma is a chronic, respiratory disease, with many known genetic and environmental risk factors. Clinical presentation varies amongst patients, and these differences are thought to be due, in part, to different underlying pathobiological mechanisms. We sought to partition the genetic component of asthma risk into different biological pathways in order to investigate their pleiotropic effects on a broad range of respiratory and non-respiratory traits. METHODS: We conducted a literature search for variants associated with asthma and confirmed association with asthma in UK Biobank participants (54,672 cases vs 341,965 controls). Independent signals were selected and mapped to genes using Functional Annotation and Mapping (FUMA), based on variant position, association with gene expression and chromatin interaction. We used Ingenuity Pathway Analysis (IPA) to identify enriched biological pathways, applying a Benjamini-Hochberg correction. All significantly associated (adjusted $p < 0.05$) pathways were subsequently used to calculate pathway-partitioned genetic risk scores and these scores were tested for trait association (~1900 traits) in a phenome-wide association study in UK Biobank. RESULTS: In total, ~1100 variants representing 255 independent signals were identified and subsequently mapped to 1666 genes. IPA found these genes were enriched in 17 canonical pathways, particularly those involved in cytokine signalling and inflammatory responses. Most of the pathways showed association with a Type 2 high inflammatory profile: high eosinophil count and positive association with nasal polyps, as well as negative association with lung function. Two pathways (C-type lectin receptors and Keratinization) however were most strongly associated with neutrophil counts (FDR = 9.19 x 10-232 and 1.08 x 10-178 respectively), suggesting a more mixed granulocytic profile. The interleukin-15 signalling pathway was the only pathway to show stronger association with adult-onset asthma compared to childhood-onset asthma as well as increased risk of hypothyroidism (FDR = 2.15 x 10-7). CONCLUSIONS: Pathway-partitioned genetic risk scores for asthma show distinct associations with respiratory traits and other comorbidities, suggesting a genetic basis for different clinical presentations of asthma. These data provide opportunities for new drug development and a more tailored approach to asthma management. We will seek to validate these findings in additional cohorts with more detailed clinical data.

Poster 5

Genome-wide analysis of school performance and overlap with psychiatric and cognitive traits

María Soler Artigas (1-4), Silvia Alemany (1-3), Judit Cabana-Domínguez (1-3), Rosa Bosch (3,5), Laura Vilar-Ribó (1-3), Natalia Llonga (1-3), Pau Carabí Gassol (1-3), Valeria Macias Chimborazo (1-3), Uxue Zubizarreta Arruti (1-3), Josep Antoni Ramos-Quiroga (1-3,6), Mireia Pagerols (5,7), Raquel Prat (5,8), Elia Pagespetit (5), Cristina Rivas (5), Gemma Español-Martín (1-3), Miquel Casas (5), Marta Ribasés (1-4)

(1) Psychiatric Genetics Unit, Group of Psychiatry Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, ES, (2) Department of Mental health, Hospital Universitari Vall d'Hebron, Barcelona, ES, (3) Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, ES (4) Department of Genetics, Microbiology, and Statistics, Faculty of Biology, Universitat de Barcelona, Barcelona, ES (5) SJD MIND Schools Program, Hospital Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, ES (6) Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Barcelona, ES (7) Unitat de Farmacologia, Departament de Fonaments Clínics, Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona (UB), Barcelona, ES, (8) Centre for Health and Social Care Research (CEES), University of Vic−Central University of Catalonia (UVic−UCC), Vic, ES

Background: Access to education is considered to be a predictor for a wide range of later life outcomes such as employment, income and health outcomes. Performance in primary school is a determinant of children's educational attainment, their socioeconomic position and health variability in adulthood. There is also evidence that individuals with mental disorders exhibit remarkable differences in their school performance. In addition, twin studies support substantial genetic influence on school performance in children. In this study we investigated the genetic factors underlying school performance and their overlap with the genetic predisposition of different psychiatric traits within the INSchool cohort. Methods: Genome-wide association studies (GWAS) were undertaken on English, primary language and mathematics scores for 4,278 children aged 6-18 years (average age 9.8 years; 57% males) from 50 different schools in Catalunya, using Haplotype Reference Consortium imputed data and ordinal regression including age, sex, socioeconomical status, school and significant ancestry principal components as covariates with the OpenMendel analysis package. Genome-wide polygenic scores (PGS) were generated with the latest summary statistics available for bipolar disorder (BP), depression,

schizophrenia (SCZ), attention deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), dyslexia, intelligence, educational attainment (EA), cognition and EA portioned by cognitive (EA_cog) and non-cognitive factors (EA_noncog) using PRScs and plinkv1.9. The same ordinal regression models as in the GWAS were used to test the association between PGS and school performance. A Bonferroni correction was applied to account for the different PGS and subjects. Disorder diagnosis, when available, was included in the model as an additional covariate to estimate the direct effect of the PGS, not mediated by diagnosis, on school performance. Results: The most significant signal for the GWAS on English performance (rs7689040; P=1.1x10-7) was an intronic variant in TENM3 gene, which encodes a protein suggested to be involved in the regulation of neuronal development. The top variant (rs9350070, P=2.81x10-7) for mathematics was an intronic variant in RNF144B, gene associated with self-reported educational attainment and for primary language (rs985280, P=1.7x10-7) it was an intronic SNP in PRICKLE2, gene associated with brain measurement and educational attainment. PGS for dyslexia were associated to worse school performance in all three subjects, and PGS for ADHD and SCZ only for worse performance in English and mathematics. PGS for intelligence, EA, cognition and EA_cog were associated with better performance in all subjects, whereas EA_noncog was associated only with better performance in mathematics and primary language. The effect of the PGS for ADHD and dyslexia were reduced when accounting for the diagnosis in the model, although the PGS remained significant in most cases.   Conclusions: Despite no genetic variant showing genome-wide significance, our preliminary findings highlighted promising genes involved in cognition, education attainment and multiple psychiatric disorders. PGS for EA, cognition and intelligence were associated with better school performance. The genetic liability for ADHD, dyslexia and SCZ showed a negative effect on school performance. ADHD and dyslexia diagnosis partially mediate the effect of the PGS.

