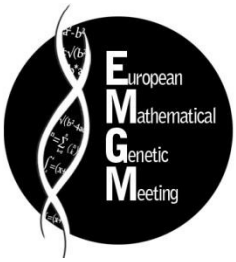
Relatedness distribution estimation between individuals in multi-population panels

Fabien Laporte, Alain Charcosset, Tristan Mary-Huard



EUROPEAN MATHEMATICAL GENETICS MEETING 2015

Brest, France

April 16-17

Organizing and Scientific Committees

- Local Organizing Committee - Inserm UMR 1078
    - Claude Férec
    - Emmanuelle Génin
    - Sophie Podeur
    - Karen Rouault
- Scientific Committee
    - Christian Dina, Nantes
    - Emmanuelle Génin, Brest
    - Anne-Louise Leutenegger, Paris
    - Hervé Perdry, Paris

Special thanks

René Vigouroux and Katell Vigouroux and Association Gaetan Saleun for their help with the organisation of the meeting and their support.

University of Bretagne Occidentale and Faculty of Humanities and Social Sciences for hosting the meeting

Océanopolis Brest for hosting the conference dinner.

Virginie Clerget for designing the EMGM logo.

A1 - Session 4 - Thursday 16 April 16:20

Comparison of methods for inferring causal pathways between genotype and phenotype preference

Ainsworth HF [a], Cordell HJ [a]

[a] Institute of Genetic Medicine, Newcastle University, UK

Many novel associations between genetic variants and human disease have been successfully identified using genome-wide association studies (GWAS). However, a typical GWAS gives little insight into the biological function through which these associated genetic variants are implicated in disease. Indeed, rather than finding variants which directly influence disease risk, the variants implicated by GWAS are typically in linkage disequilibrium with the true causal variants. Understanding the causal role the genetic variants play in disease and moving towards therapeutic interventions is not simple. Integration of additional data such as gene expression, proteomic and metabolomic data, measured in relevant tissue in the same individuals for whom we have GWAS data, could potentially provide further insight into disease pathways.

We review currently available statistical methods for inferring causality between variables that include a genetic variant, which can be used to anchor the direction of the causality. We consider Mendelian Randomisation, Structural Equation modelling, a Causal Inference Test and a Bayesian method. We present a simulation study assessing the performance of the methods under different conditions, assuming throughout that we have genotype data along with two observed phenotypes. In particular, we consider how the causal inference is affected by the presence of common environmental factors influencing the observed traits.

A2 - Poster P1

Detecting Gene-Environment Interaction using Logistic Regression with Latent Exposure (LRLE)

Alarcon F [a]*, Nuel G* [b]

[a] MAP5, UMR CNRS 8145, Université Paris Descartes, Paris, France

[b] LPMA, UPMC, Sorbonne University, Paris, France

The interest to study gene-environment (G×E) interactions has increased this last several years (Murcray et al., 2009; Thomas, 2010; Dai et al., 2012). Indeed, accounting for these interactions in Genome-Wide Association Studies (GWAS) can provide a better understanding of the disease. Two main approaches have been developed to detect G×E interaction, based on logistic regression: 1) The cases- controls (CC) approach in which the disease is considered as the response variable and a likelihood ratio test is used to test the G×E interaction. 2) The cases-only (CO) approach making the assumption of G×E independence, in which gene is considered as the response variable only among the cases.

However, to date, only few loci that interact with environment have been discovered demonstrating that the problem is very challenging due to various causes including the presence of a non observed factor acting as confounding factor. Indeed, it is very common that the causal environmental exposure is unobserved and that only proxy covariates of the causal exposure are observed. In this case, the causal exposure acts as confounding factor and causes a drastic loss of power in GE interaction detection.

In this work, we suggest to account for these unobserved confounding factors by introducing a binary latent exposure in the logistic regression. The insertion of a Ridge penalty in the likelihood allows this logistic regression with latent exposure (LRLE) model significant power gain in the detection of the GE interactions.

First, a validation of the LRLE model through a simple simulated dataset is performed as well as a calibration of the penalization parameter. After which, the performances of our new approach to detect GE interaction is studied according to the GE interaction effect and to the proxy effect. Finally, the performances of the LRLE model are compared with those of the two standards approaches (the cases- controls test and the cases-only test).

Results clearly demonstrate that the LRLE approach is much more powerful than the standard approaches.

