A benchmarking of univariate pleiotropy detection methods, with an application to epilepsy phenotypes.

Adesoji Oluyomi (1), Nothnagel Michael (1), Lerche Holger (2), May Patrick (3), Krause Roland (3)

1 - Cologne Center for Genomics [Cologne] (University of Cologne, Cologne Germany), 2 - Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research (University of Tübingen, Tübingen, Germany), 3 - Luxembourg Centre for Systems Biomedicine (University of Luxembourg, Esch-sur-Alzette, Luxembourg)

Correspondance to: oadesoji@uni-koeln.de

Pleiotropy denotes the association of a hereditary unit (SNP, locus, or gene) to multiple features or phenotypes. Detection of true biological pleiotropy, opposed to mediated or spurious locus one, may inform on the aetiology of complex diseases as well as on their categorization. Applicability of the proposed many methods for detection of pleiotropic loci, is to a large extent determined by the type of data available and the methods’ comparative performance is yet unclear. Here, we benchmarked different methods based on univariate data in non-overlapping sample sets, namely classical meta-analysis (MA), conditional false discovery rate (cFDR), a subset-based meta-analysis (ASSET), and Cross Phenotype Bayes (CPBayes). To this end, we repeatedly simulated, from a resampled population of 50,000 individuals of European ancestry based on the 1000 Genomes data and for two pleiotropic phenotype datasets of varying sizes and with different underlying aetiologies that combine various values for variant effect sizes and allele frequencies, numbers of causal SNPs, sample sizes, and the degree of genetic overlap between those phenotypes. We assessed their impact on power and false positive rate (FDR) of the methods. We discovered that sample size and effect sizes and number of simulated disease SNPs have great effects on the ability of these methods to detect pleiotropy. We also observed considerable differences in the trade-off between power and FDR between these methods. Applying the most promising methods to epilepsy sub-phenotype datasets from The International League Against Epilepsy (ILAE) consortium, we identified already known and new putative epilepsy genes.
**Pigmentation prediction in admixed Latin Americans**

**Adhikari Kaustubh (1) (2), Palmal Sagnik (3), Faux Pierre (3), Mendoza-Revilla Javier (4), Balding David (5), Ruiz-Linares Andres (3)**

1 - Department of Genetics, Evolution and Environment, University College London (University College London, London, WC1E 6BT, UK United Kingdom), 2 - School of Mathematics and Statistics, The Open University (The Open University, Milton Keynes, MK7 6AA, UK United Kingdom), 3 - Aix-Marseille Université, EFS, Faculté de Médecine Timone (Marseille, 13005, France France), 4 - Unit of Human Evolutionary Genetics, Institut Pasteur (Paris, 75015, France France), 5 - Department of Mathematics and Statistics [Melbourne] (The University of Melbourne Parkville, VIC, 3010 Australia)

Correspondance to: k.adhikari@ucl.ac.uk

There is increasing interest in the use of genetic information for predicting physical appearance traits, particularly in forensics and palaeoanthropology. We had previously conducted a GWAS of pigmentation traits for over 6,500 admixed Latin Americans in the CANDELA cohort. Here we carry forward the findings to the construction of prediction models for eye, hair and skin pigmentation based on genomic and phenotypic data.

Pigmentation is different from other complex traits such as height in the sense that a handful of loci explain a much larger proportion of phenotype variation, and the effect of external factors is lower. This has facilitated the construction of well-established pigmentation prediction tools such as HIrisPlex-S. However, such tools are primarily constructed based on studies of Eurocentric populations. Interestingly, eye and hair colour do show the largest variability in Europe, while skin colour has larger variability outside Europe.

We utilize the admixed nature of the Latin American cohort to assess the accuracy of pigmentation prediction across various ethnicities, exploring various statistical methodologies. We examined the impact on prediction accuracy of three main factors: (i) The methods of prediction, including classical statistical methods and machine learning approaches, (ii) The choice of predictors: non-genetic predictors, continental genetic ancestry, and pigmentation SNPs, in the prediction models, and (iii) Comparison of two sets of pigmentation SNPs: the commonly-used HIrisPlex-S set and novel SNP sets that we have defined based on association results in the CANDELA samples.

We find that Random Forest or regression are globally the best performing methods. Although continental genetic ancestry by itself has some relevance for the prediction of pigmentation in Latin Americans, the inclusion of pigmentation SNPs increases prediction accuracy considerably, particularly for skin colour.

For hair and eye colour, HIrisPlex-S has a similar performance to the CANDELA-specific sets, as expected from the continental nature of those traits. But we show that constructing bespoke prediction models for skin colour among participants of Native American ancestry improves prediction substantially, consistent with our findings of GWAS loci that are unique to non-European populations. Our findings reiterate the need for diversity in genetics research.
Genetic researchers become increasingly interested in modelling multivariate genetic trait architectures not only in twins but in population-based cohorts. In analogy to multivariate twin modelling strategies (Martin and Eaves, 1977; Neale and Cardon, 2013), genetic factor structures, as captured by genetic-relationship-matrices (GRM) derived from genome-wide genotyping information, can be identified using structural equation modelling techniques (SEM). This ‘GRM-SEM’ approach has been originally introduced by St Pourcain et al. (2018) and extends the methodology of estimating bivariate genetic covariance with genomewide genotyping information, as introduced by Yang et al. (2010); Lee et al. (2012) using a genomic restricted maximum likelihood (GREML) approach. Since then, multiple research groups worked on the development of GRM-SEM.

Here, we present the grmsem CRAN R package, an extension of the original gsem package St Pourcain et al. (2018) that provides applied researchers, teachers, and statisticians an easy-to-use open-source software to identify genetic and residual covariance structures in samples of unrelated individuals with genome-wide information. The grmsem package contains multiple pre-specified model structures and offers a suite of user-friendly posthoc analyses, including estimations of factorial co-heritability (the proportion of total genetic variance explained by a specific genetic factor) and bivariate heritability (the contribution of genetic factors to the observed phenotypic covariance between two measures).

During the presentation, we will, first, show the most important features of the grmsem package along with an example. Second, we will report on the model fit of multiple saturated models, including the Cholesky model and the Direct Symmetric model (Verhulst et al., 2019), and show that, contrary to twin analyses, in GRM-SEM both models provide equivalent results for heritable traits. Third, we will highlight the importance of accurately modelling residual covariance for identifying genetic structure. Fourth, we discuss differences and similarities with recently published alternative approaches estimating SEM from genomic data. We conclude the presentation by highlighting future plans for the grmsem package.
Genetic validation of Heart Failure drug targets using Mendelian randomisation

Baudier Claire (1), Fougerousse Françoise (1), Asselbergs Folkert W (2) (3), Guedj Mickaël (1), Komajda Michel (4), Kotecha Dipak (5) (6), Lumbers R. Thomas (7), Tyl Benoît (1)

1 - Institut de Recherches Internationales Servier (Suresnes France), 2 - University Medical Center [Utrecht] (Heidelberglaan 1003584 CX Utrecht Netherlands), 3 - University College London (NW1 2DA London United Kingdom), 4 - Department of Cardiology, Hospital Saint Joseph and Sorbonne university (Paris France), 5 - Institute of Cardiovascular Sciences, University of Birmingham (Birmingham United Kingdom), 6 - University Hospitals Birmingham NHS Foundation Trust (Birmingham United Kingdom), 7 - Institute of Cardiovascular Science, University College London (NW1 2DA London United Kingdom)

Correspondance to: claire.baudier@gmail.com

Drug target Mendelian randomisation studies can help to predict the impact of protein activity modulation on diseases and therefore provide useful insights for drug development. While they have successfully predicted the success or failure of drugs targeting coronary artery disease (CAD), their applicability to Heart Failure (HF) is less established. Therefore, we aimed to provide proof of principle for drug target Mendelian randomisation studies in HF by exploring the effect of adrenergic receptors modulation on the development of HF.

We performed a two-sample drug target Mendelian randomisation study using genetic variants located in the cis-region of genes encoding the nine adrenergic receptors. Since robust data on adrenoreceptor expression or protein concentration were not available, relevant downstream traits (blood pressure and heart rate) were used as proxies for receptor activity, to weight cis genetic instruments. We were able to find at least one genetic variant associated at genome-wide significance with BP and/or heart rate in published genome-wide association studies (GWAS) for four adrenergic receptors (α1A, α2B, ß1and ß2). The effects of the selected variants on risk of HF were derived from the HERMES GWAS. Mendelian randomisation analyses confirmed the protective role of a lower ß1 activity on the risk of HF and provided evidence for a protective role of a lower α1A activity. Paradoxically, a detrimental effect was found for a lower α2B activity. No effect was established for ß2 modulation.

Furthermore, since CAD can cause HF, we performed a mediation analysis to determine whether the impact of adrenergic receptors on HF was mediated partially or not by an effect on the risk of CAD. To do so, we performed Mendelian randomisation analyses with the CARDIoGRAMplusC4D CAD risk GWAS and the HERMES HF risk GWAS adjusted for CAD risk using the multi-trait Conditional and Joint Analysis (mtCOJO) method. We found a protective effect of lower ß1 or α2B activities on CAD risk, but no evidence for an effect of α1A modulation. This suggested that there is little evidence that CAD mediates the ß1 effect on HF. Besides, the beneficial effect of a lower α2B activity was insufficient to counterbalance its detrimental effect on HF.

Our results demonstrate that drug target Mendelian randomisation can explore the effects of drug modulation on the risk of HF, and check for potential mediation effect through HF risk factors such as CAD.
Flexible framework for gene-based association studies using GWAS summary statistics

Belonogova Nadezhda (1), Svishcheva Gulnara (1) (2), Tsepilov Yakov (1) (3), Axenovich Tatiana (1) (3)

1 - Institute of Cytology and Genetics (Lavrentiev av., 10 Novosibirsk, 630090 Russia), 2 - Vavilov Institute of General Genetics (Gubkin str., 3 Moscow, 119311 Russia), 3 - Novosibirsk State University (Pirogov str, 2 Novosibirsk, 630090 Russia)

Correspondance to: tatiana.aksenovich@gmail.com

Gene-based association analysis using GWAS summary statistics is an effective and popular tool for mapping of common and rare genetic variants. The methods of gene-based analysis differ in power for the traits with different genetic architecture that is likely to vary between genes across the genome. Due to it, a combination of the different methods can be a universal approach to the gene-based association analysis. To calculate one-by-one the large number of the tests and combine them, several frameworks have been proposed recently. They differ by the number of traits and methods involved, input data type (individual phenotypes and genotypes or the GWAS summary statistics), and ways to use the functional annotations. All these frameworks have a disadvantage – they are inflexible. Each of them implies a fixed set of methods, weights and combinations of the functional annotations.

We propose a flexible framework for gene-based analysis using GWAS summary statistics. We modified our sumFREGAT package (Svishcheva et al., 2019) by adding ACAT method (Liu et al., 2019) and including the probabilities of SNPs being causal (Li et al., 2020) in burden test, SKAT, ACAT-V, tests using principal components and functional data analysis approaches. Inputs for the modified sumFREGAT include GWAS summary statistics (p-values and effect sizes), correlations between genotypes calculated using the same or reference sample, the matrix of weighting functions defined by the allele frequencies and the parameters of the beta distribution, the probabilities of SNPs being causal estimated using different functional annotations. In the new framework, the user can arbitrary define a set of methods for gene-based association analysis, set of MAF-based weighting functions, set of the probabilities of SNPs being causal. Then, we developed a special function that calculates all tests defined by user and combines their p-values by the Cauchy combination test (ACAT-O). The conditional gene-based analysis can be performed by calculating the conditional GWAS summary statistics with, for example, the GCTA-COJO package (Yang et al., 2012), followed by using them in sumFREGAT as input data.

In conclusion, the modified sumFREGAT package allows users to calculate wide range of arbitrary defined gene-based association tests differing from one another by statistical method, weighting function, functional annotation, and then to combine these tests.

Availability: https://cran.r-project.org/web/packages/sumFREGAT/index.html

Funding: Russian Foundation for Basic Research (20-04-00464), and by the Ministry of Education and Science of the RF via the Institute of Cytology and Genetics (project 0259-2021-0009 / AAAA-A17-117092070032-4).
Multi-trait analysis of multiple related cardiovascular traits identifies novel loci for fibromuscular dysplasia.

**Berrandou Takiy (1) (2), Georges Adrien (1), Bouatia-Naji Nabila (1)**

1 - Paris-Centre de Recherche Cardiovasculaire (56 rue Leblanc 75015 PARIS FRANCE France), 2 - Center for Quantitative Genetics and Genomics (Blichers Allé 20, 8830 Tjele Denmark)

Correspondance to: takiy.berrandou@gmail.com

Fibromuscular dysplasia (FMD) shares common genetic determinants with several cardiovascular diseases including hypertension, coronary artery disease (CAD), migraine, cervical artery dissection (CeAD), intracranial aneurism (IA).

Here we aimed to leverage existing genetic correlation between FMD and these diseases to increase the power for loci discovery using multi-trait analysis of GWAS (MTAG), a generalisation of inverse-variance-weighted meta-analysis that accounts for incomplete genetic correlation between traits and sample overlap. We also used functional mapping and annotation using ATAC-seq datasets generated in vascular cells and tissues to characterize FMD MTAG loci.

We used existing summary statistics of GWAS meta-analyses for FMD (1962 cases and 7100 controls), migraine (10647 cases and 350494 controls), CeAD (1393 cases and 14416 controls), IA (7495 cases and 71934 controls), and CAD (60801 cases and 123504 controls) in addition to systolic blood pressure (SBP: \( n \approx 700000 \)) and low density lipoprotein (LDL: \( n = 343621 \)). Using LD score regression, we first demonstrated substantial positive genetic correlations between FMD and SBP (rg=0.44, \( P=5e^{-10} \)), migraine (rg=0.32, \( P=1e^{-4} \)), intracranial aneurysm (rg=0.34, \( P=7e^{-6} \)), and CeAD (rg=0.78, \( P=0.01 \)), although this latter did not survive correction for multiple testing. Interestingly, FMD was negatively correlated with LDL cholesterol (rg=−0.19, \( P=8e^{-3} \)) but not with CAD (rg=−0.09, \( P=0.12 \)). We benchmarked our multi-trait analysis against the standard univariate FMD analysis which only identified four loci. A multi-trait analysis with FMD, SBP, Migraine, IA, CeAD, LDL and CAD increased the number of genome-wide significant loci for FMD by adding 95 novel loci. Effect directions of all lead SNPs at the 95 MTAG loci were consistently correlated (\( r^2=0.65, p=5e^{-13} \)) with their effects from the individual FMD GWAS, where only 5 out of 95 lead SNPs were nominally significant (\( p < 0.001 \)). We found significant enrichment of FMD MTAG loci among open chromatin regions generated by ATAC-Seq in key vascular tissues including aorta, tibial artery and coronary artery, which supports that genes located in identified loci are likely to be regulated by arterial specific regulatory elements.

In conclusion, leveraging genetic correlation between carefully selected cardiovascular diseases and traits is an effective strategy to increase gene discovery pace for neglected vascular diseases like FMD, where only small samples are available. Ongoing local genetic correlation analyses between traits at the MTAG loci are expected to strengthen the current findings, and help triangulate evidence toward the causality in the diseases we investigated.
Population stratification refers to systematic genetic differences in allele frequencies between populations. Genetic association studies look for genetic differences between cases and controls that could be associated with the occurrence of a disease. Therefore, association tests may suffer from an inflation of type I error when cases and controls come from two different populations, for example two different countries. While population stratification has been studied for years, it remains poorly investigated at a fine-scale level, for example within a country. Here, we studied whole-genome sequence data from the FranceGenRef project gathering French individuals from the general population mainly recruited from the North-West of France against data on the MARTHA cohort gathering individuals with venous thromboembolism recruited in the South-East of France. Despite a comprehensive quality control, we observed a stratification between the two sets of individuals which resulted in a global inflation on common and rare variant association tests. A stringent selection of cases and controls to try homogenising the two groups also resulted in inflated associated tests, while the stratification seemed less important in the Principal Component Analysis (PCA). This leads to the conclusion that cases and controls recruited in different parts of France could not be sufficiently matched to avoid an inflation of association tests. Moreover, focusing only on the FranceGenRef sample where information on grandparent birth places is available, we were able to distinguish geographical regions within the North-West of France on PCA, showing that stratification can be detected at very fine geographic scales.

This work confirms that population stratification is present at a country-scale. If not properly taken into account, it can lead to an inflation on genetic association tests even if the stratification is not detected using classical methods such as PCA. This highlights the importance of getting geographically-matched groups of individuals in association tests, especially when the focus is on rare variants, and the need to better characterize genetic structures within countries at finer geographic scales.
Native American Ancestry, Genetic Susceptibility to critical Covid-19, and Covid-19 Case-Fatality Rates in Chile

Boekstegers Felix (1), Ruiz-Linares Andres (2), Rothhammer Francisco (3), Lorenzo Bermejo Justo (4)

1 - Institute of Medical Biometry and Informatics, Heidelberg University (Im Neuenheimer Feld 130.3 D-69120 Heidelberg Germany), 2 - Department of Genetics (Evolution and Environment, University College London, London, United Kingdom), 3 - Instituto de Alta Investigación, Tarapacá University, Arica, Chile (Tarapacá University, Arica Chile), 4 - Institute of Medical Biometry and Informatics, Heidelberg University (Im Neuenheimer Feld 130.3, 69120 Heidelberg Germany)

Correspondance to: boekstegers@imbi.uni-heidelberg.de

The Covid-19 Host Genetics Initiative and GenOMICC identified eight variants associated with critical Covid-19 illness based on individuals of European ancestry.[1, 2] Since the extent of infection, hospitalization and death rates due to Covid-19 seem to depend on ethnicity,[3, 4] we examined the relationship among Native American ancestry, genetic susceptibility to critical illness and case-fatality rates in Chile.