Poster 6

Assessment of a genetic risk score for lung function in prediction of non-recovery in COVID-19

Anne F Goemans (1), Olivia C Leavy (1,2), Beatriz Guillen-Guio (1,2), Erola Pairo-Castineira (3), Konrad Rawlik (3), J. Kenneth Baillie (3), Amisha Singapuri (2), Rachael A Evans (2), Christopher E Brightling (2), Louise V Wain (1,2), Tim CD Lucas (1), on behalf of PHOSP-COVID Collaborative Group

(1) Department of Population Health Sciences, University of Leicester, UK, (2) The Institute for Lung Health, NIHR Leicester Biomedical Research Centre-Respiratory, University of Leicester, Leicester, UK, (3) Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, UK

Background: Predicting those at highest risk of ongoing symptoms and impact on quality of life after COVID-19 remains challenging. Individuals who were hospitalised with COVID-19 with two or more comorbidities are at a higher risk of non-recovery. Other risk factors for non-recovery include age, female sex, ethnicity, deprivation index, smoking status, and BMI. We hypothesised that incorporating genetic risk for organ impairment might improve prediction of those at risk of poor recovery, and be informative about the relationship between pre-existing conditions and longer-term recovery. Aims: As COVID-19 initiates a respiratory infection, we calculate a genetic score for lung function using a previously developed GRS to determine if genetic risk for poor lung function is associated with non-recovery post COVID-19 in hospitalised patients. We use this GRS to test the hypothesis that using this genetic score improves predictive ability when added to a clinical prediction model of non-recovery. Methods: Using data from the post-hospitalisation COVID-19 study (PHOSP-COVID) and k-fold cross validation, we created a clinical prediction model for recovery status from COVID-19 at 12 months (526 cases, 528 controls) with predictors chosen from risk factors identified in the literature. We calculated individual genetic scores for lung function with PRSice-2 using 442 identified genetic signals and summary statistics from a large scale publicly available GWAS. We used AUROC methods to assess any differences in predictive ability with the addition of the genetic score. Results: The clinical model with age, sex, ethnicity, deprivation, smoking status, BMI, number of comorbidities, COVID-19 severity, and length of hospital admission had an AUC of 0.696 (95%CI: 0.664 – 0.727). Though the genetic score was not associated with overall recovery at 12 months (OR 0.004, CI: 0.000 – 23.041, p-value: 0.210), adding the genetic score and first 10 genetic principal components to the clinical model modestly improved discriminative ability (AUC: 0.715; CI: 0.685 – 0.746, p-value: 0.002) compared to the clinical predictors alone. Conclusion: This work suggests that inclusion of genetic risk scores may improve clinical prediction models for non-recovery in previously hospitalised COVID-19 patients. Further work to investigate these models with more specific cardiorespiratory outcomes is ongoing.

Poster 7

Trait genetic architecture and population structure determine model selection for genomic prediction in natural populations.

Patrick Gibbs, Jeff Paril, Alex Fournier-Level
The University of Melbourne

In plants, genomic prediction (GP) is used to predict agronomically relevant traits including crop yield, phenology, or the concentration of minerals and metabolites, which all have different genetic bases. GP models are fitted to training populations designed to maximize diversity, including genotypes with different evolutionary histories. Therefore, critical in GP is choosing the most appropriate model for a trait's distribution of genetic effects and the population's allele frequencies. The number of genome-wide markers being much larger than the number of observations, GP is canonically approached with penalized regression. However, an additive model may not be optimal for all genetic architectures and ancestral populations. Machine learning (ML) is often proposed to improve genomic prediction as it can model more complex biology such as epistasis, but has so far largely underperformed penalized regression. Nevertheless, simulation studies suggest ML approaches might be well suited for simpler traits controlled by few causal loci with strong epistatic effects. Molecular traits like the concentrations of minerals/metabolites, are agronomically important often displaying a simple genetic architecture while organismal traits such as yield combine many endophenotypes and tend to be complex. Beside biological complexity, differential selection among ancestral populations is also important in model selection. Population-specific selection on a trait causes covariance between trait value and population structure, confounding genetic effect estimation. Our study tests (1) the sensitivity between GP model and trait ontology, (2) how ancestral population specific selection on different traits affects prediction and model choice. Across 60 quantitative traits in Arabidopsis thaliana with ~1000 observations throughout Europe, penalized regression, random-forest and multilayer perceptron performance was assessed. Regression models were the most accurate except for a subset of molecular traits where random forest performed best. Next, we tested how each trait covaried with population structure, finding that complex organismal traits, particularly those related to flowering and yield, were accurately predicted with a few principal components explaining no more than 30% of genomic variance. However, molecular traits could not be predicted from a low dimensionality representation of genome structure, and rather necessitated resolution of individual markers. Finally, we showed that most of the complex organismal traits covaried geographically, particularly with longitude, while molecular traits have less covariance. This study informs the applicability of GP both in terms of modeling framework and design of training populations. First, showing that ensemble approaches can be particularly suited to simple, typically molecular traits is relevant to breeding crops with improved uptake of minerals and nutrients. Next, our results show that traits related to flowering and yield ontologies covary with population structure, likely due to differential selection, while biochemical traits tend to be less structured. Covariance between phenotypes and population structure confounds causal loci. We contend that differential ancestral selection on individual traits should be considered in the design of association panels and training populations. Moreover, covariance between complex traits and population structure enables simple prediction models to be effective, providing an additional explanation to why linear models often dominate ML in GP problems.

Poster 8

Landscape of effects of gene expression on immune-mediated diseases.

Sodbo Sharapov (1), Arianna Landini (1), Nicole Soranzo (1-3), Nicola Pirastu (1)
(1) Human Technopole, Milan, IT, (2) Wellcome Sanger Institute, Human Genetics Department, Hinxton, UK, (3) University of Cambridge, British Heart Foundation Centre of Research Excellence, Cambridge, UK

Most genome-wide association signals are suggested to act through gene expression. However, despite the extensive expression catalogs it has been difficult to attribute disease loci to these eQTLs. Here we aimed at exploring the contribution of eQTLs to disease in terms of number, primary vs secondary hit, and effect size distribution. To test for pleiotropy between eQTLGen cis-eQTLs and immune-mediated disease's loci, we used leave-one-out conditional analysis coupled with Bayesian colocalization. We applied Mendelian randomisation to estimate the effect of gene expression on disease. Of the 139 significant disease loci overlapping with at least one eQTL we were able to colocalise 51 (36,7%). We compared if the colocalization could be attributed to the top hit or to secondary conditionally independent ones, the latter were 3.61x more likely to be responsible. Despite eQTLs of large effect have been described, when colocalized with a disease, the effect was generally below 0.4 standard deviations of gene expression, suggesting that they may be constrained due to their effect on health. Finally, we have identified several loci where eQTLs for multiple genes simultaneously colocalised with the disease locus, suggesting a common local regulation of their expression. We show that disease loci are most likely driven by secondary association signals which can be detected when properly accounted for. The distribution of eQTLs puts an upper boundary to the disease relevant effect sizes that will be useful to plan future, larger eQTL studies.

Poster 9

Monocyte count and idiopathic pulmonary fibrosis: a bi-directional two-sample Mendelian randomisation analysis