A3 - Session 2 - Thursday 16 April 11:50

Interactions analysis with covariates

Ambroise B [a], O'Donovan M [a], Owen M [a], Pocklington A [a], Escott-Price V [a]

[a] Cardiff University, UK

Schizophrenia is a highly heritable disorder in which genetic factors account for ~80% of the variability in liability. Genome wide association studies (GWAS) have shown that there are potentially thousands of risk loci associated with the disease (Sullivan et al., 2003). Although it has been shown that risk for schizophrenia is partially explained by additive effects of top-ranking single SNPs, gene-gene interactions may help to explain additional heritability that contributes to risk above and beyond that explained by single SNPs (Hemani et al., 2014; Zuk et al., 2012).

When testing for statistical interactions between two independent variables, the effect size of the interaction term between those two variables is tested. However when the effect of covariates needs to be accounted for, the widely accepted method involves simply adding the covariates into the regression model.

We compared three different approaches that take into account covariates when testing for interactions. These methods were applied to GWAS data from the international Schizophrenia Consortium, consisting of a total of 3,322 cases and 3,587 controls that were drawn from 8 different European populations (International Schizophrenia Consortium, 2008).

The first method is the logistic regression analysis with simple adding the population covariates into the model. The second method adds the covariates into the model as well as the interactions terms between the covariates and the SNPs as suggested in Yzerbyt et al., 2004. The third method tests for the interaction in each population separately and then combines the interaction effects by means of a meta-analysis of the studies.

As comprehensive pair-wise SNP-SNP interaction analysis is time consuming and computationally intensive, intelligent pruning was used to reduce the number of SNPs involved in the analyses.

Comparison of the three methods will be presented, and limitations for each method will be addressed.

A4 - Session 5 - Friday 17 April 9:00

Multivariate Phenotypes, Familial Data, and Pleiotropy

de Andrade M [a], Xue L [b], Soler JMP [c]

[a] Mayo Clinic, Rochester, MN, USA

[b] Boston University, Boston, MA, USA

[c] University of São Paulo, SP, Brazil

Several models and statistical methods have been proposed to identify genetic variants with pleiotropic effects in humans, animals and plants. From the proposed models there is a variation about the number of phenotypes and the source of the data (related or unrelated). When analyzing several phenotypes using related data simultaneously there is a limitation due to the number of parameters to be estimated unless several restrictions are applied. In this paper we proposed a variation of the seemingly related regression (SUR; Zellner, 1962) to identify pleiotropic genetic variants. Our proposal consists of three steps: first remove the familial structure for each correlated phenotype, use their residuals as the new unrelated phenotypes (as in GenABEL; Aulchenko et al., 2007) and applied the SUR equations taking into account the correlation between the traits. As an application we use the Baependi family study that consists of 80 families and 1,100 subjects, genotype data from Affymetrix 6.0, and the metabolic syndrome variables as our multivariate phenotypes. Our results will be compared with the first principal component for the metabolic syndrome variables as the phenotype using the variance components approach taking into account the family structure, a recent proposed method using ITEM response theory (Fragoso et al., 2014), and with PheWAS (Phenome-Wide Association Studies) (Carroll et al., 2014).

A5 - Session 3 - Thursday 16 April 14:20

Population stratification in secondary genetic association studies

Babron M-C [a,b], Benhamou S [a,b,c], Génin E [d], Kazma R [e,f]

[a] Inserm UMR946, Variabilité Génétique et Maladies Humaines, Paris, France

[b] Université Paris-Diderot, Sorbonne Paris-Cité, UMR946, Paris, France

[c] Gustave Roussy, Service de Biostatistique et d'Epidémiologie, Villejuif, France

[d] Inserm UMR1078, Génétique, Génomique Fonctionnelle et Biotechnologies, CHU Morvan, Brest, France

[e] Centre National de Génotypage, Institut de Génomique, CEA, Evry, France

[f] Novartis Institutes for BioMedical Research (NIBR), Biomarker Development, Basel, Switzerland

Population stratification is a potential cause of false positive results in genome-wide association studies, when cases and controls are drawn from a population comprising multiple groups with different disease prevalences. To correct for stratification, the inclusion of several principal components (PCs) of genome-wide genotypes as covariates has been shown to control for the inflation of association statistics and has become standard procedure.

After testing for genetic association with a primary phenotype of interest, secondary hypotheses are often tested using a fraction of the initial sample for which the secondary phenotype is available. In practice, correction for subsample stratification is done using the PCs that were calculated for the full initial sample. This approach makes the assumption that the subsample has a population structure similar to that of the initial sample, used to calculate the PCs. However, this might not always be the case. In fact, the strategy that uses PCs calculated on the initial sample to correct for stratification, when testing secondary hypotheses on a subsample, has not yet been evaluated.

Here, we assess the robustness of the correction using PCs calculated with the full initial sample or a subsample that comprises a distribution of subpopulations that is different from the initial population.