Based on population-wide Chilean genotype data from the Consortium for the Analysis of the Diversity and Evolution of Latin America (CANDELA), we inferred individual Native American ancestry proportions using ADMIXTURE, and calculated individual weighted polygenic risk scores relying on the reported odds ratios.[5, 6] Linear robust regression was used to evaluate the association between ancestry and genetic risk to critical Covid-19. Case-fatality rates were calculated based on publicly available data from the Chilean Ministry of Science’s Covid-19 repository, [7] and their dependence on ancestry and genetic risk to critical Covid-19 was assessed based on community-aggregate data, applying bootstrapping to consider the uncertainty of the estimated communal averages.

Inspection of the allele frequencies in the 1000 Genomes Project revealed that risk variants associated with critical Covid-19 were more frequent in Latin Americans.[8] Individual polygenic risk scores showed a large variability in Chile and pointed to a lower genetic susceptibility to critical illness in the central regions of the country, which are also characterized by lower proportions of Native American ancestry. Based on 1,884 genetically admixed Chileans from the whole country, each 1% increase in the Native American ancestry proportion translated into a 0.27% higher genetic risk of critical disease (95%CI 0.12% to 0.43%, P=0.0007). The geographic distribution of case-fatality rates did not reflect the pattern of genetic susceptibility. Based on 40 Chilean counties and the Santiago Metropolitan region, which comprise 73% of the Chilean population, we found no evidence for an association between the genetic risk score and the case-fatality rate (P=0.42).

An important limitation of this study was the lack of genotype data from Chilean Covid-19 patients to examine the genetic associations with critical Covid-19 illness in non-European populations. Research to identify and address the causes of ethnic differences in Covid-19 severity is urgently needed.

References:
1. doi:10.1056/NEJMoa2020283
2. doi:10.1038/s41586-020-03065-y
3. CDC COVID-19 Hospitalization and Death by Race/Ethnicity
4. UK Biobank SARS-CoV-2 Serology Study
5. doi:10.1371/journal.pgen.1004572
6. doi:10.1101/gr.094052.109
7. Base de datos CoVID-19 Santiago de Chile: Ministerio de Ciencia, Tecnología, Conocimiento, e Innovación
8. doi:10.1038/nature15393
Can disease risk prediction hold its promises?

Clerget-Darpoux Françoise (1)

1 - Institut des Maladies Génétiques Imagine [Paris]- (24, Boulevard du Montparnasse 75015 Paris France)
Correspondance to: francoise.clerget@inserm.fr

The number of studies and papers providing Polygenic Risk Scores (PRS) for diseases with complex and unknown etiology has increased exponentially during the last decade. Their authors claim that PRS may be used as a prediction tool and even, for some diseases, that PRS prediction is equivalent to the one given to a rare mutation carrier. Apart from the very simplistic assumptions under which PRS computations are performed and numerous questions about their validity, we argue that PRS only provide a classification of individuals in terms of relative risk in a given population, but not an individual risk to be affected.
Set-based testing of proteomic features with large portions of missing values – a bootstrapping approach with optimal data usage

Debrabant Birgit (1)

1 - University of Southern Denmark (5320 Odense Denmark)

Correspondance to: bdebrabant@health.sdu.dk

Proteomic markers often exhibit large portions of missing values. This constitutes a serious challenge when a set of markers is considered as multivariate outcome in set-based analyses. Common approaches are data shortening resulting in considerable data loss, and imputation of missing values typically imposing strong assumptions on the underlying concealment process.

Here we present an alternative in the context of MANOVA with high proportions of missings in the outcome variables. Our approach makes optimal usage of the available data and relies on feature-wise estimates of the coefficient matrix. We discuss test statistics for general linear hypotheses suitable for the p>N case, i.e. when the number of outcome features exceeds the number of individuals, and present a corresponding bootstrapping approach. Simulations are used to estimate type I/II errors and to compare our method with alternative data shortening.
An assessment/comparison of ground truth datasets for the identification of non-coding variants using machine learning

Delehelle Franklin (1), Gonzalez Aitor (2), Ballester Benoît (2), Rihet Pascal (2), Berthelot Camille (1), Roest Crollius Hugues (1)

1 - Institut de biologie de l’ENS Paris, CNRS, Inserm, 46, Rue d’Ulm 75005 Paris France, 2 - Technologies avancées pour le génôme et la Clinique, Parc scientifique de Luminy - 163 avenue de Luminy 13288 Marseille cedex 9 France,

Correspondance to: franklin.delehelle@bio.ens.psl.eu

Non-coding regulatory variants represent a new frontier in human WGS data interpretation. The challenge is all the more difficult because of our poor understanding of the molecular mechanisms underlying the regulation of gene expression. Machine learning algorithms have been very helpful in capturing a range of parameters associated with such variants, but the lack of large, varied and trustworthy positive control dataset has hampered progress in this field. Two publicly available datasets offer some hope in this regard, namely the SuRE and the GTEx datasets. The SuRE protocol is an in-vitro assay that has identified 19,237 and 14,183 functional non-coding regulatory variants in K562 and HepG2 cell lines respectively. The GTEx approach identifies statistically significant associations between naturally occurring variants in genomes of human donors and the expression of nearby genes. Any statistical analysis wishing to make use of either data as positive control must account for the compromises and limitations associated with them. Here we present a comparative analysis of the two datasets on liver compared to the HepG2 line on the one hand and whole blood compared to the K562 line on the other one. We investigate their relative intrinsic strengths and weaknesses, then detail when and how their findings correlate, aiming to shed light on the pitfalls of their use in derived studies, especially regarding the discrepancies in regulatory variants counts and position.
Robust Huber-LASSO improves prediction of protein, metabolite and gene expression levels from individual genotype data

Deutelmoser Heike (1) (2), Scherer Dominique (2), Brenner Hermann (3) (4) (1), Waldenberger Melanie (5), Study Interval (6), Suhre Karsten (7) (8), Kastenmüller Gabi (8), Lorenzo Bermejo Justo (2)

1 - Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT) (Im Neuenheimer Feld 460, 69120 Heidelberg Germany), 2 - Institute of Medical Biometry and Informatics, Heidelberg University (Im Neuenheimer Feld 130.3, 69120 Heidelberg Germany), 3 - German Cancer Consortium, German Cancer Research Center (DKFZ) (Im Neuenheimer Feld 280, 69120 Heidelberg Germany), 4 - Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ) (Im Neuenheimer Feld 581, 69120 Heidelberg, Germany Germany), 5 - Institute of Genetic Epidemiology, Helmholtz Zentrum München (Ingolstädter Landstraße 1, 85764 Neuherberg Germany), 6 - University of Cambridge (The Old Schools, Trinity Ln, Cambridge CB2 1TN United Kingdom), 7 - Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar (Education City, Qatar Foundation, 24144 Doha Qatar), 8 - Institute of Computational Biology, Helmholtz Zentrum München (Ingolstädter Landstraße 1, 85764 Neuherberg Germany)

Correspondance to: deutelmoser@imbi.uni-heidelberg.de

Least absolute shrinkage and selection Operator (LASSO) regression performs a selection of variables using a penalty term and is commonly used to detect associations between single nucleotide polymorphisms (SNPs) and a molecular phenotype of interest. The performance of the standard LASSO method, which uses the least squares loss function, shows a strong dependence on a few individuals with phenotypes or genotypes that deviate from the majority of the study population – typically referred to as outliers or high-leverage observations. Robust methods have been developed to obtain more reliable statistical results, but these have rarely been applied to genetic association studies.

We investigate the benefits and limitations of a robust version of the standard LASSO that utilizes the Huber loss function. We examine the stability of the regularization parameter, the cross-iteration concordance of the model, the false-positive rate, the true-positive rate, and the predictive accuracy of standard LASSO and robust Huber-LASSO. Prediction was performed using fivefold cross-validation and its accuracy was assessed by the squared correlation between observed and predicted values. Extensive computer simulations were complemented by the analysis of real protein, metabolite, mRNA gene expression, and SNP data.

Our main finding was that robust Huber-LASSO appears to outperform standard LASSO in protein, metabolite, and mRNA prediction based on individual genotype data and thus represents a valuable alternative to standard LASSO.
Increasing the Efficiency of SNP Haplotype Regional Heritability Mapping: joint mapping of common and rare variation affecting complex traits

Eilidh Fummey (1), Bussoa Aadarsh (2), Proeme Arno (2), Navarro Pau (1), Knott Sara (3), Canela-Xandri Oriol (1), Haley Chris (1) (4)

1 - MRC Human Genetics Unit at the MRC Institute of Genetics and Molecular Medicine (University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XU, UK United Kingdom), 2 - Edinburgh Parallel Computing Centre (Bayes Centre, 47 Potterrow, Edinburgh EH8 9BT United Kingdom), 3 - Institute of Evolutionary Biology (University of Edinburgh, Edinburgh, EH9 3FL, UK United Kingdom), 4 - The Roslin Institute, Royal (Dick) School of Veterinary Studies (The University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK United Kingdom)

Correspondence to: s1533194@ed.ac.uk

Identification of rare genetic variation contributing to complex traits has led to novel biological insights. However detecting the effects of rare variants presents a statistical challenge. Thus, bespoke statistical methods for tackling this problem have been developed. As sequence data and biobank-scale cohorts become the norm, such methods must be scalable to continue to provide valuable insights.

Here we achieve a reduction in the compute time and disk space required to run one such method, SNP Haplotype Regional Heritability Mapping (SNHap-RHM). SNHap-RHM estimates the trait’s genetic variance attributable to genomic regions defined by recombination boundaries. This genetic variance is partitioned into two components dependent on different genomic relationship matrices (GRMs) for each region. One of these GRMs is constructed using relatedness scores based on SNP allele sharing between individuals; and the other uses haplotype allele sharing (Nagamine et al., 2012; Shirali et al., 2018). By incorporating haplotype-based estimates of relatedness, SNHap-RHM captures variation introduced by rare causal alleles associated with haplotype blocks.

Previously, haplotype-based GRMs have been calculated using a simple Fortran program. By reformulating the base equations used for the haplotype-based GRM calculation, the program was rewritten in C to take advantage of improved algorithms and the highly optimised linear algebra library ScalAPACK. ScalAPACK also allows for parallelisation of the calculation on distributed memory architecture. The novel implementation completed almost 300x faster than the original Fortran implementation. As well as decreasing runtime, this allows the user to scale up the compute resource as their analyses grow in size.

The reformulated algorithm was integrated into the software package DISSECT (Canela-Xandri et al., 2015). DISSECT was adapted to read and filter phased bgen files, eliminating the need for intermediate files containing individual-level data that were required by the Fortran program. This, in combination with no longer needing to store the GRMs, significantly reduces the disk space required to run an analysis. For example, this equates to a ~4GB reduction in stored files for ~28,000 individuals. Additionally, avoiding reading or writing GRMs further reduces analysis run time and the number of input-output operations.

In summary, our work increases the efficiency of SNHap-RHM in preparation for its application to sequence data and biobank-scale populations. This tool will be made publicly available in future DISSECT releases.

Identification of gene modules shared by childhood-onset asthma and Immunoglobulin-E levels by integrated network analysis of multi-omics data

Eranti Pradeep (1), Bouzigon Emmanuelle (1), Demenais Florence (1), Egea Cooperative Group And The 1 - Group of Genomic Epidemiology of Multifactorial Diseases (UMRS-1124, Inserm, Paris France)

Correspondance to: pradeep.eranti@inserm.fr

Multifactorial diseases, such as asthma, are caused by the effects of multiple genetic and environmental factors. Genome-wide association studies (GWAS) have identified many genes associated with asthma and asthma-related phenotypes (such as Immunoglobulin E (IgE), a key mediator of allergic response). However, these genetic risk variants contribute only to part of the heritable component of the disease. The remaining contribution may be partly due to variants with small marginal effects that interact with other genetic variants and the environment and are not detected by GWAS as classically used. Network-based methods, which integrate biological knowledge, i.e., molecular interactions among proteins/genes, with GWAS results (summary statistics), can be powerful to identify sets of interacting genes (gene modules) enriched in disease association signals. This approach can be applied to genome-wide studies of other types of omics data (e.g., epigenome-wide association studies (EWAS)). Furthermore, integrated network analysis of multiple omics datasets can prove successful in improving our understanding of the molecular mechanisms underlying complex diseases such as asthma.

We proposed a network-analysis based strategy that integrates the results from two omics datasets (summary statistics from childhood-onset asthma GWAS and IgE EWAS) with a publicly available biological network (STRING database) to identify two gene modules enriched in association signals towards the phenotype of interest as well as genes and biological pathways shared by these two modules. Our proposed analysis strategy employs an efficient and exact method, SigMod[1], formulated as a binary quadratic optimization problem that can be solved through graph min-cut algorithms (Liu Y et al. Bioinformatics (2017)). SigMod allows the identification of gene modules that are highly interconnected and enriched with high association signals. Biological interpretation of the identified modules was performed by enrichment analysis and interrogating the KEGG pathways database.

Our strategy identified two gene modules: a module associated with asthma containing 312 genes and 1,021 interactions; another module associated with IgE comprising 253 genes and 606 interactions. We found these two modules shared relevant genes, such as DRD3 (dopamine receptor D3), CSF3 (colony stimulating factor 3), and TRIM38 (tripartite motif containing 38), which have been reported in recent GWAS of asthma or immune response phenotypes. These two modules also shared biological pathways related to Alzheimer’s disease and immune-related mechanisms. These results support the proposed strategy, which can be extended to larger and additional omics datasets to understand further the complex mechanisms underlying asthma.

[1] Available at https://github.com/YuanlongLiu/SigMod
Polygenic analysis of the complex genetic markers of longevity

Erdman Vera (1), Tuktarova Ilsiar (1), Timasheva Yanina (1), Nasibullin Timur (1)

1 - Institute of Biochemistry and Genetics of Ufa Scientific Centre (Ufa Russia)

Correspondance to: danivera@mail.ru

Increased life expectancy with preserved physical and cognitive abilities is comprises such a complex phenomenon as longevity. Long-livers are people who reach the age exceeding the population average by far. They are resistant to common socially significant, age-related diseases (ARD). The multifactorial nature of longevity refers to the idea of the existence of complex genetic markers, certain genetic combinations, which, in senile physiological background specific for the late stages of ontogenesis, are necessary for the development of the long-liver phenotype.

With the aim to identify complex genetic markers of longevity we searched and characterized the combination of functional polymorphic loci of some genes predisposing ARD at all stages of human ontogenesis, from birth to longevity. The total group was complete of 2511 unrelated persons aged 1 to 113, residents of the Republic of Bashkortostan (Russia). In this study, we included 20 genes of antioxidant defense, glutathione metabolism and transcription factors providing redox homeostasis. The search for multilocus markers was performed in the APSampler program, using the Monte Carlo Markov Chains (MCMC) method. This program includes a set of functions necessary for the search for multilocus markers of predisposition, including the assessment of associations for the found combinations using Fisher's exact test, the ability to analyze the association for a combination of specific forcedly selected alleles or genotypes.

We identified a total of 169 multilocus patterns, related with longevity. 116 of them are predisposing longevity, 53 are counteracting ones. Allele/genotype combinations CATT+MSRA T/T+GPXIT, HIF1AC+MSRA+T+SOD1A/A, MSRA+T+SOD1A/A (among Russian), PONI G/A+PON2C+NQO1C+GSTP1G, PON2C+SOD1A/A, HIF1AT+PON2G+SOD1A+SOD2T (among Bashkir) and GSTP1A+GSR (among Tatars) were found with the most pronounced statistical indicators (OR>4, Pcor<0.01). Among individuals over 90 years old GSRT/G+SOD1A+SOD2T and NQO1C+GSTP1*G/G pattern is 7-times less common in Tatar subgroup (OR=0.137, Pcor=4.02×10^{-9}).

Thus, we defined some complex DNA-markers of longevity. Identified probable gene networks involved in the development of a stable highly adaptive phenotype allowing to avoid age-related pathologies or minimize their negative effect.