Dominic Sayers, Richard Allen and Olivia Leavy
University of Leicester, UK

Introduction: The cause of Idiopathic Pulmonary Fibrosis (IPF) is unknown, however, many studies have identified an association between a higher monocyte count and IPF. To identify whether this association is causal a bi-directional Mendelian randomisation (MR) analysis was undertaken. MR is a commonly used analysis in genetics, in which specific genetic variants (instrumental variables) are used to assess causality between an exposure group and an outcome group. If causality is found then the monocyte pathway could be a new target for the development of future treatments and interventions. Methods: Instrumental variables (IVs) were selected from previous GWAS studies on IPF and monocyte count separately. From the IPF study, 16 SNPs were used for the IPF-to-monocyte count direction and from the monocyte count study, 356 SNPs were used for the monocyte count-to-IPF direction. The inverse variance weighted (IVW) fixed effects method was used for the primary analyses. To test the MR assumptions made, sensitivity analyses were carried out using the IVW multiplicative random effects, weighted median, weighted mode, MR Egger and the robust adjusted profile score (RAPS) method.. Cochran’s Q-statistic and the I^2 statistic were performed to check for heterogeneity in the study and a leave-one-out analysis, F-statistic and the Steiger directionality test were performed to check the strength and validity of the IVs chosen. A multivariable MR including other blood cell counts (Basophil, Eosinophil, Lymphocyte and Neutrophil) was also conducted to see if this altered the relationship between monocyte count and IPF. Results: The primary analysis showed significant causal effect estimates OR=0.800, 95% CI: 0.711-0.901, p=2× 10^(-4) and OR=0.993, 95% CI: 0.990-0.996, p=3 × 10^(-7) for monocyte count to IPF and IPF to monocyte count respectively. The sensitivity analyses showed four significant effect estimates for monocyte count to IPF but only two significant effect estimates for the other direction. The multivariable MR showed no significant causal effect for monocyte count onto IPF (OR=0.813, 95% CI: 0.549-1.204, p=0.3019). None of the other four blood cell counts showed evidence for causality in the full model. Discussion: The primary univariate analyses showed that there was significant evidence that a lower monocyte count was causal of IPF, however, it also showed that there was evidence that IPF was causal of a lower monocyte count. For the IPF to monocyte count direction this causal effect was not backed up by most of the sensitivity analyses and suggests this association may not actually be causal. However, for the monocyte count to IPF direction this causal result was backed up by the sensitivity analyses, with only the weighted mode having a non-significant OR.

However, the multivariable analyses for this direction suggested that this causal effect may be due to horizontal pleiotropy

Poster 10

Personalized prediction of the risk of Type 2 Diabetes in the Estonian Biobank cohort.

Karmel Teder, Natalia Pervjakova, Kristi Läll, Lili Milani, Krista Fischer
Institute of Mathematics and Statistics, University of Tartu

As the total number of participants in the Estonian Biobank (EstBB) cohort exceeds 200,000, the sample size is sufficient for developing accurate risk prediction models. Since 2024, the participants are getting feedback on their predicted risk level for some common chronic diseases, including Type 2 Diabetes (T2D). We present the T2D model in the EstBB data and describe the process of developing an algorithm for absolute risk prediction.   As Cox proportional hazards model is not providing closed-form formulas for risk prediction (baseline hazard needs to be estimated nonparametrically), we decided to use Weibull model. We show that for our purpose - prediction of the 10-year risk, the performance of the Weibull model is adequate and comparable to the corresponding Cox model.   Separate models were fitted for individuals younger than 60 and for those of age 60 or older. As for older age groups, the issue of competing risks complicates the absolute risk estimation, the feedback on absolute risks is currently only provided to the younger age group.   As expected, the main predictors of the T2D risk are related to obesity (BMI, waist circumference), whereas some prevalent diseases (myocardial infarction, hypertension) and smoking level appear to be predictive as well. One of the strongest predictors is the Polygenic Risk Score (based on Mars et al., 2022), with the effect of one standard deviation difference in PRS   corresponding to a Hazard Ratio of about 1.6 in both younger and older age groups. In this presentation, we will demonstrate the entire process from model-fitting to the actual visual feedback provided to the participants.

Poster 11

Optimal effect estimation in genome-wide association studies with censored biomarker measurements in large-scale biobanks

Yaqi Deng, Åsa Johansson, Torgny Karlsson
Department of Immunology, Genetics, and Pathology, Uppsala University, SE

Background: In recent years, high-throughput technologies have enabled measurements of molecular phenotypes, including proteomics and other omics data types, in large-scale studies, positioning multi-omics at the forefront of the genetic epidemiology field. Genome-wide association studies (GWAS, Mendelian randomization (MR), and polygenic risk scores (PRS) are important approaches that are widely used to comprehend the relationships between genetic variations, e.g., single-nucleotide polymorphisms (SNPs), protein levels and disease traits. However, technical constraints in measurements such as noise or limit of detection (LOD), where levels below LOD are dominated by noise or cannot be measured, may lead to biased data distributions. Common strategies to handle observations below LOD include removal, assignment to a designated value, e.g., zero or LOD, or retainment (if available) regardless of a primary reflection of noise. Various regression methods have been applied in GWAS using data impaired with LOD. The choice of model directly impacts SNP-effect estimates and can potentially influence downstream analyses including MR and PRS estimations. Nonetheless, a systematic evaluation of the optimal strategy to estimate effect sizes in GWAS for censored phenotypes is lacking. In this study, we address the gap by comparing the performance of different data processing strategies and various models, including linear, Tobit, Cox, and logistic regression, for censored phenotypes. Methods: We simulated genotypes and normally distributed phenotypes with varying proportion of phenotype measurements below LOD, minor allele frequencies (MAF) and effect sizes using 1000 replicates for each combination of parameters. To assess the performance of different data processing and modeling strategies in GWAS, the sensitivity of detecting associations was calculated. For the linear and Tobit models, we also calculated relative errors in the estimated effect. Results: For the linear model, assignment of observations below LOD to the average value below LOD, as imputed by the rank-based inverse normal transformation yielded the highest sensitivity and accuracy among the three LOD-handling strategies. Sensitivities were found to be comparable and there was a high concordance between the linear, Tobit, and Cox models. The logistic model, in which the phenotypes were recoded as below (‘0’) or above (‘1’) LOD, exhibited the lowest sensitivity and a low concordance with the other models. Benchmarking showed that the linear model was computationally the most efficient, followed by Cox, logistic, and Tobit. As expected, Tobit outperformed the linear model in accuracy of estimated effects. Conclusion: At present, the computational inefficiency of the Tobit model on biobank-scale data is a burden for its implementation into the GWAS toolbox. To reach optimal precision, accuracy, and efficiency, we hence propose that the linear model, with below-LOD values set to the average of the corresponding rank-transformed values below LOD, should be used for discovery analyses of significant SNPs. The identified signals should then be reanalyzed using the Tobit model to obtain unbiased effect estimates for post-GWAS analyses such as MR or PRS construction.

Poster 12

Genome-wide pQTL analysis in the EXCEED study using data-independent acquisition mass spectrometry

B. Lim (1), N. Shrine (1), A.L. Guyatt (1), C.B. Maxwell (2), D.J.L. Jones (2), M.D. Tobin (1)

(1) Department of Population Health Sciences, University of Leicester, UK, (2) The Leicester van Geest MultiOmics Facility, Hodgkin Building, University of Leicester, UK