First, we gathered 10,876 individual genome-wide genotypes (Illumina 317k) from three large studies (MDACC, Arcage, and EuroPa) that we considered as the initial case-control sample. Then, to simulate subsamples with a different population structure, we randomly draw sets of cases and controls with different proportions from extremes. These extremes match up with the North-East and South-West quadrants of the space defined by the first two PCs calculated on the initial sample. Simulating 100 replicates of a subsample with 75% of cases and 25% of controls belonging to the North-East quadrant (the rest belonging to the South-West quadrant), the Q-Q plots show a strong type 1 error inflation ($\lambda GC=1.31$) without population stratification correction which is corrected satisfactorily with as low as 2 PCs calculated on the initial sample ($\lambda GC=1.00$) or on the subsample ($\lambda GC=0.99$). Moreover, for specific variants in genes which are strongly stratified along this axis (OCA2, RALGPS1), the correction is efficient whatever the set of PCs used.

However, this was not always the case when simulating more complex patterns of subsample stratification.

A6 - Session 1 - Thursday 16 April 9:50

Genotype imputation of family data

Böhringer S [a], Hilbers FS [b], Devilee P [b], Pfeiffer R [c]

[a] Leiden University Medical Center, Medical Statistics and Bioinformatics, Leiden, The Netherlands

[b] Leiden University Medical Center, Department of Human Genetics, Leiden, The Netherlands

[c] Biostatistics Branch, Department of Cancer Epidemiology and Genetics, NCI, NIH, Bethesda, USA

In genome wide association studies, missing genotypes and unobserved markers are often imputed. This allows inclusion of all individuals in an analysis of genetic data, and to harmonize genotype data from different studies for pooled analyses. Available software for imputation typically assumes that individuals in a dataset are unrelated individuals. Imputation approaches that accommodate family data are limited. We therefore propose and study a Bayesian approach for genotype imputation that uses an MCMC sampler to directly draw values from the posterior distribution of haplotypes in a region compatible with the respective family structure. We develop this model for general multi-generation family structures without loops. These draws can be used to compute posterior genotype dosage for a locus to be imputed or to generate multiple imputed data sets containing either genotypes or full haplotypes. Information from reference panels is included through the prior distribution of haplotype frequencies in a population.

We compare the Bayes imputation model with a naive approach that imputes family members by ignoring pedigree structure and show that imputed values from the Bayes model are unbiased. We illustrate our method by imputing missing genotypes in individuals sampled into a family study of breast cancer in the Netherlands containing up to 36 members per pedigree and relating them to the cancer outcome.

A7 - Poster P2

Cost effective assay choice for rare disease study designs

Campbell DD [a,b], Porsch RM [a], Cherny SS [a,b,d], Capra V [e], Merello E [e], De Marco P [e], Sham PC [a,b], Garcia-Barceló M-M [c]

[a] Department of Psychiatry, University of Hong Kong, China

[b] Centre for Genomic Sciences, University of Hong Kong, China

[c] Department of Surgery, University of Hong Kong, China

[d] State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, China

[e] Istituto Giannina Gaslini, Genoa, Italy

New technologies such as Next Generation Sequencing have allowed investigation of the aetiology of diseases not previously amenable to genetic analyses due to their rarity, small family size and/or locus heterogeneity. For Mendelian diseases, relatively small numbers of small pedigrees are often sufficient to reduce the candidate risk mutations to a manageable number. Of the estimated 7,000 rare severe disorders, approximately half have now had risk genes identified for them (Boycott et al., 2013). This has left a rump of rare disorders whose aetiology may be resistant to such study designs. Study design for such diseases is hampered not only by ignorance of their aetiology, but also by the number and variety of assays available, and the typically high per subject cost of such assays. Researchers of rare diseases on constrained budgets may be limited not so much by the availability of disease cases as by assay costs, making assay choice a critical issue. We have developed a framework for assay choice that maximises the number of true disease causing mechanisms 'seen', given limited resources. Although straightforward, some of the ramifications of our methodology run counter to received wisdom on study design. We illustrate our methodology with examples, and have built a website allowing calculation of quantities of interest to those designing rare disease studies.

A8 - Poster P3

Investigation of Assortative Mating in Autism Spectrum Disorders

Connolly S [a], Anney R [a], Gallagher L [a], Heron E [a]

[a] Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College, Dublin, Ireland

Assortative mating is a non-random mating system in which individuals with similar genotypes and/or phenotypes mate with one another more frequently than would be expected in a random mating system. Autism Spectrum Disorders (ASD) are considered to be heritable neurodevelopmental disorders, characterised by patterns of repetitive behaviours and deficits in language and social behaviour. This theory has been hypothesised in ASD in order to help explain some of the increase in the prevalence of ASD over the last number of years (Baron-Cohen, 2006). The hypothesis is that positive assortative mating on parents with autistic traits increases the variance of those traits, and the increase in variance causes more cases of ASD among their offspring.