The study was supported by the grant RFBR and ASA 19-54-40007 and the megagrant by the Government of Russian Federation (2020-220-08-2197).
Stabilization of cultural innovations depends on population density: the case of rock art

Eriksson Anders (1), Walker Richard (2), Ruiz Camille (3), Newton Taylor Howard (2), Casalegno Francesco (2)

1 - cGEM, Institute of Genomics, University of Tartu (Riia 23b 51010 Tartu Estonia), 2 - Blue Brain Project, Ecole Polytechnique Fédérale de Lausanne (Campus Biotech Chemin des Mines 9 1202 Geneva Switzerland), 3 - Ateneo de Manila University (Manila Philippines)

Correspondence to: anders.eriksson@ut.ee

Demographic models of human cultural evolution have high explanatory potential but weak empirical support. Here we use a global dataset of rock art sites and climate and genetics-based estimates of ancient population densities to test a new model based on epidemiological principles. The model focuses on the process whereby a cultural innovation becomes endemic in a population. It predicts that this cannot occur unless population density exceeds a critical threshold. Analysis of the data, using a Bayesian statistical framework, shows that the model has stronger empirical support than a proportional model, where detection is directly proportional to population density, or a null model, where rock art detection ratios and population density are independent. Results for different geographical areas and periods are compatible with the predictions of the model and confirm its superiority with respect to the null model. Re-analysis of the rock art data, using a second set of independent population estimates, again supports the superiority of the model. Although the available data is sparse and it is impossible to exclude all possible sources of bias, this is evidence that population density above a critical threshold may be a necessary condition for the maintenance of rock art as a stable part of a population’s cultural repertoire. Methods similar to those described can be used to test the model for other classes of archaeological artifact and to compare it against other models.
Assessment of significance of conditionally independent GWAS signals

Ghasemi Sahar (1) (2) (3), Teumer Alexander (1) (3), Wuttke Matthias (4), Becker Tim (5) (3)

1 - DZHK (German Center for Cardiovascular Research), partner site Greifswald (Germany Germany), 2 - Department of Psychiatry and Psychotherapy, University Medicine Greifswald (Ellernholzstr. 1-2 17475 Greifswald Germany), 3 - Institute for Community Medicine, University Medicine Greifswald (Abteilung SHIP-KEF Walther-Rathenau-Str. 48 Germany), 4 - Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg (Freiburg, Germany Germany), 5 - 3xValue GmbH, Ratingen (GmbH, Ratingen Germany)

Correspondence to: sahar.ghasemi@uni-greifswald.de

Motivation: Multiple independently associated SNPs within a linkage disequilibrium (LD) region are a common phenomenon. Conditional analysis has been successful in identifying secondary signals. While conditional association tests are limited to specific genomic, they are benchmarked with genome-wide scale criterion, a conservative strategy.

Method: Within the weighted hypothesis testing framework, we developed a “quasi-adaptive” method that uses the pairwise correlation (r²) and physical distance (d) from the index association to construct priority functions G = G(r², d), which assign a SNP-specific α-threshold to each SNP. Family-wise error rate (FWER) and power of the approach were evaluated via simulations based on real GWAS data. We compared a series of different G-functions.

Results: Simulations under the null hypothesis on 1100 primary SNPs confirmed appropriate empirical FWER for all G-functions. A G-function which down-weighted SNPs at higher distance step-wise and gave more emphasis on d than on r² had overall best power. It also gave the best results in application to the real data sets. As a proof of concept, “quasi-adaptive” method was applied to GWAS on free thyroxine (FT4), inflammatory bowel disease (IBD), and human height. Application of the algorithm revealed 5 secondary signals in our example GWAS on FT4, 5 secondary signals in case of the IBD, and 19 secondary signals on human height, that would have gone undetected with the established Genome-wide threshold α = 5×10⁻⁸.
A Binomial Regression Model For Genetic Association Mapping Combining Population-based And Family-based Designs

Ghosh Saurabh (1), Barick Sourojyoti (1)

1 - Indian Statistical Institute [Kolkata] (203, Barrackpore Trunk Road, Kolkata, West Bengal 700108 India)

Correspondance to: saurabh@isical.ac.in

The models suggested in Multiphen (O’Reilly et al., 2012) and BAMP (Majumdar et al, 2015) provide an alternative to study population-based genotype-phenotype association by exploring the dependence of genotype on phenotype instead of the naturally arising dependence of phenotype on genotype. This reversal of the regression model, while has no impact on the inference on association, provides the flexibility of incorporating multiple phenotypes without the requirement of making any a priori assumptions on the correlation structure of the vector of phenotypes. In this project, we aim to extend the BAMP (Binomial regression-based Association of Multivariate Phenotypes) model to include family-based data in addition to population level data. The family-based Binomial Regression model involves the conditional distribution of identity-by-state scores on the absolute or squared sib-pair phenotype differences. Since the marginal distributions of the response variables in the two regressions (that is, under no genetic association) are different, we explore an integrated test constructed jointly on the estimated regression parameters corresponding to the two designs. We extend the framework to multivariate phenotypes and carry out extensive simulations under different genetic models to evaluate the power of the proposed integrated test.
Dissecting the shared genetics between type 2 diabetes and depression using a multi-phenotype GWAS approach

Gichohi Jared (1), Balkhiyara Zhanna (2), Kaakinen Marika (2), Nouwen Arie (3), Prokopenko Inga (2)

1 - University of Lille (1 place de verdun 59045 Lille Cedex France), 2 - University of Surrey (University of Surrey, Guildford Surrey GU2 7XH United Kingdom), 3 - Middlesex University [London] (Hendon Campus, The Burroughs, Hendon;NW4 4BT;London United Kingdom)

Correspondance to: jared.maina@univ-lille.fr

Type 2 diabetes (T2D) and depression are two common multi-factorial diseases with a huge global disease burden. Epidemiological evidence suggests a bi-directional relationship exists between the two. Furthermore, analysis based on GWAS results of the two diseases has shown a positive genetic correlation between T2D and depression. However, the mechanisms underlying their co-morbidity remain undefined. Using the UK Biobank, we employed the MTAG software to perform a multi-phenotype GWAS (MP-GWAS) of T2D and depressive phenotypes to increase the power for loci identification, improve effect size estimates and provide suggestions for multi-phenotype effects, such as pleiotropy. We compared two diagnostic criteria for depression: clinically diagnosed major depressive disorder (MDD) (n = 91,796; 5,357 cases, 86,439) and depressive symptoms based on self-report assessed using the PHQ-9 questionnaire (n = 153,567) and evaluated how each performs in a MP-GWAS with T2D (n = 482,985; cases = 19,344 cases, 463,641 controls). We applied functional transcriptome-wide analysis (TWAS) based on various tissue prediction models in the GTEx database. MP-GWAS increased the number of genome-wide significant signals for depressive symptoms from three to eight, while no signals were identified for MDD compared to the respective single-phenotype GWAS. Seven of the eight loci identified for depressive symptoms after MP-GWAS were established for T2D including TCF7L2, CDKAL1 and IGF2BP2. TWAS results revealed enrichment of significant gene associations in the same tissues for depressive symptoms and T2D including the frontal cortex (CDKAL1), substantia nigra (BETL1), amygdala (HSPA1B), hypothalamus (EIF2S2P3), whole blood (HLA-DRB1), adipose subcutaneous tissue (IRS1), and muscle (HLA-DRA). These genes are involved in biological processes and canonical pathways (e.g., cell transformation, immune system and glucose-insulin signalling pathways) that are important for the pathogenesis of T2D and depression. Our results indicate that analysis of depression defined based on self-reported depressive symptoms performs better in a MP-GWAS with T2D than the strictly defined MDD phenotype. Also, genes in the shared depression/T2D loci exert tissue-specific expression. Our results boost our understanding of the underlying mechanisms of these two complex disorders.
Genetic factors driving immune differences between monozygotic twins clinically discordant for multiple sclerosis

Herrera Rivero Marisol (1), Janoschka Claudia (2), Stoll Monika (1), Klotz Luisa (2)

1 - Department of Genetic Epidemiology, Institute of Human Genetics, University of Münster (Albert-Schweitzer-Campus 1, Building D3, Domagkstr. 3, 48149 Münster Germany), 2 - Department of Neurology with Institute of Translational Neurology, University Hospital Münster (Albert-Schweitzer-Campus 1, Building A1, 48149 Münster Germany)

Correspondence to: marisol.herrera@uni-muenster.de

Multiple sclerosis (MS) is a demyelinating autoimmune disease of the central nervous system. The immunopathology of MS is complex, heterogeneous, and involves many subpopulations of adaptive as well innate immune cells. It is yet unknown what causes the immune system to attack myelin sheaths. However, both genetic and environmental factors play important roles. In monozygotic twins, there is 25% chance that both co-twins will develop MS when one of them has been diagnosed with the disease. A previous study of 43 monozygotic twins clinically discordant for MS found that the immune variation between MS-affected/-unaffected co-twins could be better explained by twinship (56%) than by MS diagnosis (1-2%). Here, using SNP array, FACS and RNA-seq data from the same twin cohort, we studied (i) how genetic variants associate with differences in immune cell populations and with the CD4+ T cell transcriptome, and (ii) how the CD4+ T cell transcriptome changes between MS-discordant co-twins. We found 5,867 genome-wide significant (p<5x10^-8) quantitative trait loci (QTLs) for 57 FACS cell subsets. Variant annotations and functional enrichment analyses were obtained for FACS-QTLs in phenotypes clumped in the following categories: top ranking (24 phenotypes prioritized according to the level of support of the identified QTL signals), innate, adaptive, B, CD4 and CD8 cells. Specific features between innate and adaptive immune cells, as well as between B, CD4 and CD8 cells were identified. In CD4+ T cells, 12,581 genes were differentially expressed between MS-discordant co-twins. These genes were enriched in immune and brain tissues, as well as in immune and stress response pathways. Genetic associations with variance in gene expression were only found at the suggestive threshold of genome-wide significance (p<1x10^-5), with 45 independent signals. From these, there were 8 overlaps with FACS-QTLs in monocytes and 29 genes differentially expressed in CD4+ T cells. Overall, we confirmed a strong involvement of adaptive immunity in MS. Moreover, our genetic analyses revealed important enrichments in gene sets related to neuronal development and function. Taken together, our results suggest that an intrinsic neuronal manifestation of MS may exist, whose interactions with immune genes might determine the development of the disease.
Can imputation in a European country be improved by local reference panels? 
The example of France

Herzig Anthony (1), Velo-suárez Lourdes (1), Redon Richard (2), Deleuze Jean-François (3), Génin Emmanuelle (1), Consortium Frex, Consortium Francegenref

Population based genotype imputation remains a widely used technique for enriching datasets of genotyped or low-coverage sequenced individuals. Advances in software capabilities have been rapid, enormous haplotype reference panels have been assembled, and dedicated computation servers have been created.

Previous comparisons of different strategies have shown the benefits of public reference panels that are large and diverse as well as the importance of ‘local reference panels’ or ‘study specific panels’ (SSPs). Leveraging both resources simultaneously is unfortunately often impractical. This is due to the fact that the largest public imputation panels are accessible only through online servers which can present certain limitations for the user. Here, using two panels of sequencing data from France, we evaluate the trade-offs between server-based imputation and pursuing an in-house imputation pipeline using an SSP.

France has a population with extensive internal fine-structure; and while public reference panels contain an abundance of European genomes, there include few French genomes. Intuitively, using an SSP for France would therefore be likely beneficial. To investigate, we impute 550 French individuals with array and whole-exome sequencing data from six different regions in France, using either the University of Michigan imputation server with the Haplotype Reference Consortium panel, or in-house using a combined panel of 850 whole-genome sequenced French individuals with haplotypes that are publically available through the 1000 Genomes Project.

With approximate geo-localization of both our target and SSP reference individuals we are able to pinpoint the relevance of difference groups within the proposed SSP for different groups of target individuals. This helped to illustrate different scenarios where SSP-based imputation would be preferred over server-based or vice-versa. We could also show to a high degree of resolution how the proximity of the reference panel to a target individual determined the accuracy of both haplotype phasing and genotype imputation.

Finally, we put forward a pragmatic solution where server-based and SSP-based imputation outcomes can be combined based on comparing posterior genotype probabilities. If the SSP used is appropriate, such a combined approach can give a level of imputation accuracy markedly in excess of what could be achieved with either strategy alone.
Non-additive genetic variance for complex traits is notoriously difficult to estimate in non-laboratory species, including humans, since the environment cannot be controlled so that non-additive genetic variance can be confounded with environmental covariance among relatives. In principle, non-additive variance attributable to common DNA variants can be estimated from a random sample of unrelated individuals with genome-wide SNP data. In this design, the proportion of phenotypic variance explained by epistatic effects can be quantified in a classical variance component framework using the Hadamard product of a Genomic Relationship Matrix (GRM) with itself. Here, we propose to jointly estimate the proportion of variance explained by additive, dominance and additive-by-additive genetic variance in a single analysis model. We derive the sampling variances of the additive-by-additive genetic variance from the properties of the Hadamard product of the GRM and apply the model to 70 complex traits using 254,679 unrelated individuals from the UK Biobank and 1.1M genotyped and imputed SNPs. While we found strong evidence for additive variance (average additive variance across traits = 0.207 +/- 0.001), we did not find evidence of significant non-additive variance across traits (average dominance variance across traits = 0.001 +/- 0.001, and additive-by-additive variance = 0.058 +/- 0.032). Our results provide new evidence that genetic variance for complex traits is predominantly additive, and that sample sizes of many millions of unrelated individuals are needed to estimate epistatic variance with sufficient precision.
Computational prediction of the pathogenic status of cancer-specific somatic variants

Hu Pingzhao (1), Feizi Nikta (1)

1 - University of Manitoba [Winnipeg] (Winnipeg, Manitoba Canada R3T 2N2 Canada)

Correspondence to: pingzhao.hu@umanitoba.ca

Background: In-silico interpretation approaches are shown to be promising in increasing the clinical utilization of genetic tests by classifying variants into well-defined pathogenic and non-pathogenic clinical groups. Current prediction tools are mostly trained based on the characteristics of germ-line variants and utilized for classifying both germline and somatic variants, which may result in biased predictions in the event of classifying somatic variants. Considering the critical role of somatic variants in cancer occurrence and progression establishing cancer-specific prediction tools which are designed solely based on the characteristics of somatic variants is of high importance.

Methods: We established a gold standard dataset exclusively for cancer somatic single nucleotide variants (SNVs) collected from the catalog of somatic mutations in cancer (COSMIC). We portrayed the genomic characteristics of the somatic variants located in coding and non-coding regions of the genome by defining 80 and 65 different genomic features in two distinct gold standards, respectively, which were used to train two support vector machine classifiers trained exclusively based on the characteristics of cancer somatic single nucleotide variants for coding and non-coding regions, respectively.

Results: Using the support vector machine (SVM) classification method we designed two distinct models achieving the AUC (area under the ROC curve) of 0.94 and 0.89 in classifying the pathogenic status of somatic variants located in coding and non-coding regions of the genome, respectively. We compared the performance of our models with two well-known classification tools including FATHMM-FX and CScape which are not originally designed for classifying cancer somatic variants. Our models outperformed both tools in classifying variants located in both coding and non-coding regions of the genome. Furthermore, we applied our models to predict the pathogenic status of somatic variants identified in young patients (under 45 years of age) with breast cancer from METABRIC and TCGG-BRCA studies.

Conclusion: The computational models designed originally based on the characteristics of cancer somatic variants are proven to have high discriminative power for classifying the variants into pathogenic and non-pathogenic groups. In addition, the potential application of the models in revealing the functional and clinical impacts of cancer somatic variants is strongly suggested through our study.
Complex gene regulation: higher-order interactions in single cell expression data

Del Debbio Luigi (1), Jansma Abel (2) (1), Khamseh Ava (2) (1) (3), Ponting Chris (2)

1 - Higgs Centre for Theoretical Physics, School of Physics and Astronomy (The University of Edinburgh, James Clerk Maxwell Building, Kings Buildings, Mayfield Rd, Edinburgh EH9 3JZ United Kingdom), 2 - MRC Human Genetics Unit, Institute of Genetics and Cancer (University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU United Kingdom), 3 - School of Informatics (University of Edinburgh, Edinburgh EH8 9AB United Kingdom)

Correspondance to: a.a.a.jansma@sms.ed.ac.uk

Traditional techniques to infer gene regulatory networks are based only on properties of pairs of genes. Common examples include (partial) correlation and mutual information. This simplification not only hinders the correct estimation of pairwise interactions, it also ignores the complexity of higher-order interactions and combinatorial gene regulation. Furthermore, these pairwise quantities are introduced to serve as proxies for the more interesting but abstract notion of a causal interaction that corresponds to a biological process.

We consider a model-independent nonparametric estimator of symmetric interactions that coincides with the definition of interaction in statistical physics. This estimator can be shown to both analytically and numerically recover the ground truth in simulated binary systems on a lattice, and it separates each order of interaction. We apply this estimator to extract (higher-order) interactions from binary single cell gene expression data.

Estimating the interactions directly from data requires a deeply sampled state space, which is generally not available in expression data where the state space grows exponentially with the number of genes. Exploiting the conditional independencies between the variables can make inference much less data-hungry. We randomly select 10k cells from each of four cell types in the 10X Million Cell Dataset of an embryonic mouse brain, keeping only the 500 most highly variable genes, and use two causal discovery algorithms, the constraint-based PC algorithm and a hybrid MCMC method, to discover conditional independencies amongst genes and obtain the causal graphs. Knowledge of these graphs makes the inference of interactions from expression data tractable and we find significant genetic interactions at first, second, and third order, with differences in interaction graph structure and density exhibited between cell types. Across cell types, we find hundreds of significant triplet interactions where the 95% confidence interval does not include zero.