Introduction: Genetic variation affects both the abundance and structure of proteins in the body. The identification of protein-quantitative trait loci (pQTLs) permits the study of causal disease mechanisms, which may in turn inform our understanding of potential drug targets. A recently released platform for ultra-high-throughput serum proteomics uses data-independent acquisition mass spectrometry (DIA-MS) to measure blood protein biomarker quantities (1). DIA-MS identifies and quantifies proteins in an untargeted fashion, allowing global, unbiased protein profiling potentially uncovering proteins that would not be identified in targeted methods, such as antibody-based methods.   We generated genome-wide associations across protein groups in the UK-based EXCEED cohort, in order to construct a pQTL mapping resource.   Methods: Protein quantification has been undertaken in serum extracted from blood prick samples of 1,918 EXCEED participants, using high-throughput DIA-MS (1). Protein groupings (N=319) were identified and retention time normalised using Data-Independent Acquisition-Neural Network (DIA-NN) software v1.3.1 (2). Batch effects were visually assessed using protein group-specific boxplots and UMAP, and then formally assessed using an ANOVA test of proteomic principal components. We performed genome-wide association studies of single-nucleotide polymorphism (SNP)-protein associations (under an additive model) across 319 protein groups, in up to 641 participants with genomic data, using REGENIE v3.4.1. Protein group quantities were first adjusted for age, sex and batch effects. Residuals were then rank inverse-normal transformed. Ancestry principal components were included as covariates during association testing. We annotated SNP-protein associations with P<5×10−8. We defined cis-pQTLs as those on the same chromosome as their cognate gene, and within 1Mb of the transcription start site. Otherwise, associations were defined as trans-pQTLs. Results We found strong evidence of batch effects (ANOVA P=2.51×10−54), necessitating adjustment in association analyses. In total, there were 88 SNP-protein associations at P<5×10−8, grouped into 25 loci (defined as lead SNP ±1Mb). These associations were across 23 different protein groups. All SNP-protein associations were annotated as trans, and all but one (APOF; Q13790) are on a different chromosome. Implications and discussion: We present initial pQTL mapping across 319 protein groups in 641 individuals within EXCEED using a high-throughput DIA-MS method. We are currently genotyping additional EXCEED participants, which should improve statistical power for detection of pQTL associations. Future work will include comparing the associations found in the EXCEED pQTL resource to other population-based pQTL resources generated on other platforms.

Poster 13

Correcting for fine-scale population structure in genome-wide association studies: the French POPGEN Project as a proof-of-concept

Gaëlle Marenne (1), POPGEN Study group (2), Anthony F Herzig (1), Emmanuelle Génin (1,3)

(1) Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, FR, (2) Inserm, Brest, FR, (3) CHRU Brest, F 29200 Brest, FR

Population structure is a major confounding factor in genome-wide association analyses (GWAS). A common best-practice is to check for ancestry homogeneity through principal component analyses (PCA) and include the first few principal components (PCs) as covariates to control for structure and avoid spurious association signals. An alternative is linear mixed modelling, where a genetic relationship matrix (GRM) is used to describe the covariance between individuals. This essentially allows all principal components to enter the model with random

effects and can therefore control for both the population structure and cryptic relatedness.   Here, we used genotyping data from 9,598 individuals of the French POPGEN project. By design, this sample represents the French general population at beginning of 20th century; with a precise geographical anchorage for each individual. We investigated the impact of different adjustment methods to control for fine-scale population structure on GWAS when the phenotype of interest is confounded by geography. To do so, we here modelled latitude and longitude as response variables; these two co-ordinates being of course geographically rather than genetically determined; but for which heritability estimates are known to be significantly different from zero in modern populations.   As expected, we obtained large genomic inflation factors in GWAS of latitude and longitude without adjusting for population structure, with more than 2,000 and 300 genome-wide significant loci all along the genome for latitude and longitude respectively. Our results show that the sample on which the PCA has been computed is crucial to capture the fine-scale diversity present among the individuals contributing to the GWAS and to accurately adjust the analysis. First, we show that PCs computed on a world-wide sample are much less efficient in correcting the inflation than PCs computed on the association analysis sample. Adjusting on the first 4 PCs, more than 400 and 90 genome-wide significant loci remained for latitude and longitude respectively when correcting with world-wide PCs, while no genome-wide significant signal remained when correcting with the study sample PCs. Second, we show that working with haplotype data rather than independent SNP data further improves the inflation correction. Third, we show that using a linear mixed model seems to appropriately correct for the global inflation, but signals of genome-wide significant association remained; notably for genomic regions known to be under recent selective pressure in European populations.   In conclusion, our results highlight the importance of capturing the genetic diversity within the sample of analysis to correctly adjust for fine-scale population structure. Our results also warn on using a linear mixed modelling in complex-trait genetics in human populations as this does not account for certain patterns of correlation between phenotypes and geography.

Poster 14

Correcting genomic inflation using Genomic Control and LD-Score regression leads to loss of power in GWAS meta-analysis

Archit Singh (1,2,3), Lorraine Southam (1), Konstantinos Hatzikotoulas (1), Ken Suzuki (4,5,6), Henry J Taylor (7,8,9), Xianyong Yin (10,11), Ravi Mandla (12,13), Alicia Huerta-Chagoya (12), Nigel W Rayner (1), T2D-GGI, Andrew P. Morris (1,4), Eleftheria Zeggini (1,14)*, Ozvan Bocher (1)*

(1) Institute of Translational Genomics, Helmholtz Zentrum München- German Research Center for Environmental Health, Neuherberg, DE, (2) Munich School for Data Science (MUDS), Helmholtz Zentrum München- German Research Center for Environmental Health, Neuherberg, DE, (3) Doctoral program of Experimental Medicine and Health Sciences, TUM School of Medicine and Health, Technical University of Munich, Munich, DE, (4) Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, University of Manchester, Manchester, UK, (5) Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, University of Tokyo, Tokyo, JP, (6) Department of Statistical Genetics,

*Osaka University Graduate School of Medicine, Suita, JP, (7) Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA, (8) British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, (9) Heart and Lung Research Institute, University of Cambridge, Cambridge, UK, (10) Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, CI, (11) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA, (12) Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA, (13) Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, (14) TUM School of Medicine and Health, Technical University of Munich and Klinikum Rechts der Isar, Munich, DE, * These authors contributed equally to the study*

Genome-wide association study (GWAS) meta-analyses are widely used to describe genetic variants associated with complex traits. However, these studies can be prone to systematic bias arising for example from population stratification. To reduce the resulting genomic inflation and false positives, genomic control (GC) correction computes the λ-value by comparing the observed χ2-distribution of tested variants to its theoretical counterpart. This method might not effectively correct inflation in large, well-powered genetic studies of traits with significant polygenic background, given assumptions about the genetic structure of complex traits, where most of the genome is not expected to show association. LD-score regression (LDSR) improves upon GC by utilizing LD information to discern trait polygenicity and confounding, contingent on the reference panel's population representation. Here, we compare these two correction methods and assess their impact on genetic discovery in GWAS meta-analysis. Utilizing type 2 diabetes as an exemplar polygenic disease, we conduct a comparative analysis across three consecutive, overlapping GWAS meta-analyses, two published by the DIAMANTE consortium in 2018 and 2022 ("DIAMANTE-18" and "DIAMANTE-22"), and the recent largest meta-analysis published by the T2DGGI consortium, representing effective sample sizes of 231K, 492K and ~1.5M, respectively. We ascertain the proportion of true associations that are consistently replicated across these studies and assess whether genomic inflation correction leads to loss of true positives. We calculate the false positive association proportion and the true positive association loss before and after applying correction methods by measuring false positive rate (FPR) and true positive rate (TPR). We find that GC-correction in the DIAMANTE-22 study leads in a loss of over ~3000 associated variants. These variants map to 40 independent loci that are not tagged by other significantly associated variants and are therefore lost after correction. We observe that association of over 90% of these variants is recovered in the T2DGGI study, highlighting that they are true signals recovered by a better powered GWAS meta-analysis. Correction of summary statistics using the LDSR-intercept leads to similar loss of variants and independent loci. When applying GC-correction, we find limited improvement to FPR and a decline in TPR. For example, in the DIAMANTE-22 study, the FPR improves slightly from 4×10-5 to 2×10-5 after GC-correction, while the TPR declines from 0.31 to 0.26. We observe similar changes when correcting using the LDSR-intercept. Although the intention behind employing these correction methods is to minimize the false positive rate (FPR), we observe minimal evidence of FPR reduction and an important shrinkage of independent loci, suggesting an overall loss of association signals when employing

these methods. Our findings underscore drawbacks of using these correction methods in large meta-analyses, namely the inappropriate GC-correction on traits with high polygenicity, which is more comprehensively captured as sample size and, hence, power grows; and the limitation of the LDSR-correction when incompletely representative LD reference panels are used. We advise caution in applying these methods and emphasize the necessity for further research to effectively address inflation in GWAS meta-analyses.