By simulating systems on small causal graphs, we see that different underlying causal dynamics lead to different structures in the inferred interactions. This relationship between interactions and causal structure allows us to predict combinatorially interacting gene triplets, and distinguish various types of causal interactions.
The Impact of Cross-Species Gene Flow on Species Tree Estimation

Jiao Xiyun (1), Flouri Tomas (2), Rannala Bruce (3), Yang Ziheng (2)

1 - Southern University of Science and Technology (Shenzhen, Guangdong China China), 2 - University College London (University College London Gower Street London WC1E 6BT United Kingdom), 3 - University of California, Davis (Davis, CA, USA United States)

Correspondence to: jiaoxy@sustech.edu.cn

Recent analyses of genomic sequence data suggest cross-species gene flow is common in both plants and animals, posing challenges to species tree inference. We examine the levels of gene flow needed to mislead species tree estimation with three species and either episodic introgressive hybridization or continuous migration between an outgroup and one ingroup species. Several species tree estimation methods are examined, including the majority-vote method based on the most common gene tree topology (with either the true or reconstructed gene trees used), the UPGMA method based on the average sequence distances (or average coalescent times) between species, and the full-likelihood method based on multi-locus sequence data. Our results suggest that the majority-vote method is more robust to gene flow than the UPGMA method and both are more robust than likelihood assuming a multispecies coalescent (MSC) model with no cross-species gene flow. A small amount of introgression or migration can mislead species tree methods if the species diverged through speciation events separated by short time intervals. Estimates of parameters under the MSC with gene flow suggest the Anopheles gambia African mosquito species complex is an example where gene flow greatly impacts species phylogeny.
The effect of synaptic plasticity gene variants on working memory in mentally healthy individuals under gene-by-environment interactions

Kazantseva Anastasiya (1) (2), Enikeeva Renata (1), Davydova Yuliya (1), Takhirova Zalina (3), Mustafin Rustam (4), Lobaskova Marina (5), Malykh Sergey (5), Khusnutdinova Elza (1)

1 - Institute of Biochemistry and Genetics of Ufa Scientific Centre (Ufa Russia), 2 - Federal State Budgetary Educational Institution of Higher Education “Ufa State Petroleum Technological University” (Ufa Russia), 3 - Russian Academy of Education (Moscow Russia), 4 - Department of Medical Genetics and Fundamental Medicine, Bashkir State Medical University (Ufa Russia), 5 - Psychological Institute of the Russian Academy of Education (Moscow Russia)

Correspondence to: kazantsa@mail.ru

The mechanisms underlying individual predisposition to individual cognitive functioning, including working memory (WM), are complex, and genetic and environmental factors play a significant role. Together with the role of apolipoprotein E (APOE) in cognitive decline, which is frequently associated with an impaired working memory, cognitive processes were shown to be affected by synaptic plasticity genes. However, mathematical gene-environmental models, which can explain individual variability in working memory in non-demented population of young age, have to be established. The present study aimed to estimate the main effect of examined gene polymorphisms together with gene-environment interactions in individual differences in working memory in non-demented young adults.

The study included 897 mentally healthy individuals (79% women; 19.74±1.51 years) of Caucasian origin from Russia. The assessment of working memory was performed using Corsi block-tapping test under the Battery of cognitive tests (InLab, Goldsmiths, University of London). Examined environmental parameters included: birth order, sibship size, place of residence, preterm birth, socio-economic status, rearing in a complete/incomplete family, childhood adversity, rearing in a bilingual family, smoking, physical activity, weight at birth, maternal and paternal age at birth. The genotyping of BDNF rs6265, NRG1 rs6994992, SNAP25 rs363050, NRXN1 rs4971648, NRXN1 rs1045881, KIBRA rs17070145, APOE rs429358, rs7412, DNMT1A rs10418707, DNMT3A rs1550017, DNMT3B rs2424932 gene SNPs was performed using PCR-based KASP genotyping technology on “CFX96” DNA Analyzer (BioRad, USA). Statistical analysis included multiple linear regression followed by FDR-correction for multiple testing (PLINK v.1.09).

While controlling for ApoE alleles, the association of NRXN1 rs1045881 (β=-0.51, P=0.036) and NRG1 rs6994992 (β=1.12, P=0.012) with working memory was revealed, which indicated that rs1045881 T-allele and rs6994992 C-allele carriers (under the additive model) scored lower on working memory. Stratification by ApoE E2-allele showed stronger association with rs1045881 T-allele (β=-1.38, P=0.011) and WM, thus indicating a higher strength of association under the conditions of dementia-protective ApoE allele.

Subsequent multiple linear regression analysis revealed the GxE models, which indicated a modulating effect of bilingual rearing (β=-1.14, P=0.014) and maternal age at delivery (β=-0.09, P=0.015) on association of NRG1 rs6994992 T-allele with decreased WM in healthy young adults.

The present study evidenced in the necessity to correct for the ApoE genotype in statistical models, when analyzing non-demented individuals. We observed the involvement of synaptic plasticity genes (NRXN1, NRG1) together with rearing specificity in individual variation in working memory in healthy individuals.

The study was supported by the megagrant by the Government of the Russian Federation (2020-220-08-2197).
Identifying association between genotypes and patient care trajectories using patients’ networks

Lambert Judith (1) (2), Labatiet Louise (3), Leutenegger Anne-Louise (4), Jannot Anne-Sophie (1) (5), Baudot Anaïs (6) (7)

1 - Centre de Recherche des Cordeliers (15 rue de l’École de Médecine - Bâtiment B - 75270 Paris Cedex 06 France), 2 - Marseille medical genetics - Centre de génétique médicale de Marseille (Faculté de Médecine - Timone27, boulevard Jean Moulin13385 Marseille cedex 5 France), 3 - Maladies neurodéveloppementales et neurovasculaires (Hôpital Robert Debré48 Bd Serurier75019 Paris France), 4 - INSERM U1141 (Hôpital R. Debré, 48 Bd Serurier, Paris France), 5 - HEGP - Informatique - Département d’Informatique et Santé Publique [CHU HEGP] (20, Rue Leblanc 75908 PARIS CEDEX 15 France), 6 - Marseille medical genetics - Centre de génétique médicale de Marseille (Faculté de Médecine - Timone27, boulevard Jean Moulin13385 Marseille cedex 5 France), 7 - Barcelona Supercomputing Center - Centro Nacional de Supercomputacion (Torre Girona c/ Jordi Girona, 31, 08034 Barcelona Spain)

Correspondance to: judith.lambert@inserm.fr

Finding groups, or clusters, of patients based on longitudinal data is a fundamental task in clinical research. Different approaches have been developed, focusing on raw data, feature or model clustering, but overall, clustering patients based on longitudinal data, in particular when the patients are followed at different ages, remains a major challenge.

The aim of our study is to develop network approaches to obtain patient clusters from longitudinal clinical data. Networks indeed offer the opportunity to integrate heterogeneous data, are robust to handle missing data and could help preserve patient privacy. As a use case, we analysed drug prescriptions contained in the Echantillon Généraliste des Bénéficiaires (EGB) French database. The EGB is a sample of nearly 660,000 beneficiaries of the French health insurance followed over time and containing information about beneficiaries and drug prescriptions. We extracted the prescriptions for antithrombotic agents (ATC class : B01) from 2008 to 2018 and filtered the data to select individuals with at least one prescription for at least two consecutive months. This concerns a total of 51,615 beneficiaries.

We constructed different networks, for every age ranging from 0 to 70 years old. Each network contains nodes corresponding to beneficiaries of a given age and edges representing the similarity between beneficiaries. We calculated this similarity from the number of drugs prescribed for each beneficiary, using a cosine similarity, and considered only values greater than or equal to 0.8. The obtained networks have, on average, 2,800 nodes and 4 million edges.

The detection of patient clusters in each network is performed with the unsupervised Markov cluster algorithm (MCL), which is based on stochastic flow simulation. We then characterized each cluster with its number of beneficiaries, sex-ratio and by identifying the most prescribed drugs. Overall, our network approach allows identifying clusters of beneficiaries clinically relevant, highlighting long-term treatments and underlying pathologies.

We will then construct a longitudinal multi-layer network by connecting the networks constructed at different ages. We will develop a method to identify clusters from this special type of multi-layer networks. We thereby expect to identify clusters of patients from a longitudinal perspective. On a longer term, we expect to associate patient clusters built from longitudinal clinical data and genetic information, to seek for variants involved in longitudinal phenotype.
Same role but different actors: genetic regulation of post-translational modification of two different proteins


1 - Centre for Global Health Research, Usher Institute, University of Edinburgh (Edinburgh United Kingdom), 2 - Genos Glycoscience Research Laboratory (Zagreb Croatia), 3 - MRC Human Genetics Unit, Institute for Genetics and Cancer, University of Edinburgh (Edinburgh United Kingdom), 4 - Laboratory of Glycogenomics, Institute of Cytology and Genetics (Novosibirsk Russia), 5 - Faculty of Pharmacy and Biochemistry, University of Zagreb (Zagreb Croatia)

Correspondance to: arianna.landini@ed.ac.uk

Post-translational modifications (PTMs) are essential mechanisms used by cells to diversify protein functions and dynamically coordinate their signaling networks. While PTMs are known to be involved in regulating almost all cellular events, genetic regulation of PTMs themselves has not been extensively investigated. Protein N-glycosylation, one of the major PTMs, has been linked to the ageing process and a wide variety of diseases, ranging from rheumatoid arthritis, type 2 diabetes, Crohn's and Parkinson's disease to cancer. Nevertheless, genetic regulation of N-glycosylation is yet not fully understood. To answer the question whether the same PTM of two proteins is performed by the same enzymes and regulated by the same genes, we performed genome-wide association meta-analysis of glycosylation of two proteins - transferrin (35 N-glycan traits, N=1890) and immunoglobulin G (IgG) (24 N-glycan traits, N=2020). In the first ever GWAS of transferrin N-glycosylation we identified 10 significantly associated loci (P < 1.43 ×10^-9), while 13 loci were associated with IgG N-glycosylation (P < 2.08×10^-9). Three of the genomic regions associated with transferrin glycosylation (TF, FOXI1 and MSR1) were never previously associated with the N-glycome of any protein, while seven of them were previously associated with either IgG or total plasma proteins glycosylation. Overall, we found 7 genomic regions containing glycosyltransferases, enzymes directly involved in the biochemical pathway of glycosylation, and 2 of these (FUT8 and FUT6) were associated with both IgG and transferrin glycosylation. Using colocalisation methods, we showed that while these genes encode fucosyltransferases altering both proteins, there is strong support for a different causal variant in each gene influencing the glycosylation of each protein. Moreover, we also suggest that core fucosylation of the two proteins is regulated by different transcription factors, IKZF1 for IgG and HNF1A for transferrin. By uncovering the genes responsible for transferrin N-glycosylation and comparing them with those underlying IgG glycosylation, our results are beginning to unravel the genetic regulation of glycosylation, one of the most common PTMs. We show that both unique and shared genes are involved in glycosylation of two proteins and suggest not only that different transcription factors regulate the same enzyme in different tissues, but also that different underlying causal variants in the same gene are regulating glycosylation in a substrate-specific manner.
BactGWAS: a new R package to perform GWAS for bacteria

Laporte Fabien (1) (2), Maury Mylène (3), Patin Etienne (4), Hafner Lukas (3), Leclercq Alexandre (3), Quintana Murci Lluis (4), Chikhi Rayan (2), Lecuit Marc (3), Aschard Hugues (2)

1 - Mer, Molécules, Santé, Université Catholique de l’Ouest - Angers (3 Place André Leroy, 49000 Angers France), 2 - Department of Computational Biology, Institut Pasteur (25-28 rue du Docteur Roux, 75015 Paris France), 3 - Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria, Institut Pasteur, Inserm U1117 (25-28 rue du Docteur Roux, 75015 Paris France), 4 - Human Evolutionary Genetics, Institut Pasteur, UMR 2000, CNRS (25-28 rue du Docteur Roux, 75015 Paris France)

Correspondance to: fabien.laporte@outlook.com

Introduction: Genome-wide Association Studies (GWAS) have been central to identify genetic variations associated with complex human phenotypes. There is now tremendous interest for implementing GWAS-like approaches to genomes of pathogenic bacteria in order to advance our understanding of infectious disease risk and underlying bacterial virulence factors. However, bacterial genomes harbour complex structure making such analyses extremely challenging.

Materials and Methods: We used real genetic data from 3718 strains of Listeria monocytogenes collected in the context of Listeria surveillance in France. Their sequences were represented as a De Bruijn graph to obtain an unitig genetic matrix. We simulated phenotypes with and without unitig effects to compare our approach with classical bacterial GWAS softwares. We also compared alternative ad hoc algorithms for deriving pairwise genetic relatedness between strains and assessed a range of linear mixed models (LMM) with various random effects. The LMM inference was performed using Restricted Maximum Likelihood estimator.

Results & conclusion: We identified multiple critical issues in existing approaches, resulting in invalid GWAS results in our simulation. We developed solutions for each of those issues, implementing an adaptive inference of relatedness between strains and a stepwise LMM model selection. The new approach displays a uniform p-values distribution under the null scenario, and unbiased variance estimates and high relative power under the alternative one. The package will be available soon.
SNPs tagging structural variations help to reveal widespread role of CNVs in complex traits and gene expression

Lepamets Maarja (1), Võsa Urmo (1), Alasoo Kaur (2), Mägi Reedik (1)

1 - Estonian Genome Centre, Institute of Genomics, University of Tartu (Riia 23b, Tartu Estonia), 2 - Institute of Computer Science, University of Tartu (Narva str 18, Tartu Estonia)

Correspondance to: maarja.lepamets@ut.ee

Introduction: Genome-wide association studies (GWAS) have been successful at identifying links between single nucleotide polymorphisms (SNPs) and complex traits. Pinpointing true causal markers, however, has been challenging, and can be even more so in loci where SNP effects reflect the underlying structural variation. In order to recognise such loci, we identify copy number variable regions (CNVr) that are tagged by flanking SNPs and evaluate the abundance of those tag-SNPs among known GWAS and expression quantitative trait loci (eQTLs).

Methods and Results: We discovered that approximately 18% (N=1,991) of 35,422 non-overlapping CNVr detected from ~2,300 Estonian whole-genome sequencing samples had tag-SNPs (Pearson’s R2>0.8) in their 500kb radius. We compared them to known associations from GWAS Catalog and found that for 190 CNVr the top tag-SNP was in high LD (r2>0.8) with a GWAS hit. The most frequent associations were with BMI (17 unique CNVr) and different cognitive ability traits (in total 23 CNVr). We also detected high level of pleiotropy (>10 unique GWAS traits) for CNVr at 6p21.32 and 17q21.31. Furthermore, we analysed the overlap of tag-SNPs and 95% credible sets (CS) for eQTLs from eQTL Catalogue. We discovered 618 CNVr-gene pairs corresponding to 338 unique CNVr for which the tag-SNPs covered over 90% of CS in at least one dataset/tissue. We found that 72 (21%) of those CNVr overlapped with an exon of the corresponding gene.

Conclusions: Altogether, we found that 453 (23%) CNVr which were tagged by SNPs showed evidence of being an underlying variant behind either GWAS or eQTL SNP associations. Such information is essential when designing functional follow-up experiments to identify causal variants.
The singleton missense variant found in a rare adult-onset cancer- SDHB/SDHD as an exemplar for quantifying phenotypic specificity.

Loveday Chey (1), Garrett Alice (1), King Laura (1), Butler Samantha (2), Robinson Rachel (3), Maher Eamonn (4), Buffet Alexandre (5), Burnichon Nelly (5), Gimenez-Roqueplo Anne-Paule (5), Choi Subin (1), Durkie Miranda (6), Burghel George (7), Drummond James (8), Berry Ian (2), Wallace Andrew (7), Woodward Emma (7), Izatt Loise (9), Tischkowitz Marc (4), Whiffin Nicky (10), Ware James (11), Hanson Helen (1), Turnbull Clare (1)

1 - The institute of cancer research [London] (United Kingdom), 2 - Birmingham Women’s and Children’s NHS Foundation Trust (Birmingham United Kingdom), 3 - Leeds Teaching Hospitals NHS Foundation Trust (Leeds United Kingdom), 4 - University of Cambridge (United Kingdom), 5 - Université de Paris (Paris France), 6 - Sheffield Children’s NHS Foundation Trust (Sheffield United Kingdom), 7 - Manchester University NHS Foundation Trust (Manchester United Kingdom), 8 - Addenbrooke’s Cambridge University Hospitals NHS Foundation Trust (Cambridge United Kingdom), 9 - Guy’s and St Thomas’ NHS Foundation Trust (London United Kingdom), 10 - Wellcome Centre for Human Genetics, University of Oxford (Oxford United Kingdom), 11 - National Heart and Lung Institute and MRC London Institute of Medical Sciences (London United Kingdom)

Correspondence to: chey.loveday@icr.ac.uk

Background: We frequently encounter for the first time a very rare missense variant (VRMV) in an affected proband with the “correct” phenotype in what is generally deemed a “quiet” cancer susceptibility gene. Prior to adoption of the 2015 ACMG framework, such an observation was often in itself taken as sufficient to classify the variant as pathogenic. From the 2015 ACMG framework, PP2, PM1, PP4 and PS4 have been variously used in relation to this scenario. SDHB and SDHD are prime examples. Phaeochromocytoma/paragangliomas (PCC/PGL) are very rare tumours, of which 10-20% are associated with germline mutation of SDHB/SDHD. What evidence weight should be given to a single observation of a SDHB/SDHD variant in an affected proband?

Methods: We amassed variant data from clinical testing spanning >10 years from two UK clinical laboratories and one French clinical series. Combined with the Cancer Genome Atlas (TCGA) these comprised 4993 PCC/PGL. We combined our case datasets with gnomAD and 1000Genomes population datasets to calculate the frequency for VRMVs in SDHB and SDHD in the PCC/PGL case series compared to control series. We then used a clustering algorithm to examine for regional case-control enrichment of VRMVs. We then performed case-only comparison for clinical and molecular sub-phenotypes.

Results: We derived a likelihood ratios towards pathogenicity for observation of a VRMV in PCC/PGL for SDHB of LR=76.0 (54.5-106.0) and for SDHD of LR= 12.9 (7.3-22.0). From the clustering analysis we could up-/downgrade for regional enrichment of VRMVs inside/outside the SDHB cluster region (aa 46-255) to LR= 93.3 (63.2-137.8) and 29.9 (15.1-59.3) respectively (p=0.0004), and those inside/outside the SDHD cluster region (aa 70-114) to LR=29.3 (12.2-70.1) and 5.3 (2.2-13.0) respectively (p=0.0003). Both malignant disease (SP-LR=4.49 (2.96-6.80)) and multiple tumours (SP-LR=3.37 (1.97-5.76)) were predictive for SDHB mutational status compared to SDH wild-type mutational status. Both head-and-neck disease (SP-LR=4.04 (3.38-4.82)) and multiple tumours (SP-LR= 11.63 (9.74-13.90)) were predictive for SDHD mutational status compared to SDH wild-type mutational status. Loss of staining on SDHB immunohistochemistry conferred for SDHB LR=17.9 (14.7-21.8) towards pathogenicity and for SDHD LR= 18.1 (16.6-19.8).

Conclusion: Our results illustrate that even with fairly modest overall attribution of the gene(s) to the phenotype, for very rare phenotypes even a single observation of a rare variant may warrant significant evidence weightings towards pathogenicity. We propose that adoption for rare disease of the methodologies illustrated would improve consistency and accuracy of quantitative estimation of PP2, PM1 and PP4.
Identifying causal genetic variants for rare diseases is crucial to provide better diagnosis, treatments and genetic counseling. Sequencing data from unrelated case and control samples are usually used to identify genes with a higher burden of rare variants in cases than in controls. However, this strategy lacks of power when dealing with highly heterogeneous rare diseases or oligogenic inheritance. In this work, we explore how knowledge about gene functional networks can help in identifying digenism (subset of oligogenism) in a case example of a rare disease: the Cerebral Small Vessel Disease (CSVD). Despite 11 causal genes identified in CSVD, 4 of them in the core matrisome associated gene set, 80% of familial cases lack of molecular diagnosis.

Here, we used case-control whole-exome sequencing data in 207 CSVD cases and 566 controls from the general population. We investigated pairs of genes simultaneously carrying functionally high impact variants in a single individual. Using external information about functional networks (protein-protein interaction, Reactome pathways, molecular complexes), we determined how many of these pairs of genes were functionally connected. Focusing on described sets of genes from the GSEA project (http://www.gsea-msigdb.org), we compared the proportion of connected pairs in cases and controls.

Our results show a significantly higher proportion of connected pairs of highly-impacted genes in CSVD cases than in controls in the core matrisome gene set (p=0.0034), which encourage further exploration of digenism. We also show that statistical power to identify causal pair of genes will be even more challenging than to identify causal genes, with a higher number of pair of genes to test and smaller numbers of events for each test.

In conclusion, our results support the hypothesis of digenism involved in the development of CSVD, and probably of other underdiagnosed rare diseases, and highlight the need of specific methodology to overcome the statistical challenges of such research.
Efficient multiple changepoint procedure for the detection of local signatures of selection

Mary-Huard Tristan (1) (2), Rigaill Guillem

1 - Institut national de la recherche agronomique [Paris-siège] (147 rue de l’Université 75338 Paris Cedex 07 France), 2 - UMR INRAE GQE Le Moulon / UMR INRAE MIA Paris (Génétique Quantitative et Évolution - Le Moulon Université Paris-Saclay, INRAE, CNRS, AgroParisTech Ferme du Moulon 91190 Gif-sur-Yvette, France France)

Correspondance to: maryhuar@agroparistech.fr

The detection of genomic region involved in local adaptation is, arguably, one of the main challenges of modern population genetic. Many statistical procedures have been developed to detect such regions, most of them being slow or computationally demanding. This hampers their application to high-throughput SNP genotypic datasets. Most procedures rely on a divergence score based on population allelic frequencies measured at each marker (e.g. Fst). This score is then analyzed either (i) through multivariate procedures that perform a joint analysis of the score but do not account for the spatial arrangement of the markers (e.g. Bayesian Factor models) or (ii) through genome scan procedures that account for the spatial distribution of the signal (e.g. BayeScan). Following this last line of research we propose a new genome scan approach based on the robust Fst estimator recently proposed by Bhatia. We demonstrate that the problem of segmenting the genome into Fst-homogeneous regions can be recast as a weighted changepoints detection problem. When comparing two populations we propose to use the quasi-linear, yet exact fpop-algorithm, to recover the penalized maximum likelihood estimator. This approach drastically reduces the computational burden while retaining strong statistical guarantees. The procedure is illustrated on real Human datasets (HGDP and/or 1 KG datasets).
PathWAS analysis sheds a new light on the biology of complex traits

May-Wilson Sebastian (1), Macdonald-Dunlop Erin (1), Consortium Scallop (2), Wilson James (1), Pirastu Nicola (1)

1 - University of Edinburgh (Usher Institute for Population Health Sciences and Informatics University of Edinburgh Teviot Place Edinburgh EH8 9AG United Kingdom), 2 - Karolinska Institute, Stockholm (Sweden)

Correspondance to: s.may-wilson@ed.ac.uk

Rationale: With the aim of understanding complex traits and multifactorial disease, there has been an increasing focus on studying omics alongside genetic data from GWAS. Studies such as TWAS have incorporated gene expression with GWAS in order to improve the power of discovery for causal associations with complex traits. These studies could, however, be potentially limited by examining the effects of individual genes acting in isolation and not in the context of broader biological networks. Incorporating multiple genes, grouped by pathways, has the potential to further increase power while also improving our understanding of the underlying biology.

Method: We selected genes from known biological pathways, obtained from online databases such as KEGG, and then created polygenic risk scores (PRS) for genes from these pathways using available QTL data (eQTLs from eQTLgen and pQTLs from INTERVAL). These PRS are created using betas weighted by LD structure using the package LDpred2. The relative contribution of each gene on overall pathway functionality is estimated by fitting a multivariable Mendelian randomisation (MR) using the *QTL SNPs as exposures against a measured protein "end-point" from the SCALLOP consortium. The PRS and MR results are then combined to create an overall pathway PRS, validated in an independent sample. The significant pathway scores were then tested, using PheWAS, against multiple complex traits in UK Biobank.

Results: Our method successfully predicted the end-point protein level in 8 pathways. These pathways are primarily immune response pathways, such as NOD-like receptor signalling and Toll-like receptor signalling. From these results, PheWAS identified numerous associations between these pathways and traits in UK Biobank such as lymphocyte and leukocyte count but also height, weight and lung-function traits.

Conclusion: Pathway scoring offers the prospect of more powerful and holistic analysis of GWAS results, with the potential to discover relevant causal pathways for complex traits. By incorporating additional QTL and proteomics datasets, along with improvements made in analysis of pathways, in future it may be possible to further improve the prediction of pathway functionality for any individual, thus allowing potential examination of pathway relevance in many complex traits and diseases.
Evaluating the interpretability of SNP effect size classes in Bayesian genomic prediction models

Mollandin Fanny (1), Rau Andrea (1), Croiseau Pascal (2)

1 - Génétique Animale et Biologie Intégrative (Domaine de Vilvert F-78350 Jouy-en-Josas France), 2 - Génétique Animale et Biologie Intégrative (Domaine de Vilvert F-78252 Jouy-en-Josas France)

Correspondence to: fanny.mollandin@inrae.fr

Genomic prediction models are widely used as an evaluation tool for genomic selection in plant and animal breeding, and for the calculation of polygenic risk scores for human diseases. Among other approaches, non-linear Bayesian models represent an attractive approach to perform genomic prediction, due in part to their flexibility and ability to perform variable selection. In recent years, a suite of models in the so-called Bayesian alphabet have been proposed to find a compromise between reality, which may correspond to an omnigenic model for complex traits, and computational convenience. In particular, the BayesR model (Erbe et al., 2012) strikes this balance by modeling single nucleotide polymorphisms (SNPs) as a mixture of markers with null, small, medium, or large variance. Although these approaches have been shown to achieve improved prediction accuracy in a variety of scenarios, there is still a need to evaluate the extent to which the assignment of SNPs to specific effect size classes (small, medium, large) reflects the true underlying genetic architecture and is meaningful for downstream SNP selection. Based on a set of real genotypes, we generated simulated data under a wide variety of genetic architectures and phenotype heritabilities, and we evaluated the impact of excluding or including causal markers among the genotypes. We define several statistical criteria for QTL mapping, including several based on sliding windows to account for linkage disequilibrium, and we evaluated their ability to accurately prioritize known causal markers in various scenarios. Finally, we provide some insight into how the use of additional prior biological information can contribute to the interpretability of the SNP classes identified by BayesR. In particular, BayesRC (MacLeod et al., 2016), an extension of BayesR, categorizes SNPs according to their functional annotations, each of which is independently modeled using BayesR to allow for a varying enrichment of QTLs. The use and extension of the BayesRC model in conjunction with appropriate QTL mapping criteria is thus a promising approach for incorporating and better exploiting heterogeneous biological information into powerful and interpretable genomic prediction models.
Correction for sample overlap, Winner’s curse and weak instruments bias in two-sample Mendelian Randomization

Mounier Ninon (1) (2), Kutalik Zoltán (3) (1) (2)

1 - Swiss Institute of Bioinformatics (Lausanne Switzerland), 2 - University Center for Primary Care and Public Health (Lausanne Switzerland), 3 - Department of Computational Biology (University of Lausanne Switzerland)

Correspondance to: ninon.mounier@unil.ch

Introduction: Inverse-variance weighting (IVW) two-sample Mendelian Randomization (MR) is the most widely used method to estimate the causal effect of an exposure on an outcome. However, the resulting causal effect estimates may suffer from different biases due to sample overlap, Winner’s curse and weak instruments.

Methods: Assuming a spike-and-slab genomic architecture, we analytically derived the bias of such estimate, which can be quantified using only summary statistics. The bias depends on the true causal effect size, the strength of confounding, the sample sizes of the exposure and outcome GWASs, the sample overlap (that does not need to be known and can be estimated from cross-trait LD score regression), the threshold used to select instruments and the genetic architecture of the exposure (polygenicity and per-SNP heritability). Hence, we propose a correction of the IVW-MR estimate and compared it against its uncorrected counterpart under a wide range of simulations settings. Finally, we used UKBB data to compute summary statistics for body mass index (BMI) and systolic blood pressure (SBP) from samples with variable overlap and used these for IVW-MR.

Results: Using simulated data, we observed that when the confounder and the causal effect are acting in the same direction, IVW-MR effects are overestimated for fully-overlapping samples and underestimated for non-overlapping samples. When they are acting in opposite directions, observed effects are underestimated for all overlaps because the three sources of biases are towards the null. In all the explored scenarios, our correction reduced bias up to 30 folds. Using summary statistics for real data, our method revealed that IVW-MR causal effect of BMI on SBP from non-overlapping samples was significantly underestimated (by 30%) because of weak instruments bias and Winner’s curse and that IVW-MR causal effect of SBP on BMI was significantly overestimated when using fully-overlapping samples (by 20%).

Conclusions: We developed a method to correct causal effect estimates for sample overlap, weak instrument bias and Winner’s curse simultaneously using only summary statistics.
Retro-prospective modelling of recurrent events

Munsch Gaëlle (1), Ibrahim-Kosta Manal (2) (3), Gouimidi Louisa (3), Deleuze Jean-François (4) (5), Morange Pierre-Emmanuel (2) (3), Trégouët David-Alexandre (1), Jacqmin-Gadda Hélène (1)

1 - Bordeaux population health (Université de Bordeaux - Case 11 - 146 rue Léo Saignat 33076 Bordeaux cedex France), 2 - Centre de recherche en CardioVasculaire et Nutrition (Faculté de Médecine de la Timone 27 boulevard Jean Moulin 13385 MARSEILLE CEDEX 5 France), 3 - Laboratoire d'hématologie (278 rue Saint Pierre, Marseille 13385 France), 4 - Centre d'Étude du Polymorphisme Humain (27, rue Juliette Dodu, 75010 Paris France), 5 - Centre National de Recherche en Génomique Humaine (91057, Evry France)

Correspondance to: gaelle.munsch@u-bordeaux.fr

Risk factors analysis of recurrent events is a research topic frequently encountered in biomedical domains. When the event of interest occurs before the inclusion in the study, the information is generally not used in statistical analyses aimed at identifying risk factors for a recurrent event. Indeed, to avoid selection bias, only post-inclusion information is used, although the integration of pre-inclusion data could increase the statistical power when the explanatory variables are non-time-dependent. We here propose a weighted survival model allowing to study both prospective and retrospective events. This work is motivated by the analysis of venous thromboembolism (VTE) recurrence risk factors in the MARTHA cohort.

The MARTHA study was initially designed to investigate VTE risk factors. From 1994 to 2010, 1,459 VTE patients were recruited at La Timone Hospital in Marseille. Among them, 397 had already suffered more than one VTE event. Between 2013 and 2018, 773 patients were contacted to gather information on post-inclusion VTE, which led to the identification of 151 additional recurrences. Our total sample then consisted of 1,459 individuals including 548 recurrences. In this work, we aim at assessing the impact of non-time-dependent variables including gender, age at first VTE recurrence, and genetically determined ABO blood groups (O, A, B, AB) on the risk of VTE recurrence. The association was modelled through a weighted Cox model integrating recurrent events pre and post-inclusion. Weights were calculated from the inverse of the survival probability up to the date of recurrence data collection using a delayed-entry Cox model applied to the available death event.

In the prospective analysis of 773 subjects including 151 recurrences, male gender (HR=1.6±0.18, p=8.10-3) was significantly associated with an increased risk of recurrence, but not age (HR=1.4±0.28, p=0.26) nor ABO blood groups despite a trend for increased risk of recurrence in A group (HR=1.4±0.21, p=0.09) compared to O group. The retro-prospective analysis allowed to refine these observations. The regression coefficients obtained were consistent and the standard deviations almost divided by 2: HR=1.7±0.09 (p=7.10-8) for male gender, HR=1.3±0.10 (p=0.01) for age, and HR=1.3±0.11 (p=0.02) for A blood group.

The proposed methodology enables to optimize the analysis of non-time-dependent risk factors of a recurrent event by integrating both pre- and post-inclusion data. This modelling has an immediate field of application in the context of genetic association studies on the risk of a disease's recurrence where studied DNA polymorphisms are fixed at birth.
Converting single nucleotide variants between genome builds: from cautionary tale to solution

Ormond Cathal (1), Ryan Niamh M. (1), Corvin Aiden (1), Heron Elizabeth A. (1)

1 - Neuropsychiatric Genetics Research Group (Dept. of Psychiatry, Trinity College Dublin Ireland)
Correspondance to: caormond@tcd.ie

Introduction

Next generation sequencing studies are dependent on a high-quality reference genome for single nucleotide variant (SNV) calling. Although the two most recent builds of the human genome are widely used (GRCh37 and GRCh38), position information is typically not directly comparable between them. Re-alignment gives the most accurate position information, but since this procedure is computationally expensive, tools such as liftOver and CrossMap are used to convert data from one build to another. However, the positions of converted SNVs do not always match SNVs derived from aligned data and in some instances, SNVs are known to change chromosome when converted. Occasionally, multiple variants map to the same position on the target build, so there is not always a reversable mapping between positions. This inconsistency can be a significant problem for bioinformatics analyses when compiling sequencing resources or comparing results across studies. For example, the SNV chr22:c.15690247T>A on GRCh38 is in the top percentile of variants ranked by CADD. When converted to GRCh37, this variant maps to chr14:c.19553586T>A but is now in the bottom percentile.

Methods

Here we describe a novel algorithm to identify positions that are unstable when converting between the two most commonly used human genome reference builds. Application of this algorithm resulted in 11.3Mb of unstable positions with GRCh37 as the source and 20Mb with GRCh38 as the source. These positions were largely observed to overlap segmental duplications and contigs which are known to differ between the builds. The unstable positions are detected independent of the conversion tools and are determined by the chain files.

Results

As a proof of principle, we examined SNVs derived from whole genome sequencing of two publicly available genomes aligned to both builds. Pre-excluding SNVs at these unstable positions prior to conversion results in SNVs that are stable. This simple procedure gives the same final list of stable SNVs as applying the algorithm and subsequently removing variants at unstable positions. We compared the converted variants to data aligned to the same build and noted discordance rates of 1-3%. Variants at unstable positions, although fewer in number, had higher discordance rates than those at stable positions. This work highlights the care that must be taken when converting SNVs between genome builds and provides a simple method for ensuring higher confidence converted data. Source code for the algorithm as well as BED files for the unstable positions are available on github (http://github.com/cathaloruaidh/genomeBuildConversion).
SNP Fine-Mapping and Gene Prioritisation in the PGC3-Schizophrenia Study


1 - MRC Centre for Neuropsychiatric Genetics and Genomics (Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University. Hadyn Ellis Building, Maindy Road, CF244HQ, Cardiff United Kingdom)

Correspondance to: pardinas@cardiff.ac.uk

Genome-wide association studies (GWAS) are a powerful framework for resolving the genetic basis of human complex traits. However, the linkage disequilibrium (LD) structure of the human genome does not allow for straightforwardly inferring whether the source of a given association signal is a specific SNP or gene within its boundaries. Statistical fine-mapping techniques, which leverage local LD to establish likely causal (“credible”) sources of associations, are thus necessary precursors to statistical or functional studies of associated regions. The latest GWAS meta-analysis from the Schizophrenia Working Group of the Psychiatric Genomics Consortium ("PGC3") has established 270 genome-wide significant loci, which collectively span more than 1,000 genes. Narrowing down this list could provide a clearer picture of the biological processes that confer risk of this disorder, or even pinpoint potential actionable targets.