Poster 15

Overcoming Ultra Low Sequencing Depth Challenges: Classification Based on Motif Copy Number Variations. Determining Y-Chromosome Haplogroups Using Low Quality Short Read WGS Data.

Tarmo Puurand, Märt Möls, Toomas Kivisild, Maido Remm
University of Tartu, Institute of Genomics, Estonian Genome Centre, EE

Accurate prediction and classification of individuals from short read whole genome sequencing (WGS) data with low coverage, such as 0.01x coverage, pose significant challenges due to the scarcity of reads for SNP variant calling. However, leveraging information on motif copy number variations of tandem repeats, such as centromere lengths, can provide a solution even in the face of low sequencing depth. We demonstrate the efficacy of utilizing motif copy number of Y-chromosome specific tandem repeats (e.g., VNTR, CEN, HET) for precise prediction of Y-chromosome haplogroups, even with degraded low-coverage DNA samples. Our proposed fast k-mer based method do not require mapping of the reads and is robust against environmental contamination, offering a practical solution for accurate classification. We will detail the modelling process, including pre-selection of k-mers, estimation of the classification model, and calculation of classification accuracy, providing insights into the practical implementation of this approach.

Poster 16

Pathway Analysis in Cognitive Decline using Longitudinal Data with Respect to Common Variants

Elaheh Vojgani (1,4), Kumar Parijat Tripathi (2), Luca Kleineidam (3), Alfredo Ramirez Zuniga (2), Michael Nothnagel (1,4,5)
(1) Statistical Genetics and Bioinformatics Group, Cologne Center for Genomics (CCG), University of Cologne, Cologne, DE, (2) Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry, Medical Faculty, University of Cologne, Cologne, DE, (3) Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, DE, (4) University Hospital Cologne, Cologne, DE, (5) West German Genome Center, Cologne site, University of Cologne, Cologne, DE

Background/Objectives: Biological pathways intricately govern cellular processes, and understanding their role in neurological disorders' cognitive decline is paramount. Alzheimer's disease (AD) is a multifaceted neurological disorder characterized by cognitive decline, often manifesting early in life. Investigating the underlying biological mechanisms triggering cognitive impairment before dementia onset is crucial.   Methods: In this study, we employ a pathway-driven statistical approach, considering common genetic variants across pathways, to generate pathway scores. These scores aim to assess their interaction with time in driving cognitive decline, utilizing longitudinal data with temporal cognitive scores. Our approach involves two key components: first, developing a pipeline for pathway score generation with simulation studies involving various degrees of variant effect sizes within a pathway as well as different proportions of protective variants within a pathway to evaluate its efficacy, and second, applying the scores in a linear mixed model to examine their association with cognitive decline using real data from available cohorts.   Results: Preliminary results indicate that with large proportions of deleterious variants in a pathway and a strong correlation between genetic effect size and CADD score of a variant, the linear model is very well able to predict the phenotype, with coefficients of determination consistently exceeding 0.5. Conclusion: The proposed model captures phenotypic variation in simulated data effectively through pathway scores. This approach promises to enhance the understanding of biological processes driving the neurological disorders. In future, we will implement an R-package pipeline to enable users to obtain CADD-based pathway scores.

Poster 17

Time-to-Event Data Analysis in Genome-Wide Association Studies: A Simulation Study

Anastassia Kolde, Merli Mändul, Krista Fischer

University of Tartu, Institute of Mathematics and Statistics, University of Tartu, Institute of Genomics, EE

The recent surge in sample size in Genome-Wide Association Studies (GWAS) has been matched by the valuable expansion of data through extended follow-up time and linkage of omics databases with electronic health records (EHR) in large-scale population-based biobanks. A large proportion of GWASs is mainly focused on discovery of genetic variants associated with the risk of incident diseases. For that purpose, one needs to apply a regression modeling methodology that is designated for censored time to event data. Here, the Cox Proportional Hazards (CPH) has become a standard approach, as an alternative to simpler methods like linear or logistic regression models. A widely known fundamental assumption of the CPH model is that the effect of each covariate in the model needs to be proportional with respect to time and with respect to the other covariates in the model, but there are other issues to consider that are specific to biobanks and studies of observational nature. Firstly, since biobanks are mostly volunteer-based then there is no natural origin time for the start of the study, hence one needs to carefully consider the choice of the time-scale for the analysis and consider the adjustments needed for accounting for left truncation. Secondly, assumption of independence of subjects are scrutinized in biobank settings and imposing the need to account for possible genetic relatedness. Thirdly, in the situation of the abundance of covariates yielding from

EHRs and omics databases one needs to be mindful about choice of the covariates. We conducted a simulation study, mirroring the Estonian Biobank cohort, that investigates the performance of different models under varying time-scale choices, the implications of familial relationships, and the impact of omitted covariates. Our principal findings underscore that while adjustments for left-truncation are needed for unbiased results, they do not hold paramount importance in discovery studies, nor does accounting for relatedness. However, it is imperative to exercise caution with omitted covariates, as their absence can lead to significant bias.

Poster 18

Genome-wide landscape of blood traits pleiotropy

Arianna Landini (1), Sodbo Z. Sharapov (1), Nicola Pirastu (1), Nicole Soranzo (1,2,3)

(1) Human Technopole, IT, (2) British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK, (3) Human Genetics Department, Wellcome Sanger Institute (WT), Hinxton, UK

The regulation of blood cell counts results from the complex interplay between genetics and physiological status. Despite previous studies uncovered numerous genetic associations, distinguishing between cell type-specific and shared signal remains unclear. We address this gap by exploring the pleiotropic landscape of 29 blood cell count traits. We performed GWAS of TOPMed-imputed genotypes from ~350k UK Biobank (UKBB) individuals of European ancestry. We explored the landscape of pleiotropy by using leave-one-out conditional approach, obtaining signals with a single underlying causal SNP, followed by Bayesian colocalisation analysis. We identified 13,027 significant trait-loci pairs (p-value < 5e-8, MAF > 1e-4), corresponding to 1328 overlapping genomic regions. Among these, 247 were trait-specific, while the remaining were associated with multiple traits, up to 13 linked to all 29 traits. Conditional analysis revealed 23,904 independent association signals, which were grouped in 4435 colocalisation groups containing few signals specific to a single trait (3283). Despite the high level of pleiotropy, about half (42.5%) of these groups involved only two traits, while few (1.3%) encompassed over half of the 29 blood traits. Individual signals near ABO and SLC7A5 influenced 25 traits with different patterns. While the firsts showed concordant effects across most traits, the seconds showed divergent effects (increasing red blood cells and platelets while decreasing white blood cells), suggesting different roles in blood cells regulation. These results deepen our understanding of blood cell genetic regulation, highlighting gene and locus pleiotropy and intricate local regulation, likely reflecting differences in variant activity across cell types and along the differentiation process.

Poster 19

Interpretable genomic predictions via effect propagation in gene regulatory network.

Natalia Ruzickova, Michal Hledik, Gasper Tkacik

Institute of Science and Technology Austria, AT

As their statistical power grows, genome-wide association studies (GWAS) have identified an increasing number of loci underlying quantitative traits of interest. These loci are scattered throughout the genome and are individually responsible only for small fractions of the total heritable trait variance. The recently proposed omnigenic model provides a conceptual framework to explain these observations by postulating that numerous distant loci contribute to each complex trait via effect propagation through intracellular regulatory networks. We formalize this conceptual framework by proposing the “quantitative omnigenic model" (QOM), a statistical model that combines prior knowledge of the regulatory network topology with genomic data. By applying our model to gene expression traits in yeast, we demonstrate that QOM achieves similar gene expression prediction performance to traditional GWAS with hundreds of times less parameters, while simultaneously extracting candidate causal and quantitative chains of effect propagation through the regulatory network for every individual gene. We estimate the fraction of heritable trait variance in cis- and in trans-, break the latter down by effect propagation order, assess the trans-variance not attributable to transcriptional regulation, and show that QOM correctly accounts for the low-dimensional structure of gene expression covariance. We furthermore demonstrate the relevance of QOM for systems biology, by employing it as a statistical test for the quality of regulatory network reconstruction and linking it to the propagation of non-genetic, environmental effects.

Poster 20

A Bayesian functional principal component analysis model for longitudinal gene expression data

Selima Jaoua (1,2), Daniel Temko (1), Hélène Ruffieux (1)
(1) MRC biostatistics Unit, University of Cambridge, Cambridge, UK, (2) EPFL Lausanne, Lausanne, CH

A Bayesian functional principal component analysis model for longitudinal gene expression data Selima Jaoua, Daniel Temko, Hélène Ruffieux Introduction: Stereotyped synchronised patterns of gene expression are a feature of human tissue differentiation and of human immune responses in the context of disease. Existing approaches to modelling this data rely heavily on statistical dimensionality reduction techniques designed for studying either single time-point data or univariate functional data, such as principal component analysis and functional principal component analysis (FPCA), respectively. However, appropriate bespoke methods for modelling coordinated patterns of temporal variation are currently lacking. Results We present a novel Bayesian model, factor FPCA, which is designed to capture correlations across genes and time in these settings. Our model posits that samples vary in their levels of activity of a small number of functional principal components. As in standard FPCA, each of these components is associated with a smoothly varying temporal loading function. However, in contrast to standard FPCA, in the factor FPCA model each component is also associated with a vector of gene-specific loadings, encoding the gene-wise weighted contribution of the component. We couple our model with an efficient mean-field variational inference scheme. We show through extensive simulations that our combined modelling and inference framework is (i) capable of accurately recovering true underlying component loadings and sample-specific component scores, and (ii) is scalable to realistic genomic dataset sizes.

Implications Overall, our approach encodes useful priors in a novel Bayesian model, which is coupled with an efficient inference scheme. Through applications to developmental and disease immune-response data, we expect that this framework will shed new light on processes relevant to human health, and that these insights could provide clinically relevant leads to inform future drug development.

Poster 21

Transcriptomic risk scores for attention deficit/hyperactivity disorder

J. Cabana-Domínguez (1,2,3,4,9), N. Llonga (1,2,9), L. Arribas (1,2), S. Alemany (1,2), L. Vilar-Ribó (1,2,3), P. Carabí (1,2), D. Demontis (6,7,8), C. Fadeuilhe (1,2,3,5), M. Corrales (1,2,3,5), V. Richarte (1,2,3,5), A. Børglum (6,7,8), J.A. Ramos-Quiroga (1,2,3,5), M. Soler Artigas (1,2,3,4,10), M. Ribasés (1,2,3,4,10)
(1) Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, ES, (2) Department of Mental Health, Hospital Universitari Vall d'Hebron, Barcelona, ES, (3) Biomedical Network Research Centre on Mental Health (CIBERSAM), Madrid, ES, (4) Department of Genetics, Microbiology, and Statistics, Faculty of Biology, Universitat de Barcelona, Barcelona, ES, (5) Department of Psychiatry and Forensic Medicine, Universitat Autònoma de Barcelona, Barcelona, ES, (6) Department of Biomedicine/Human Genetics, Aarhus University, Aarhus, DE, (7) The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, DE, (8) Center for Genomics and Personalized Medicine, Aarhus, DE, (9) These authors contributed equally, (10) These authors jointly supervised this work

Background: Attention deficit hyperactivity disorder (ADHD) is a highly heritable neurodevelopmental disorder. The proportion of phenotypic variance explained by all measured genetic variants is 14% and the polygenic risk score (PRS) for the disorder explains around 5.5%. The aim of the present study is to estimate transcriptomic risk scores (TRSs) using transcriptome-wide association study (TWAS) results from the latest ADHD genome-wide association study meta-analysis (GWAS-MA) and transcriptomic profiles from peripheral blood mononuclear cells (PBMCs) of ADHD patients and controls, and assess whether combining polygenic risk scores (PRS) and TRS improves ADHD prediction over PRS alone. Methods: TWAS was performed using the S-PrediXcan method, integrating data from the latest ADHD GWAS-MA (38,691 individuals with ADHD and 186,843 controls) and transcriptomic imputation models from the joint-tissue imputation (JTI) approach using GTEx v8 whole blood and 13 brain tissues. Colocalization analyses were performed for the 14 studied tissues using CAUSALdb and fastENLOC software. TRSs were constructed based on gene expression data from PBMCs in an in-house sample of 222 medication-naïve adult ADHD cases and 269 controls as the sum of the expression of each gene weighted by its signed z-score value in the TWAS. Different TRSs were estimated by selecting genes according to several P-value thresholds from the TWAS on ADHD. PRS were constructed based on the GWAS-MA on ADHD in the same clinical sample with PRScs and PLINK1.09. Logistic regression models were used to test the association between TRS or PRS and ADHD, including age, sex, GWAS batch and first 10 principal components as covariates. Finally, likelihood ratio test was used to compare the goodness of fit of the model with the PRS and covariates with the model that also included the TRS.