We used FINEMAP to infer sets of credible SNPs from the PGC3 associations, reflecting both statistical uncertainty and the underlying LD structure of the data. On the resulting fine-mapped SNPs we then applied a set of thresholds based on functional, eQTL and Hi-C annotations, to generate a list of "prioritised" genes. Enrichment analyses were then used to test the association between these prioritised genes and other gene sets found to be related to schizophrenia in common and rare contexts, such as Loss-of-Function (LoF) intolerant genes and genes associated to neurodevelopmental disorders.

Our prioritisation criteria highlighted 65 protein-coding genes, and we also defined a wider set of 477 protein-coding genes by selecting all those tagged by at least one credible SNP. Genes in both sets were found to be twice as likely to be LoF-intolerant than other genes within GWAS loci, after accounting for the gene length. Prioritised genes were also found to be informative about the genetic architecture of this disorder, being between 2-11 times more likely than other genes within GWAS loci to be within gene sets enriched for schizophrenia rare variation in the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) project.

By coupling fine-mapping analyses with gene-based prioritisation criteria based on convergent lines of functional and cellular evidence, we have defined a narrow set of genes from a large amount of potential GWAS-based candidates, which might be of interest for future functional or pharmacological studies. The relationship between these genes and others previously associated to schizophrenia not only supports their identification, but informs on the overlap between the common and rare genetic architectures of this disorder.
Is systolic blood pressure a causal risk factor for intraocular pressure?: a Mendelian randomisation study.

Plotnikov Denis (1), Huang Yu (2), Guggenheim Jeremy A (3)

1 - Kazan State Medical University (420012 49 Butlerova Street, Kazan Russia), 2 - Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People’s Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China (106 Zhongshan Second Road, Guangzhou, Guangdong Province China), 3 - Cardiff University (Cardiff University, Maindy Rd, Cardiff CF24 4HQ United Kingdom)

Correspondence to: plotnikovd@cardiff.ac.uk

Elevated intraocular pressure (IOP) is a common finding in ophthalmic practice and is considered a major risk factor for primary open angle glaucoma – a leading cause of blindness in certain age groups. The association between blood pressure (BP) and IOP has been shown in observational studies; however, the causality of this relationship has not been assessed yet. We thus applied a Mendelian randomisation (MR) analyses (both linear and non-linear) to test the hypothesis of BP being a causal risk factor for IOP elevation. We performed a 1-sample MR analysis in 71,069 UK Biobank individuals using polygenic risk score for BP as a single instrumental variable. The assumption of linearity in the relationship between BP and IOP was tested in a non-linear MR framework in the same sample of 71,069 UK Biobank participants stratified by the residual SBP into 10 quantiles. We performed an inverse variance-weighted (IVW) linear 2-sample MR analysis; and as a sensitivity analyses we carried out the 2-sample MR-Egger, weighted-median MR and MR-PRESSO analyses using GWAS for BP (n=226,997) and GWAS for IOP (n = 71,069) summary statistics data sets.

The observational analysis estimated effect of BP on IOP was 0.28 mmHg [95%CI 0.26 to 0.29 per 10mmHg of SBP; p <2E-16]. 1-sample linear MR yielded a causal effect estimate of 0.07 [95%CI - 0.02 to 0.15 per 10 mmHg SBP; p = 0.14]. The first stage regression F-statistics of 2240.86 rejected a hypothesis a weak instrument bias presence.

IVW MR analysis estimated a causal effect of SBP on IOP as 0.06 [95%CI -0.1 to 0.2 per 10mmHg SBP; p = 0.29]. MR-PRESSO analysis identified 8 SNP as outliers; and the estimated effect corrected for outliers was higher than that from the IVW MR (0.12 mmHg IOP per 10 mmHg SBP; p = 0.01); and was similar to one obtained by the weighted median MR (0.17 mmHg IOP per 10 mmHg SBP; p = 0.05). The results of the non-linear MR analysis did not support the hypothesis of the relationship between BP and IOP being non-linear (I2 = 5.13%; Cochran’s Q = 9.32, p = 0.41).

The results of the current study only tentatively support a causal role of SBP on IOP.
Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome

Porcu Eleonora (1) (2) (3), Sadler Marie C. (2) (3), Lepik Kaido (4) (5), Auwerx Chiara (1) (2) (3), Wood Andrew R. (6), Weihs Antoine (7), Ribeiro Diogo M. (8) (3), Bandinelli Stefania (9), Tanaka Toshiko (10), Nauck Matthias (11) (12), Volker Uwe (12) (13), Delaneau Olivier (8) (3), Metspalu Andres (14), Teumer Alexander (12) (15), Frayling Timothy (16), Santoni Federico A. (17), Reymond Alexandre (1), Kutalik Zoltán (2) (3) (6)

1 - Center for Integrative Genomics (Lausanne Switzerland), 2 - University Center for Primary Care and Public Health (Lausanne Switzerland), 3 - Swiss Institute of Bioinformatics (Lausanne Switzerland), 4 - Institute of Computer Science, University of Tartu (Tartu Estonia), 5 - Estonian Genome Centre, Institute of Genomics, University of Tartu (Tartu Estonia), 6 - Genetics of Complex Traits, College of Medicine and Health, University of Exeter (Exeter United Kingdom), 7 - Department of Psychiatry and Psychotherapy, University Medicine Greifswald (Greifswald Germany), 8 - Department of Computational Biology, University of Lausanne (Lausanne Switzerland), 9 - Local Health Unit Toscana Centro (Florence Italy), 10 - Clinical Res Branch, National Institute of Aging, (Baltimore, MD United States), 11 - Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald (Greifswald Germany), 12 - DZHK (German Centre for Cardiovascular Research), partner site Greifswald (Greifswald Germany), 13 - Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald (Greifswald Germany), 14 - Estonian Biobank, University of Tartu (Tartu Estonia), 15 - Institute for Community Medicine, University Medicine Greifswald (Greifswald Germany), 16 - University of Exeter Medical School, University of Exeter (Exeter United Kingdom), 17 - Endocrine, Diabetes, and Metabolism Service, Lausanne University Hospital (Lausanne Switzerland)

Correspondence to: eleonora.porcu@unil.ch

Comparing transcript levels between cases and controls allows the identification of differentially expressed genes, which may be causes, consequences or mere correlates of the disease under scrutiny. Whereas we have previously integrated summary-level data from GWAS and eQTLs in a Transcriptome-Wide Mendelian Randomization (TWMR) approach to elucidate gene expression to trait causal effects, here we propose a reverse TWMR (revTMWR) to shed light on causal imprints of complex traits on transcript levels.

Combining TWMR and revTWMR results revealed that whole-blood gene expression-trait correlation is mainly driven by causal effect from the phenotype on the expression rather than the reverse. For example, BMI- and triglycerides-gene expression correlation coefficients robustly correlate with trait-to-expression causal effects ($r=0.09, P=1.54\times10^{-39}$ and $r=0.09, P=1.19\times10^{-34}$, respectively), but not detectably with expression-to-trait effects.

Genes implicated by revTWMR confirmed known associations, such as rheumatoid arthritis and Crohn's disease induced changes in expression of TRBVand GBP2, respectively. They also revealed intricate feedback loops controlling the levels of clinical biomarkers. For instance, we observed that high levels of HDL cholesterol lower the expression of genes involved in cholesterol biosynthesis (SQLE, FDFT1) and increase the expression of genes responsible for cholesterol efflux (ABCA1, ABCG1), two key molecular pathways in determining HDL levels.

In conclusion, our method disentangles the gene expression-phenotype relationship and reveals that complex traits have more pronounced impact on gene expression than the reverse, demonstrating that studies comparing the transcriptome of cases and controls are more prone to reveal disease-induced gene expression changes rather than disease causing ones.
Probabilistic inference of genetic architecture, functional enrichment, and improved prediction of complex traits and common disease age-at-onset.

Robinson Matthew (1)

1 - Institute of Science and Technology [Austria] (IST Austria Am Campus 1 3400 Klosterneuburg Austria Austria)
Correspondance to: matthew.robinson@ist.ac.at

Due to the complexity of linkage disequilibrium (LD) and gene regulation, understanding the genetic basis of common complex traits remains a major challenge. We develop a series of Bayesian models implemented in a hybrid-parallel algorithm that scales to whole-genome sequence data on many hundreds of thousands of individuals, taking 22 seconds per iteration to estimate the inclusion probabilities and posterior effect size distributions of 8.4 million markers and 78 SNP-heritability parameters in the UK Biobank. Unlike naive penalized regression or mixed-linear model approaches, our software hydra accurately estimates annotation-specific genetic architecture, determines the underlying joint effect size distribution, provides a probabilistic determination of association within marker groups, and improves genomic prediction, in a single modelling step. For height, body mass index, cardiovascular disease, and type-2 diabetes, we find that on average 24% of all cis and coding regions of each chromosome are associated with each trait, with over 3,100 independent exonic and intronic regions and over 5,400 independent regulatory regions having ≥ 95% probability of contributing ≥ 0.001% to the genetic variance of these four traits. Only ≤10% of the SNP-heritability is attributable to proximal regulatory regions within 10kb upstream of genes, while 12-25% is attributed to coding regions, up to 40% to intronic regions, and 22-28% to distal 10-500kb upstream regions. ≤60% of the variance contributed by these exonic, intronic and distal 10-500kb regions is underlain by many thousands of common variants, each with larger average effect sizes compared to the rest of the genome. We also find evidence for an infinitesimal contribution of many thousands of common genomic regions to variation in the onset of common complex disorders of reproductive timing, high blood pressure, cardiac disease, and type-2 diabetes, and for the genetic basis of age-at-onset reflecting the underlying genetic liability to disease. For all traits, we improve prediction accuracy over both individual-level and summary statistic approaches, within and across biobanks.
Assessing the impact of the non-Gaussian dependence structure in the REML estimation of the multivariate genetic model

Rohmer Tom (1), Ricard Anne (2) (3), David Ingrid (1)

1 - Génétique Physiologie et Systèmes d’Élevage (INRAE Auzeville - 24, chemin de Borde-Rouge - Auzeville Tolosane 31326 Castanet Tolosan France), 2 - Génétique Animale et Biologie Intégrative (Domaine de Vilvert F-78252 Jouy-en-Josas France), 3 - Département Recherche et Innovation (Exmes, France)

Correspondance to: tom.rohmer@inrae.fr

In multiple-trait animal models, variance components are estimated using REstricted Maximum Likelihood (REML) method. Best Linear Unbiased Prediction (BLUP) of breeding values is then performed and used for selection. Such an approach is based on the assumption of normality of each bivariate phenotypes even if, in practice, this hypothesis is not always realistic. Indeed, in a multivariate setting, even if the marginal distributions of each phenotypes are normally distributed, the multi-normality assumption may be violated due to the non-normality for the cross-sectional dependence structure, that is to say the copula of the multivariate distribution fails to be Gaussian.

We assessed by simulation, the impact of a non-Gaussian copula for the residual term of the mixed model on the REML estimations of the genetic parameters (heritability, correlation) and on the BLUP of breeding values. Large populations of 18510 animals over 6 generations were simulated and bivariate phenotypes were generated considering different copulas (symmetric, asymmetric, with tail dependence or not) for the dependence structure between the error components. Results obtained highlighted the robustness of REML estimation of genetic and environmental variance-covariance components and prediction of breeding values in a multivariate mixed model context when the Gaussian hypothesis is not verified.
I'am hiQ – A Novel Accuracy Index for Imputed Genotypes

Rosenberger Albert (1), Viola Tozzi (1), Hung Rayjean J (2), Amos Christopher I (3), Bickebölle Heike (1), Consortium Integral-Ilcco (4)

1 - Institut für Genetische Epidemiologie, Universitätsmedizin Göttingen (Humboldtallee 32 37073 Göttingen Germany), 2 - Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto (600 University Ave, Toronto ON MSG 1X5 Canada), 3 - Baylor College of Medicine (1 Baylor Plaza, Houston, TX 77030, États-Unis United States), 4 - homepage (https://ilcco.iarc.fr/Links/links.php France)

Correspondance to: arosenb@gwdg.de

Background: Imputation of untyped markers is a standard tool in genome-wide association studies to close the gap between directly genotyped and other known DNA variants. However, high accuracy with which genotypes are imputed is fundamental. Several accuracy measures have been proposed and some are implemented in imputation software, unfortunately diversely across platforms (e.g. info implemented in IMPUTE2). We introduce I’am hiQ, an independent pair of accuracy measures that can be applied to dosage files, the output of all imputation software. I’am (imputation accuracy measure) quantifies the average amount of individual-specific versus population-specific genotype information in a linear manner. hiQ (heterogeneity in quantities of dosages) addresses the inter-individual heterogeneity between dosages of a marker across the sample at hand. Results: Applying both measures to a large case-control sample of the International Lung Cancer Consortium (ILCCO), comprising 27,065 individuals, we found meaningful thresholds for I’am and hiQ suitable to classify markers of poor accuracy. We demonstrate how Manhattan-like plots and moving averages of I’am and hiQ can be useful to identify regions enriched with less accurate imputed markers, whereas these regions would be missed when applying the accuracy measure info. Conclusion: We recommend using I’am hiQ additionally to other accuracy scores for variant filtering before stepping into the analysis of imputed GWAS data.
A systematic evaluation of low-coverage whole genome sequencing imputation across human populations

Rubinacci Simone (1), Deleaneau Olivier (1)

1 - Department of Computational Biology, University of Lausanne (Quartier Sorge, 1015 Lausanne Switzerland)
Correspondance to: simone.rubinacci@unil.ch

Recent advances in Low-Coverage Whole Genome Sequencing (LC-WGS) imputation allow the use of reference panels containing thousands of individuals, similarly to SNP array imputation. A key feature of LC-WGS is that it mitigates well-known problems of SNP arrays, such as ascertainment bias and the lack of a substantial proportion of rare variants. Imputation methods for LC-WGS require a matrix of Genotype Likelihoods (GL) as an input and use a reference panel of haplotypes to refine the sparsely mapped reads.

Here, we evaluate the performance of LC-WGS at three distinct levels: (i) genotype calling, (ii) choice of reference panels, and (iii) ancestry.

First, we benchmark seven popular GL calling tools together with a new method, GLIMPSE_call, which we specifically developed for LC-WGS. We use 52 individuals from 26 populations of the 1000 Genomes project for a range of 0.1x–4x coverage. We show that our tool is as accurate as the best methods for SNP calling but runs 10–100 times faster. Additionally, we assess the performance at small indels. We found that only few tools are able to call indels and usually this comes with an increased running time. Similarly to what is shown for SNPs, GLIMPSE_call performs as accurately as the best callers at indels, with a minimal increase of the running time.

Second, we use data generated in the previous step to benchmark all the main publicly available reference panels in the context of LC-WGS imputation. We show the impact of high-coverage reference panels using different metrics, such as the dosage r², genotype concordance and the accuracy at heterozygous sites. As expected, we find that the main differences are in the rare-variant spectrum, where high-coverage reference panels increase the accuracy of at least 10%. Within only high-coverage reference panels, we show that the version of 1000 Genomes Project phased with TOPMed offers the best performances.

Finally, we assess the quality of imputation for 127 world-wide populations from the Simon’s diversity project in order to provide the most complete benchmark of LC-WGS across human diversity. We find differences between populations, but overall our pipeline offers a solid strategy for imputation for each human population.

Our extensive benchmark of imputation across different genotyping tools, reference panels, and human populations highlights the feasibility of LC-WGS independently of ancestry. All techniques used in this study will be available as part of the next release of the GLIMPSE tool suite.
A two-stage approach to identify high-order epistatic interactions - An application to cardiac ageing in Drosophila

Saha Saswati (1), Perrin Laurent (1) (2), Spinelli Lionel (1), Brun Christine (1) (2)

1 - Theories and Approaches of Genomic Complexity (Parc scientifique de Luminy - 163 avenue de Luminy 13288 Marseille cedex 9 France), 2 - Laboratoire Information Génomique et Structurale (CNRS) (Parc Scientifique de Luminy, 163 Avenue de Luminy - Case 934, 13009 Marseille France)

Correspondence to: saswati.saha@inserm.fr

Understanding the relationship between genetic variations and variations in complex and quantitative trait phenotypes in any population remains an ongoing challenge. Genome-wide association studies (GWAS) provide the standard tool for identifying the associations. However, most existing GWAS studies focus on disease risk. Relatively less attention is paid to longitudinal phenotypes like disease-progression (or severity), which have growing importance to understand the dynamics of biological systems. Furthermore, identifying epistatic interactions, the interactions between genetic variants associated with phenotypes, is also vital because a genetic factor primarily functions through a complex mechanism that involves multiple genes. This effect might be missed if the gene is examined in isolation without allowing for its potential interactions with other unknown factors. GWAS techniques for detecting epistasis do not consider longitudinal data, and current methods are still limited to binary epistatic interactions and cannot handle well complicated higher-order interactions. Hence, it is important to develop methods that not only take into account the dynamics of longitudinal data but also offer to reveal the relationships between multiple variants, revealing extensive networks of interactions.