Results: In the multiple-tissue TWAS, 56 genes were differentially expressed between ADHD and controls in at least one of the 14 tissues studied. We found significant association between ADHD and TRSs derived from gene expression profiles in 11 out of 13 brain areas. The TRS remained significantly associated with ADHD in four brain tissues after multiple comparison correction (amygdala, caudate, cortex and frontal cortex). Additionally, the TRSs showed increased proportion of variance explained as more strict association P-values were used to select genes from TWAS. When restricting the analyses to the subset of colocalized genes in the four brain tissues after multiple comparison correction, the association with ADHD was stronger and the proportion of variance explained increased in amygdala, caudate and frontal cortex. In addition, TRSs and PRS were not correlated, and the combination of TRSs and PRS significantly improved the proportion of variance explained beyond the PRS-only model. Conclusions: We found association between ADHD and TRS in PBMCs constructed using TWAS results from multiple brain areas, showing that individuals with ADHD carry a higher burden of TRSs than controls. All models combining TRS and PRS improved the fit of the model over PRS alone, pointing to the complementary predictive potential of transcriptomic profiles.

Poster 22

A novel life-time risk model to estimate the age-dependent penetrance of rare and common variants in complex diseases and application to Alzheimer disease

Catherine Schramm (1), Emmanuelle Génin (2), Olivier Quenez (3), ADES consortium, Gaël Nicolas (3), Camille Charbonnier (4)

(1) Univ Rouen Normandie, Inserm U1245, Normandie Univ, Rouen, FR, (2) Inserm UMR1078, Brest, FR, (3) Univ Rouen Normandie, Inserm U1245, Normandie Univ, CHU Rouen, Department of genetics, Rouen, FR, (4) Univ Rouen Normandie, Inserm U1245, Normandie Univ, CHU Rouen, Department of biostatistics, Rouen, FR

Recent association studies on exome sequencing datasets unveiled rare genetic risk factors with stronger levels of association with complex diseases than most of their common counterparts. Although each of them are rare or ultra-rare, together they affect a non-negligeable part of the population. It is thus essential to accurately estimate their absolute penetrance, stratified for major common risk factors and/or genetic risk score (GRS) strata, before putative use in a clinical setting. This task represents a methodological challenge since the rare variants in question are too rare to be studied through traditional prospective cohorts and collect pedigree data for each category of variant seems unrealistic. To solve this problem, we developed a model combining information from prospective cohorts and case/control sequencing data into a Kaplan-meier-like estimator to assess age-dependent penetrance of different rare variant categories. For each age interval (a) and each common genetic risk strata (C), the probability of disease (D) for rare variant carrier (R) is computed using the Bayes formula as $P(D|a,C) \times P(R|a,C,D)/P(R|a,C)$ where $P(D|a,C)$ is derived from prospective cohort and $P(R|a,C,D)$ is computed on our case-control data. In practice, several parametrizations of this model were designed and evaluated. To test for epistasy between common and rare variants, we proposed to compare estimations with and without interaction using bootstrap. In a simulation study, we generated piecewise constant variant effects and showed that our method is able to correctly estimate their associated penetrance curves when P(R|a,C) is computed in controls only. After validation through extensive simulation studies, we applied our model to Alzheimer disease, integrating the approximately 25,000 samples of the ADES-ADSP case-control sequencing dataset with large published prospective data. We derived Alzheimer disease penetrance curves for rare but recurrent TREM2-p.R47H, TREM2-p.R62H, ATP8B4-p.G395S and ABCA7 truncating variants, stratified for the common APOEe4 allele and GRS tertiles. Cumulated effects of TREM2-p.R47H and APOEe4/e4 led to almost complete penetrance by age 85, although TREM2-p.R47H alone and APOEe4/e4 alone conferred less than 30% and 50% risk by age 85, respectively. Our model offers a perspective to estimate the penetrance of combinations of risk factors integrating prospective and case/control data. Results on Alzheimer disease data show the importance of accounting for digenic dimensions, at least, and that the background effect of GRS appears negligeable, in comparison.

Poster 23

Unravelling the interplay between type 2 diabetes, genetics and metabolite levels

Ozvan Bocher (1), Archit Singh (1,2,3), Ana Luiza Arruda (1,2,3), Ene Reimann (4), Urmo Võsa (4), Andrei Barysenka (1,5), William Rayner (1,5), Reedik Mägi (4), Eleftheria Zeggini (1,5)

(1) Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, DE, (2) Technical University of Munich (TUM), TUM School of Medicine, Munich, DE, (3) Helmholtz Association - Munich School for Data Science (MUDS), (4) Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, EE, (5) Technical University of Munich (TUM) and Klinikum Rechts der Isar, TUM School of Medicine, Munich, DE

Numerous metabolite levels have been associated with the occurrence of type 2 diabetes (T2D), but their causal role in T2D development and the involvment of genetics in mediating those relationships remain to be elucidated. We sought to investigate the interplay between genetics, metabolomics and T2D risk in the UK Biobank cohort by using bidirectional two-sample Mendelian Randomization (MR) and interaction QTL analyses. In the forward MR, we describe 63 metabolites with a causal effect on T2D, including glucose, apolipoprotein B and various lipid classes. In the reverse direction, we report a causal effect of T2D liability on 178 metabolite levels (with p-values down to 10-175), including an increase in alanine, valine and glucose levels, and a decrease in levels from cholesterol classes. Secondly, we describe 14 metabolites which exhibit a different genetic regulation between T2D cases and controls. Four of these metabolites, L_VLDL_FC_pct, L_VLDL_TG_pct, L_LDL_PL_pct and S_HDL_FC_pct, were replicated in the Estonian Biobank with p-values down to 9.88x10-13 and 8.1x10 4 in the discovery and replication cohort respectively. These variants reside in two different genomic regions and are

significant QTLs for the corresponding metabolites in healthy individuals but not in individuals with T2D. Additionally, these variants are not associated with T2D, suggesting that the different genetic regulation at these loci is a consequence rather than a cause of T2D development. This work provides a better understanding of the metabolic changes induced by the occurrence of T2D and provide potential directions to investigate T2D consequences and subsequent complications.

Poster 24

Nested UK biobank Exome-wide association study of Recurrent Pregnancy Loss – exploiting the multi-modal data in large population biobanks to define complex health outcome.

Chia-Yi Chu (1), Yevheniya Sharhorodska (2,3), Inga Prokopenko (3)
(1) People-Centred Artificial Intelligence Institute, University of Surrey, Guildford, UK, (2) Department of Clinical and Experimental Medicine, School of Biosciences and Medicine, University of Surrey, Guildford, UK, (3) Department of clinical genetics, Institute of Hereditary Pathology, National Academy of Medical Sciences, Lviv, UA

Introduction: Recurrent pregnancy loss (RPL) affects around 2.6% of women, with nearly half of cases being idiopathic, i.e., lacking known causes. Though previous large-scale genome-wide association studies (GWAS) attempted to identify genetic variants associated with generic miscarriage, only three associations with low frequency SNPs near TLE1, NAV2 and SIK1 genes are reported. Objective: We explored the RPL susceptibility using UK biobank (UKBB) multi-modal data by establishing stringent case and control definitions. Methods: In UKBB, the RPL cases were defined either by diagnosis as habitual aborter or by two or more diagnoses of missed/spontaneous abortion in primarily hospital admission records, supplemented by self-reported data. The controls were women with a healthy live birth and without a medical record or self-reported pregnancy loss. We defined 409 RPL cases and 1,639 controls among European ancestry women. We used the whole-exome-sequencing data and performed association analysis in PLINK v2.00 for 26,388,328 SNVs, assuming a log-additive model of inheritance. Results: We identified a suggestive (P-value<10-5) association with RPL risk at KIR2DS1 gene (rs377462001, OR[95%CI]=4.64[2.46-8.78], P-value=2.28´10-6),which involves in immune response and aligns with previous reports on RPL susceptibility. We also observed nominal effects at NBPF3 (rs145079058, OR[95%CI]=1.62[1.29-2.04], P-value=3.05´10-5), previously associated with vitamin B6 level in GWAS, as well as at FCRL5 (rs6679793, OR[95%CI]=0.63[0.50-0.80], P-value=9.80´10-5) and LGALS1 (rs77181504, OR[95%CI]=1.63[1.29-2.05], P-value=3.62´10-5) genes, both implicated in immune response. Conclusion: Implementing stringent case and control definitions in RPL studies demonstrate potential to identify more variants associated with RPL and advance our understanding of its aetiology. Funding: RaR\100084.

Poster 25

Efficient completion of genotype correlation matrix combining imputed and

sequenced data

Svishcheva G.R. (1,2), Kirichenko A.B. (1), Belonogova N.M. (1), Tsepilov Y.A. (1), Zorkoltseva I.V. (1), Axenovich T.I. (1)
(1) Federal Research Centre, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, RU; (2) Vavilov Institute of General Genetics, Russian Academy of Sciences, RU

The problem of reconstructing genotype correlation matrices arises when imputed and sequenced data are combined for joint gene-based association analysis. Usually, gene-based association analysis is performed using GWAS summary statistics and correlations between genotypes of variants within a gene. These data are freely available for each of the imputed and sequenced genotypes.   We propose an efficient analytical method of reconstructing correlations between the genotypes of two variants, one of which is imputed and the other is sequenced. The method is based on maximizing the matrix determinant. It has a number of useful properties and has an analytical solution for our task. To test the proposed method, we compared the structure of original and reconstructed matrices of correlations and the results of gene-based association analyses (performed using the Burden, PCA and SKAT tests) of body mass index from the UK biobank using these matrices. The original matrix contained pairwise Pearson's correlations calculated using real genotypes from the UK biobank, and the reconstructed matrix was derived via correlations between imputed genotypes and correlations between sequenced genotypes. The results showed a high quality of matrix completion: the difference between the correlation coefficients and the difference between the results of the gene-based association analysis were very low. Acknowledgements This research has been conducted using the UK Biobank resources under application number 59345. The work of Svishcheva, Belonogova and Zorkoltseva was supported by Russian Science Foundation (RSF) grant No. 23-25-00209. The work of Kirichenko, Tsepilov and Axenovich was supported by the budgetary project of IC&G SB RAS No. FWNR-2022-0020.   Conflict of interest The authors declare that they have no conflict of interest.

The poster can only be accessed here.

Poster 26

Statistical integration of multi-omics and drug screening data from cell lines

Jeanine Houwing-Duistermaat (1,2), Said el Bouhaddani (3)
(1) Department of Mathematics, Radboud University Nijmegen, NL, (2) Department of Statistics, University of Leeds, UK, (3) Department of Data science & Biostatistics, UMC Utrecht, NL

A novel computational workflow to summarise multi-modal datasets in a unified way will be presented. This work was motivated by a study with transcriptomics, proteomics, and drug screening datasets that were measured in LUHMES cell lines and controls. LUHMES cell lines show α-synuclein aggregation and are used to study biological mechanisms underlying neurodegenerative diseases such as multiple system atrophy (MSA) and Parkinson’s disease. Our aim is to identify potentially druggable pathways and genes involved in MSA.   The workflow comprises a novel probabilistic data integration method, named POPLS-DA, and a data linking part using functional databases to integrate the drug screen data. POPLS-DA is a latent variable model which addresses heterogeneity by including data-specific components. Since the different datasets are not measured on the same cells, the unit of the model is a gene, i.e. datasets are linked via genes. For prioritizing genes using these datasets, the performance of POPLS-DA is compared to other single- and multi-omics approaches. We applied the workflow to the multi-modal data. POPLS-DA appeared to perform better than other integration approaches. The output of POPLS-DA was used to construct an integrated interaction network where the drug screen data was incorporated to highlight druggable genes and pathways in the network. Finally, a functional enrichment analyses are performed to identify clusters of synaptic and lysosome-related genes and proteins targeted by the protective drugs. We found that HSPA5, a member of the heat shock protein 70 family, was one of the most targeted genes by the validated drugs, in particular by AT1-blockers. HSPA5 and AT1-blockers have been previously linked to α-synuclein pathology and Parkinson’s disease, showing the relevance of our findings. Our computational workflow identified new directions for therapeutic targets for MSA. POPLS-DA provided a larger interpretable gene set than other single- and multi-omic approaches. An implementation based on R and markdown is freely available online.

Poster 27

Comparing the performance of clustering methods to highlight fine-scale genetic structure using POPGEN data

Mael Guivarch (1) *POPGEN Study Group (1), Emmanuelle Génin (1,2), Anthony F. Herzig (1),*, Aude Saint Pierre (1),(* 1). Univ. Brest, Inserm, EFS, UMR 1078, GGB, F-29200, Brest, France 2. CHU Brest, F-29200 Brest, France * Equal Contribution*

Introduction: In order to highlight genetic risk factors associated with complex diseases, it is crucial to analyse patients’ genomes alongside those of genetically similar individuals from the general population; underscoring the importance of understanding the fine-scale genetic structure of the population. Employing clustering techniques is a commune approach for describing such structure, with the aim of categorizing individuals into groups based on their genetic profiles. These methods are furthermore essential for various applications in population genetics, including accurately calculating allele frequencies. Nevertheless, efficiently identifying groups and selecting optimal clustering approaches from a wide variety of algorithms can be particularly challenging. Methods: In this context, we offer a comparative examination of diverse clustering methodologies, emphasizing hierarchical methods like fineSTRUCTURE, model-based clustering approaches such as Mclust, network-based approaches such as the Leiden algorithm, and aggregation-based clustering techniques. Additionally, we explore the effects of various similarity metrics derived from haplotype-sharing methods on the outcomes of clustering. Results: We enhance previous comparative studies by harnessing the capabilities of tree sequence toolkits to simulate data according to controlled demographic scenarios. The simulated datasets are created to best fit the demographics scenarios of real data sourced from POPGEN, a project encompassing the entire metropolitan region of France with genotyping data for 9598 individuals. Our simulations enable us to assess the reliability and accuracy of different clustering methods in controlled environments. Conclusion: Our study highlights the performance of various clustering approaches on datasets simulated from controlled demographic scenarios, offering insights to interpret fine-scale population structure and to select appropriate methods. We explore techniques for simulating data with fine-scale structure as well as giving examples of clustering results based on the POPGEN project. Funding: This work is funded by the French Ministry of Research for the POPGEN project in the framework of the French initiative for genomic medicine (Plan France Médecine Génomique 2025; PFMG 2025; https://pfmg2025.aviesan.fr). The CONSTANCES cohort benefits from grant ANR-11INBS-0002 from the French National Research Agency.