In this regard, we propose a novel two-stage approach combining a linear mixed-effects model (LMM) and a mixed-effect conditional random forest (MERF) model to identify from population longitudinal studies (i) the genetic variants associated with a phenotypic variation, (ii) reveal the epistatic interactions networks of genetic variants associated with the phenotypic variation.

In the first stage, to alleviate the computational burden, LMM using single marker GWAS filters the single nucleotide polymorphism (SNP) markers that have a significant effect on phenotype. Then in the second stage, a random forest (RF) model (MERF) is applied to the SNPs filtered, and the tree structure in RF facilitates us to identify n-way SNP interactions between potential explanatory SNPs that contribute to phenotypic variability.

This method has been developed in the framework of a project aiming to identify the genetic architecture linked to cardiac aging in Drosophila. Coupled with some biological validations, it will help in revealing the gene/protein network(s) of genetic interactions and regulations underlying the phenotypic variation. It will also help in understanding the cellular and molecular pathways involved in complex age-related cardiovascular disorders and the obtained results will be presented. Importantly, though the above method is applied here for cardiac aging in Drosophila, it can be generalized to other longitudinal as well as cross-sectional population studies in any model organism.
Pgainsim: A method to assess the mode of inheritance for quantitative trait loci in genome-wide association studies

Scherer Nora (1), Sekula Peggy (1), Pfaffelhuber Peter (2), Köttgen Anna (1), Schlosser Pascal (1)

1 - Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center – University of Freiburg, Germany (Hugstetter Straße 49, D 79106 Freiburg Germany), 2 - Faculty of Mathematics and Physics, University of Freiburg, Germany (Ernst-Zermelo-Straße 1, D 79104 Freiburg Germany)

Correspondence to: nora.isabella.scherer@uniklinik-freiburg.de

Background: When performing genome-wide association studies (GWAS) conventionally an additive genetic model is used to explore whether a single nucleotide polymorphism (SNP) is associated with a quantitative trait regardless of the actual mode of inheritance (MOI). Recessive and dominant genetic models are able to improve statistical power to identify non-additive variants. Moreover, the actual MOI is of interest for experimental follow-up projects. Here, we extend the concept of the p-gain statistic (Petersen et al 2012, PMID 22672667) to decide whether one of the three models provides significantly more information than the others.

Methods: We define the p-gain statistic of a genetic model by the comparison of the association p-value of the model with the smaller of the two p-values of the other models. Considering the p-gain as a random variable depending on a trait and a SNP in Hardy-Weinberg equilibrium under the null hypothesis of no genetic association we show that the distribution of the p-gain statistic depends only on the allele frequency (AF).

To determine critical values where the opposing modes can be rejected, we developed the R-package pgainsim (https://github.com/genepi-freiburg/pgainsim). First, the p-gain is simulated under the null hypothesis of no genetic association for a user-specified study size and AF. Then the critical values are derived as the observed quantiles of the empirical density of the p-gain. For applications with extensive multiple testing, the R-package provides an extension of the empirical critical values by a log-linear interpolation of the quantiles.

Results: We tested our method in the German Chronic Kidney Disease study with urinary concentrations of 1,462 metabolites with the goal to identify non-additive metabolite QTLs. For each metabolite we conducted a GWAS under the three models and identified 119 independent mQTLs for which pval_add>min(pval_rec,pval_dom)<4.6e-11. For 38 of these, the additive modelling was rejected based on the p-gain statistics after a Bonferroni adjustment for 1 Mio5492 tests. These included the LCT locus with a known dominant MOI, as well as several novel associations. A simulation study for additive and recessive associations with varying effect sizes evaluating false positive and false negative rates of the approach is ongoing.

Conclusion: This new extension of the p-gain statistic allows for differentiating MOIs for QTLs considering their AF and the study sample size, even in a setting with extensive multiple testing.
Penetration estimation of SORL1 loss-of-function variants using a family-based strategy adjusted on APOE genotypes suggests a non-monogenic inheritance

Schramm Catherine (1), Charbonnier Camille (1), Zarea Aline (2), Wallon David (2), Lacour Morgane (2), C N R-M A J Collaborators, Alarcon Flora (3), Campion Dominique (2) (1), Nuel Gregory (4), Nicolas Gaël (2) (1)

1 - Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ (22 boulevard Gambetta, 76183 ROUEN France), 2 - Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Neurology and CNR-MAJ (22 boulevard Gambetta, 76183 ROUEN France), 3 - Mathématiques Appliquées Paris 5 (UFR Mathématiques et Informatique, 45 rue des Saints-Pères 75270 PARIS CEDEX 06 France), 4 - Laboratoire de Probabilités, Statistiques et Modélisations (Campus Jussieu Tour 16-26, 1er étage, Place Jussieu 75005 Paris / Bâtiment Sophie Germain 5ème étage Avenue de France 75013 Paris France)

Correspondence to: cath.schramm@gmail.com

Introduction: There is an increasing interest in rare genetic variants with a moderate to strong effect in the etiology of complex disorders. Estimating the age-related penetrance associated with such variants is essential. However, rarity and co-occurrence with other risk factors make such estimations difficult. In Alzheimer Disease, a common complex disorder, the APOE4 allele is the main risk factor (allele frequency ~14%, odds ratios of 3.4 and 14, for heterozygous and homozygous carriers, respectively). Recently, rare loss-of-function variants (LoF) with a strong effect (odds ratios >7) were reported in the SORL1 gene. We propose a family-based methodology to estimate the penetrance of SORL1 rare LoF variants adjusted for APOE4.

Material and methods: In a survival analysis framework, we modeled an instantaneous risk f(t)=f0(APOE,t)*exp(beta(t)) combining: (i) a baseline, f0(APOE,t), for non-carriers of SORL1 LoF variants, stratified by APOE genotypes derived from a large previously published cohort study and (ii) an additive effect of SORL1 LoF variants (beta(t)) estimated from our family cohort. We supposed f0(APOE,t) to be piecewise constant and deduced its associated parameters with their variability from published data. We estimated beta(t), also piecewise constant, from our familial cohort including 42 pedigrees with probands carrying a rare (frequency <1%) LoF (protein-truncating or a missense variant with in vitro LoF evidence) SORL1 variant and with disease onset <75 years. Data include phenotypic information and genotypes of about 459 and 52 relatives, respectively. To accommodate for missing genotypes, we embedded our model into an Expectation-Maximisation (EM) algorithm. This algorithm alternates until convergence between replacing unknown genotypes by their posterior distribution based on current age-related penetrance (E-step) and updating the age-related penetrance for each genotype based on observations and last computed posterior genotype distributions (M-step). 95% confidence intervals were computed by bootstrap accounting for both f0(APOE,t) and beta(t) variability. To correct for ascertainment bias, proband phenotypes were omitted in penetrance estimation. The robustness of this method was assessed in a simulation study.

Results: We provide penetrance estimation curves associated with SORL1 LoF variants at the digenic level. By age 75, we estimate SORL1 LoF variants to reach 100% penetrance only among homozygous APOE4 carriers, 89%[76%-96%] among APOE4 heterozygous carriers and 57%[43%-72%] among APOE4 non carriers.

Conclusion: This method allows for computing penetrance of rare variants at the digenic level by taking into account robust estimation of baseline.
Advanced Bayesian meta-analysis methods for investigating pleiotropy effect

Baghfalaki Taban (1), Sugier Pierre-Emmanuel (2) (3), Truong Thérèse (3), Pettitt Antony (4) (5), Mengersen Kerry (4) (5), Liquet Benoit (2) (6)

1 - Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University (Tehran Iran), 2 - Laboratoire de Mathématiques et de leurs Applications [Pau] (Bâtiment IPRA - Université de Pau et des Pays de l'Adour Avenue de l'Université - BP 1155 64013 PAU CEDEX France), 3 - Centre de recherche en épidémiologie et santé des populations (16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France France), 4 - ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (Brisbane, Queensland Australia), 5 - School of Mathematical Sciences, Queensland University of Technology (Brisbane, Queensland Australia), 6 - Department of Mathematics and Statistics, Macquarie University (Sydney, New South Wales Australia)

Correspondance to: pe.sugier@univ-pau.fr

Genome-wide association studies (GWAS) focus on testing association between millions of genetic markers (or single nucleotide polymorphisms, SNPs) and a phenotype in an agnostic way, in a single-SNP approach. One major finding from GWAS era is that pleiotropy – that occurs when one gene influence two or more unrelated traits - is a widespread phenomenon in human complex traits. Several methods were proposed to combine results across studies of different phenotypes in order to improve the power of detecting pleiotropic associations at SNP level. It is well-established that incorporating prior biological knowledge as gene or biological pathways structures to consider complex mechanisms can help to discover additional genetic risk factors. However, only few methods take into account this prior information on biological knowledge in the context of pleiotropy.

We propose different Bayesian meta-analysis approaches named GCPBayes (Group-level Bayesian Meta-Analysis for Studying Cross-Phenotype Genetic Associations) (1) that take into account gene (or pathway) structure by considering multivariate continuous and Dirac spike and slab priors for group selection. Also, Bayesian sparse group selection with hierarchical spike and slab priors enables us to select important variables both at the group level (gene or pathway) and within group (e.g., at the SNP level). GCPBayes is based on Markov chain Monte Carlo (MCMC) through Gibbs sampling for both correlated and uncorrelated cases. These approaches can be applied to multiple phenotypes of different shapes and allow to take into account for heterogeneity in size and direction of effects.

Simulation studies have shown that our methods outperform benchmark approaches such as ASSET and CPBayes, for finding pleiotropic associations both at SNP-level and gene (or pathway) level. The developed statistical approaches has been applied for enriching our insights about the genetic mechanisms of thyroid and breast cancer types in a candidate pathway approach. Thoses analyses allowed us to detect 9 pleiotropic effects at gene-level, and 4 pleiotropic effects at SNP level in xxx selected pathways. Benchmark approaches did not detect any signal.

The Bayesian meta-analysis approaches that we have developed have made it possible to highlight new genes affecting both thyroid and breast cancers, which would not have been demonstrated with the current reference methods. The R-package for GCPBayes approaches is available in https://github.com/tbaghfalaki/GCPBayes.

**COMET: an R package to identify sample cross-contamination in whole genome sequencing studies.**

**Thiéry Alexandre (1), Zeller Tanja (2), Blankenberg Stefan (2), Ziegler Andreas (1) (2) (3)**

1 - Cardio-CARE, Medizincampus Davos (Herman-Burchard-Strasse 1, 7265 Davos Wolfgang Switzerland), 2 - University Heart Center Hamburg, University Medical Center Hamburg-Eppendorf (Martinistraße 52, 20246 Hamburg Germany), 3 - School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal (Pietermaritzburg South Africa)

Correspondance to: alexandre.thiery@cardio-care.ch

Identification of sample cross-contamination is crucial in next generation sequencing (NGS) studies because undetected contamination may lead to bias in association studies. In PCR-free germline multiplexed whole genome sequencing, sample cross-contamination may be investigated by studying the excess of non-matching reads at homozygous sites compared to the expected sequencing error rate. In this presentation, we propose a probabilistic method to infer contaminated samples and their contaminant for low levels of contamination. The distance on the well plate between the contaminant and the contaminated sample may be penalized. The method is implemented in a free of charge R package. We compare it with the three alternative methods: ART-DeCo, VerifyBamID2, and the built-in function in the Illumina’s DRAGEN platform and demonstrate its accuracy on simulated data. We illustrate the method using real data from the pilot phase of a large-scale NGS experiment with 9000 whole genome sequences. In the real data, our method was able to successfully identify cross-contamination. Sample cross-contamination in NGS studies can be identified using a simple-to-use R package.
Large scale genomic data identifies common biological mechanisms of cardiometabolic traits

Timasheva Yanina (1) (2) (3), Zudina Liudmila (4), Balkhiyarova Zhanna (4), Prokopenko Inga (4)

1 - Bashkir State Medical University (Lenina 3 Ufa 450007 Russia), 2 - Section of Genomics of Common Disease Department of Medicine Imperial College London (Hammersmith Hospital Campus Burlington Danes Building Du Cane Road, London, W12 0NN United Kingdom), 3 - Institute of Biochemistry and Genetics of Ufa Scientific Centre (Ufa Russia), 4 - University of Surrey (University of Surrey, Guildford Surrey GU2 7XH United Kingdom)

Correspondence to: ianina_t@mail.ru

Arterial hypertension, coronary artery disease (CAD), type 2 diabetes (T2D), obesity, systemic inflammation often occur together, forming a cardiometabolic continuum. An overlap between the risk factors for these conditions along with their overwhelming comorbidity suggests that they may share common pathophysiological mechanisms. Large-scale genome-wide association studies (GWAS) have revealed hundreds of genetic loci associated with traits used in the definition of arterial hypertension, including systolic, diastolic, and pulse pressure (SBP/DBP/PP) and other phenotypes indicative of metabolic health. Our aim was to dissect shared genetic factors contributing to their comorbidity applying multi-phenotype approach.

Using summary statistics from published GWAS available in public domain, we analysed 3420 single nucleotide variants (SNVs) associated with BP traits, including SBP/DBP/PP, and 60 other cardiometabolic phenotypes (defined as related to cardiovascular and metabolic diseases). We identified 252 unique BP association signals, and found 301 BP SNVs overlapping with 540 SNVs reported for other phenotypes and located within 100kb from one another. Further analysis revealed 22 SNVs associated with two and more cardiometabolic phenotypes. Specifically, rs13107325 in SLC39A8 gene was associated with mean arterial pressure, body mass index (BMI), and high density lipoprotein cholesterol; rs9368222 in CDKAL1 was associated with SBP, fasting glucose (FG), and T2D; rs3184504 in SH2B3 influenced DBP, risk of stroke, CAD, and ICAM1 levels; rs11556924 in ZC3HC1 affected DBP/SBP, and susceptibility to CAD; rs757081 in NUCB2 had an effect on PP/SBP, height and T2D; rs10830963 in MTNR1B altered PP, T2D risk, FG and corrected insulin response; rs653178 in ATXN2 influenced changes in DBP, ICAM1 level, and type 1 diabetes; and rs12454712 in BCL2 was related to higher SBP, waist-to-hip ratio adjusted for BMI, and T2D susceptibility.

Overall, we highlighted pathways associated with cardiometabolic traits, including circadian rhythm regulation (MTNR1B), pancreatic islet function (CDKAL1), appetite control and nutritional regulation (NUCB2 and ATXN2), immune and inflammatory signaling (SH2B3 and SLC39A8), cell cycle regulation and apoptosis (ZC3HC1 and BCL2), thus suggesting shared biological mechanisms between hypertension endophenotypes and cardiometabolic traits.

Polygenic patterns of susceptibility to multiple sclerosis

Timasheva Yanina (1) (2) (3), Nasibullin Timur (3), Erdman Vera (3), Tuktaraova Ilsiar (3), Galiullin Timur (1), Zaplakhova Oksana (1), Bakhtiiaroa Klara (1)

1 - Bashkir State Medical University (Lenina 3 Ufa 450007 Russia), 2 - Section of Genomics of Common Disease Department of Medicine Imperial College London (Hammersmith Hospital Campus Burlington Danes Building Du Cane Road, London, W12 0NN United Kingdom), 3 - Institute of Biochemistry and Genetics of Ufa Scientific Centre (Ufa Russia)

Correspondance to: ianina_t@mail.ru

Multiple sclerosis (MS) is a chronic demyelinating disease of the central nervous system resulting from the combined influence of genetic and environmental factors. Genome-wide association studies have identified more than 400 genetic variants associated with MS. Our aim was to search for complex genetic markers of MS using polygenic approach.

The study group consisted of 644 patients with MS and 1382 representatives of the control group (301 Bashkirs, 774 Russians, 951 Tatars by ethnic origin). The polygenic effect of the studied loci on MS risk was analysed using the APSampler algorithm. A permutation test was used to assess the statistical significance.

In the group of Russian women, the combination most significantly associated with MS was RPS6KB1 rs180515T/T + CD40 rs6074022C/T (OR = 0.27, Pperm = 0.0005); in Russian men, RPS6KB1 rs180515C/C + PVT1 rs759648C (OR = 7.96, Pperm = 0.0008); in women of Tatar ethnic origin - MANBA rs228614A/A + PVT1 rs759648A + CD40 rs6074022C (OR = 0.1, Pperm = 0.001), in Tatar men - RPS6KB1 rs180515C + CLEC16A rs12708716G/G + CD58 rs OR2300747G (OR = 14.88, Pperm = 0.0016), in Bashkir women PVT1 rs759648C + CD58 rs23000747A + CD40 rs6074022T/T (OR = 4.92, Pperm = 0.001), in Bashkir men SOX8 rs2744148A + CD40 rs6074022*T (OR = 0.06, Pperm = 0.002).

Our results suggest that the patterns of genetic predisposition to MS may vary depending on the ethnicity, and the effects of the genetic loci are influenced by the presence of other susceptibility alleles.

The study was supported by the RFBR/ASA grant (19-54-40007) and the mega-grant from the Government of the Russian Federation (2020-220-08-2197).
Genetically independent phenotype analysis identifies LPA and VCAM1 as drug targets for human ageing

Timmers Paul (1) (2), Tiys Evgeny (3) (4), Sakaue Saori (5), Akiyama Masato (6) (7), Kiiskinen Tuomo (5) (8), Zhou Wei (9) (5) (10), Hwang Shih-Jen (11) (12), Yao Chen (11) (12), Deelen Joris (13) (14), Levy Daniel (11) (12), Ganna Andrea (5) (8), Kamatani Yoichiro (6) (15), Okada Yukinori (16), Joshi Peter (2), Wilson James (1) (2), Tsepilov Yakov (3) (17)

1 - MRC Human Genetics Unit, MRC Institute for Genetics and Cancer, University of Edinburgh (Edinburgh United Kingdom), 2 - Centre for Global Health Research, Usher Institute, University of Edinburgh (Edinburgh United Kingdom), 3 - Laboratory of Theoretical and Applied Functional Genomics, Novosibirsk State University (Novosibirsk Russia), 4 - Laboratory of Glycogenomics, Institute of Cytology and Genetics (Novosibirsk Russia), 5 - Program in Medical and Population Genetics, Broad Institute of Harvard and MIT (Cambridge, Massachusetts United States), 6 - Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences (Yokohama Japan), 7 - Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University (Fukuoka Japan), 8 - Institute for Molecular Medicine Finland, University of Helsinki (Helsinki Finland), 9 - Analytic and Translational Genetics Unit, Massachusetts General Hospital (Boston, Massachusetts United States), 10 - Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT (Cambridge, Massachusetts United States), 11 - Framingham Heart Study (Framingham United States), 12 - Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health (Bethesda, Maryland United States), 13 - Max Planck Institute for Biology of Ageing (Cologne Germany), 14 - Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne (Cologne Germany), 15 - Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo (Tokyo Japan), 16 - Department of Statistical Genetics, Osaka University Graduate School of Medicine (Suita Japan), 17 - Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics (Novosibirsk Russia)

Correspondance to: paul.timmers@ed.ac.uk

The length and quality of life is important to us all, yet identification of promising drug targets for human ageing using genetics has had limited success. Here, we combine six large European-ancestry genome-wide association studies (GWAS) of human ageing traits—healthspan, father and mother lifespan, exceptional longevity, frailty index, and self-rated health—in a principal component framework that maximises their shared genetic architecture. The first principal component (GIP1) is more heritable than the original studies and shows strong genetic correlations with length of life as well as multiple indices of mental and physical wellbeing. We identify 27 genomic regions associated with GIP1, and provide additional, independent evidence for an effect on human ageing for loci near HTT and MAML3 using a study of Finnish and Japanese subject survival. Across the genome, GIP1 associations are enriched in genes involved in haem metabolism and pathways related to transcription, neurogenesis, homeostasis, proteolysis, intracellular signalling, immunity, and the muscle system. Finally, using proteome-wide two-sample Mendelian randomisation and colocalisation, we provide robust evidence for a detrimental effect of blood levels of apolipoprotein(a) (LPA) and vascular cell adhesion molecule 1 (VCAM1) on GIP1. Together, our results demonstrate that combining multiple ageing traits using genetic principal components enhances power to detect biological targets for human ageing.
Reconstructing KIR haplotypes taking ambiguous and missing data into account

Van Der Burg Lars (1), De Wreede Liesbeth C. (1) (2), Baldauf Henning (2), Sauter Jürgen (2), Schetelig Johannes (2) (3), Putter Hein (1), Böhringer Stefan (1)

1 - LUMC, Biomedical Data Sciences (Leiden Netherlands), 2 - DKMS (Dresden/Tübingen Germany), 3 - University Hospital Carl Gustav Carus, Department of Internal medicine I (Dresden Germany)

Correspondence to: l.l.j.van_der_burg@lumc.nl

Background

The role of KIR (killer-cell immunoglobulin-like receptors) genes in improving outcomes after allogeneic hematopoietic stem cell transplantation is under debate. It is hard to assess, since the KIR region is genetically complex; exhibiting copy number variations, a huge allelic diversity and alleles often measured with ambiguities. For better modelling the biological impact of KIR genes, haplotypes have to be reconstructed based on these complex data[1].

Objectives

We present a method to reconstruct haplotype content of the highly complex KIR gene cluster with several levels of missingness: allelic ambiguity, missing phase, and parental configuration of duplications or null alleles.

Methods

We develop an EM-algorithm for haplotype frequency estimation that iteratively adds genes to the reconstruction. To account for exponentially growing number of haplotypes, heuristic grouping strategies are employed, leading to a collapsed haplotype. To properly propagate ambiguities through this algorithm, we profile the likelihood for the collapsed haplotype. We evaluate our method in simulations and compare to a naïve method choosing the most likely haplotype. We also analyze a dataset of 3500 individuals genotyped for 7 KIR genes.

Results

Estimated haplotype frequencies with the new method are unbiased in simulations. Real data analysis shows estimates close to those obtained from independent data[1]. Our simulation study shows that the EM-algorithm can borrow information across loci when linkage disequilibrium is high, whereas increased ambiguity levels increase standard errors. In comparison with the naïve imputation method, the EM-algorithm is more efficient.

Conclusions

Our new EM-algorithm is the first to account for the full genetic architecture of the KIR loci. This algorithm can handle several levels of missingness, and allows to control the overall complexity of the reconstruction using collapsing. Combining information from multiple genes improves the accuracy of estimated haplotype frequencies and can help to deal with the ambiguities on the single gene level.

Shared risk alleles with discordant polygenic effects: Disentangling the genetic overlap between ASD and ADHD

Verhoef Ellen (1), Grove J. (2) (3) (4) (5), Shapland C.y. (6) (7), Demontis D. (2) (3) (4), Burgess S. (8) (9), Rai D. (10) (11) (12), Børglum A.d. (2) (3) (4), St Pourcain B. (1) (10) (13)

1 - Max Planck Institute for Psycholinguistics (Wundtlaan 1 6525 XD Nijmegen Netherlands), 2 - The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (Fuglesangs Allé 26, 8210 Aarhus Denmark), 3 - Department of Biomedicine (Human Genetics) and Centre for Integrative Genomics, University of Aarhus (Bartholins Allé 6, building 1245, 8000 Aarhus C Denmark), 4 - Center for Genomics and Personalized Medicine, Aarhus (- Denmark), 5 - Bioinformatics Research Centre, Aarhus University (C.F. Møllers Allé 8, Building 1110, DK-8000 Aarhus C Denmark), 6 - MRC Integrative Epidemiology Unit, University of Bristol (Oakfield House Oakfield Grove Bristol BS8 2BN United Kingdom), 7 - Population Health Sciences, University of Bristol (University of Bristol, Beacon House Queens Road Bristol, BS8 1Q United Kingdom), 8 - MRC Biostatistics Unit, University of Cambridge (Cambridge Biomedical Campus Cambridge, CB2 0SR United Kingdom), 9 - Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge (Cambridge Biomedical Campus, Cambridge, CB2 0SR United Kingdom), 10 - Centre for Academic Mental Health, Bristol Medical School, University of Bristol (University of Bristol Beacon House Queens Road Bristol, BS8 1Q United Kingdom), 11 - NIHR Biomedical Research Centre, University of Bristol (University of Bristol Beacon House Queens Road Bristol, BS8 1Q United Kingdom), 12 - Avon and Wiltshire Mental Health Partnership NHS Trust Bath NHS House, Newbridge Hill, Bath BA1 3QE United Kingdom), 13 - Donders Institute for Brain, Cognition and Behaviour, Radboud University (Heyendaalse weg 135, 6525 AJ Nijmegen Netherlands)

Correspondence to: ellen.verhoef@mpi.nl

Background: Autism Spectrum Disorder (ASD) and Attention-Deficit/Hyperactivity Disorder (ADHD) are genetically complex co-occurring neurodevelopmental disorders. Their genetic architectures reveal striking similarities, but also differences. Despite an overall positive genetic correlation between ASD and ADHD, genetic links with educational attainment (EA) are discordant, with positive genetic associations for ASD and negative genetic associations with ADHD. Here, we investigate how EA-related polygenic variation is encoded within ASD and ADHD genetic architectures using a bidirectional multivariable regression (MVR) framework, a multivariate methodology adopted from a causal modelling approach.

Methods: We carried out MVR analyses using genome-wide summary statistics based on 10,610 to 766,345 individuals. Without making causal inferences, as we allow for biological pleiotropy, we simultaneously estimated polygenic ASD and ADHD associations with EA as aggregate effects of independent genetic variants, controlling for both the position of the studied SNPs and the direction of the genetic effects at the SNP level. MVRs were carried out for ASD- and ADHD-related variant sets that were selected according to guidelines established for polygenic scoring methods (Pthr<0.05 and Pthr<0.0015). Each risk-related marker allele was aligned to an ASD, ADHD and EA SNP estimate and the aggregate association effect across all selected alleles with EA was simultaneously estimated for ASD and ADHD risk.

Results: Both, ASD- and ADHD-related variant sets captured discordant association patterns with EA. For example, ASD-related variants (Pthr<0.0015, N-SNPs=1973) showed a 0.009 increase (SE=0.003) in years of education per log-odds ASD liability. Conditional on these effects, the same alleles captured a 0.029 decrease (SE=0.004) in years of education per log-odds ADHD liability. Findings remained robust, when studying SNPs with the same risk increasing allele for both ASD and ADHD (~80% of the original variant sets) and were replicated using different ASD summary statistics. The largest discordant association effects with EA were estimated for

Conclusion: Discordant EA-related association profiles are encoded by the same alleles at the same genetic markers in ASD and ADHD genetic architectures, consistent with either biological pleiotropy and/or co-localisation of different risk variants. Our findings suggest local negative genetic covariance for EA-related genetic variation that may contribute to the overall positive genome-wide correlation between ASD and ADHD.
Enhanced SpliceMap and RNA-seq from clinically accessible tissues improves outlier prediction for non-accessible tissues

Wagner Nils (1), Çelik Muhammed Hasan (2), Hözlzwimmer Florian (1), Gagneur Julien (1)

1 - Technische Universität München (Boltzmannstraße 3, 85748 Garching bei München Germany), 2 - University of California Irvine (Irvine, CA 92697, United States United States)

Correspondance to: wagnern@in.tum.de

Aberrant splicing is a major cause of genetic disorders. However, the affected tissues are often not easily accessible, preventing direct detection of many relevant aberrant splicing events in clinical diagnostics. Deep learning models like SpliceAI and MMSplice have improved prioritizations of rare variants found by whole genome sequencing (WGS). However, their performance on predicting aberrant splicing in any given tissue has not been assessed so far. Here, we developed the first benchmark dataset for aberrant splicing prediction by applying the aberrant splicing caller FRASER on 6931 RNA-seq samples from 48 GTEx tissues. We constructed a tissue-specific splicing map (SpliceMap) improving annotations of the GENCODE annotation by excluding unexpressed splice sites and quantifying tissue-specific reference levels of alternative splicing. Next, we built new models combining i) state-of-the-art deep learning models together with SpliceMap for WGS-only diagnostics and ii) furthermore including RNA-seq from clinically accessible tissues (CATs) for integrated WGS and RNA-seq based diagnostics. At 20% recall, out-of-the-box application of existing deep learning models, yield a precision less than 2%. However, when combined with SpliceMap, we reached 20% precision. Integrating furthermore RNA-seq of CATs (blood and skin) led to 50% precision. These results replicated in an independent mitochondriopathy cohorts of WGS with matched RNA-seq samples. Altogether, our approach significantly increases performances of predicting aberrant splicing events in clinically non-accessible tissues, reaching a sufficiently high precision to guide non-invasive genetic diagnostics.
A Genome-Wide Association Study of the Longitudinal Course of Executive Functions

Wendel Bernadette (1), Papiol Sergi (2) (3), Andlauer Till F. M. (4), Dannlowski Udo (5), Kircher Tilo (6) (7), Falkai Peter (3), Schulze Thomas G. (2) (3), Bickeböller Heike (1), Heilbronner Urs (2)

1 - Department of Genetic Epidemiology, University Medical Center Göttingen, Georg-August-University Göttingen (Göttingen, 37073 Germany), 2 - Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich (Munich, 80336 Germany), 3 - Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich (Munich, 80336 Germany), 4 - Department of Neurology, University Hospital rechts der Isar, School of Medicine, Technical University of Munich (Munich, 81675 Germany), 5 - Department of Psychiatry, University of Münster (Münster, 48149 Germany), 6 - Department of Psychiatry and Psychotherapy, Philipps-University Marburg (Marburg, 35039 Germany), 7 - Centre for Mind, Brain and Behaviour, University of Marburg (Marburg, 35032 Germany)

Correspondance to: bernadette.wendel@med.uni-goettingen.de

We investigated executive functions (EFs), a group of higher-level cognitive capabilities that control and coordinate mental processes. They rank among the most heritable psychological traits. We focused on the time course of two core EF skills: set-shifting and updating, assessed by two different cognitive tests: The Trail Making Test, part B (TMT-B), and the Verbal Digit Span backwards, respectively. We studied the genetic basis of these time courses by applying data of the transdiagnostic PsyCourse Study which comprise patients of the affective-to-psychotic spectrum (mainly bipolar and schizophrenic patients) and healthy controls. In our analysis, 1338 genotyped individuals were included in which the time course of EFs was assessed at four measurement points, each six months apart. Behaviorally, EFs improved across the different diagnostic groups, when compared to the initial evaluation.

To identify single nucleotide polymorphisms (SNPs) associated with the change over time in EF skills, we performed a genome-wide association study (GWAS). We applied linear mixed models (LMMs) in our GWAS to model the dependence structure of the longitudinal data sample and tested the SNP-by-time interactions for significance, adjusting for age at baseline, gender, diagnosis and study center. In total, nine genome-wide significant SNPs on chromosome 5, all in strong linkage disequilibrium with each other, were identified for the TMT-B. SNP rs150547358 was the most significant variant with a p-value of $7.2\times10^{-10}$ and an effect estimate beta of 1.16 (95% CI: 1.11, 1.22). All of the nine significant SNPs had a positive effect estimate, reflecting an association with a decrease in the set-shifting performance across time. We replicated the association using the data of the FOR2107 consortium and analyzing the SNPs in 1795 individuals with the two available measurement points two years apart. The variant rs150547358 had again the lowest p-value ($p = 0.015$), with a similar albeit smaller effect estimate beta of 0.85 (95% CI: 0.74, 0.97).

We did not obtain significant results regarding the Verbal Digit Span backwards phenotype.

Our study is a first step towards analyzing the genetic basis of longitudinal phenotypes.
Shared Genetic Susceptibility between Type 2 Diabetes and cancer

Zudina Liudmila (1), Balkhiyarova Zhanna (1) (2), Demirkan Ayse (1), Pupko Igors (1), Gunter Mark (3), Kaakinen Marika (1) (2), Prokopenko Inga (1)

1 - Department of Clinical & Experimental Medicine, University of Surrey, Guildford, United Kingdom (University of Surrey, Guildford GU2 7XH United Kingdom), 2 - Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom (South Kensington Campus London SW7 2AZ United Kingdom), 3 - Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France (150 Cours Albert Thomas 69372 Lyon CEDEX 08, France France)

Correspondance to: l.zudina@surrey.ac.uk

Type 2 Diabetes (T2D) is a risk factor for several cancers, including postmenopausal breast, colorectal and pancreatic. Epidemiological studies suggest a lower risk of prostate cancer in T2D patients. We dissected the genetic factors shared between T2D and the four aforementioned cancers by using the polygenic risk score (PRS) approach and by partitioning the associated genetic variants through hierarchical clustering according to their effects on intermediary phenotypes. We constructed PRS for each among breast, colorectal, prostate and pancreatic cancers and T2D using 171, 90, 154, 21 and 403 established (P<5×10^-8) independent SNPs for each disease, respectively. Each PRS was weighted by the effect sizes for the disease it was associated with. The effect sizes were calculated in a dataset different from where the PRS were tested. We then evaluated the association between each of the PRS and the risk of T2D, and the effect of T2D PRS on the risk of four cancers in 14,553 individuals from the European Prospective Investigation into Cancer cohort including 2,120 and 499 with colorectal and pancreatic cancer, respectively; and 12,719, 12,650 and 16,671 individuals from the UK Biobank with breast cancer, prostate cancer and T2D, respectively.

We demonstrated that the T2D PRS is significantly associated (P < 0.0125) with all four cancers, while only breast and prostate cancers PRSs are significantly associated with T2D risk (OR[CI]=1.16[1.14-1.17], P=3.91×10^-79 and OR[CI]=1.06[1.05-1.08], P=2.08×10^-14, respectively).

We then used hierarchical clustering to partition variants for four cancers and T2D that were used in PRS construction, and 97 established glycaemic traits variants by evaluating their effects on related traits, including lipid levels, sex hormone levels, glycaemic traits, anthropometric measurements and inflammation markers, into the optimal number of groups that define biological pathways. After the QC and LD clumping, we obtained three clusters of independent signals, representing glycaemic pathway (N_SNPs = 80), lipid and hormonal pathway (N_SNPs = 139) and a mixed group (N_SNPs = 346). We constructed PRS based on each of these three groups and tested them against risk of T2D and four cancers. Variants classified in hormonal pathways were associated with breast (OR[CI]=1.06[1.04-1.08], P=3.83×10^-9) and prostate (OR[CI]=1.07[1.04-1.10], P=7.46×10^-7) cancers, but not colorectal and pancreatic cancers. Glycaemic group variants affect prostate (OR[CI]=1.07[1.04-1.09], P=1.71×10^-6) and pancreatic (OR[CI]=1.11[1.01-1.21], P=0.025) cancers, but not colorectal and breast cancers.

Using data from the UK Biobank, EPIC and GECCO studies, we provide evidence for shared genetic factors between T2D, and four cancers.