As assortative mating can be explored through both phenotypic and genotypic data, we are interested in investigating both approaches to determine whether or not there is evidence of assortative mating in ASD.

The genotype data that we will be investigating is genome-wide Single Nucleotide Polymorphisms (SNP) data on trio families (Autism Genome Project (AGP) data and Simons Simplex Collection (SSC) data). We are investigating evidence of excess of genotypically similar mating pairs. There are a number of means of detecting this excess, such as the Mating Type Distortion Test (MTDT) (Sebro et al., 2010) and Identical By Descent (IBD) approaches. Applications of these approaches to the AGP and SSC datasets will be presented here.

The phenotype data available on the parents of the ASD offspring that facilitates exploration of non-random mating are the Broad Autism Phenotype Questionnaire (BAPQ) and the Social Responsiveness Scale (SRS). These measures aim to capture ASD traits within the parents, if there is evidence of these traits within the parent-pairs this would indicate potential non-random mating. We examine both of these phenotypic measures for assortative mating in these two datasets.

A9 - Session 8 - Friday 17 April 16:20

The impact of population stratification on genomic heritability estimation

Dandine-Roulland C [a,b], Bellenguez C [c,d], Génin E [e,f], Perdry H [a,b]

[a] Université Paris-Sud, France

[b] Inserm U1178, Villejuif, France

[c] Inserm U1167, Lille, France

[d] Institut Pasteur de Lille, France

[e] Inserm U1078, Brest, France

[f] Université de Bretagne occidentale, France

The heritability of a quantitative trait is the proportion of its variance which is explained by genetic factors in a linear model with additive allelic effects, assuming there is no epistasis, no gene × environment interaction, and independence between genetic and environmental factors. Heritability has long been estimated on familial data; however, in the last few years, it has been proposed to use mixed-models procedures to estimate the heritability of quantitative traits using genome-wide SNP data from unrelated individuals. In this approach, all SNPs of the genome are supposed to have additive allelic effects, modeled as independent and identically distributed normal random variables.

A flaw of this method is its vulnerability to the presence of population structure in the sample, which can create a upward bias in the heritability estimation. For example, the mean level of an environmental factor can vary significantly with latitude at the scale of a country (such as the sun exposure, the diet, etc); this, together with the existence of a north-south population stratification, will bias upward the heritability of any phenotype depending on such an environmental factor.

It has been proposed to take into account population structure by incorporating the first k principal components of the genomic data as fixed effects in the model, together with normal random allelic effects. We show that this procedure is equivalent to a regression on all principal components, with fixed effects for the k first components and random effects for the others. Its behavior (in particular, its sensitivity to the value of k) is investigated on simulated and on real data sets.

A10 - Poster P4

Exome array GWAS in 10,000 Germans identifies association between MUC22 and multiple sclerosis

Dankowski T [a,*], Buck D [b,*], Andlauer TF [c], Antony G [d], Bayas A [e], Bechmann L [f], Berthele A [b], Bettecken T [c], Chan A [g], Franke A [h], Gold R [g], Graetz C [i], Haas J [j], Hecker M [k], Herms S[l,m,n,o], Hohlfeld R [p], Infante-Duarte C [q], Jöckel K-H [r], Kieseier B C [s], Knier B [b], Knop M [t], Lichtner P [u,v], Lieb W [w], Lill C M [i], Limmroth V [x], Linker R A [y], Loleit V [b], Meuth S [z], Moebus S [r], Müller-Myhsok B [c], Nischwitz S [t], Nöthen M M [l;m], Friedemann P [q], Pütz M [aa], Ruck T [z], Salmen A[g], Stangel M [ab], Stellmann J-P [ac], Strauch K [ad,ae], Stürner K H [ac], Tackenberg B [aa], Then Bergh F[af], Tumani H [ag], Waldenberger M [ah,ai], Weber F [t], Wiendl H [z], Wildemann B [j], Zettl U K [k], Ziemann U [aj], Zipp F [i], Hemmer B [b,ak,+], Ziegler A [a,al,am,+] on behalf of the German Competence Network for Multiple Sclerosis (KKNMS) **equal contribution as first author, +equal contribution as senior author*

[a] *Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany*
[b] *Department of Neurology, Technische Universität München, Munich, Germany*
[c] *Max-Planck-Institut für Psychiatrie, Munich, Germany*
[d] *Central Information Office (CIO), Philipps University Marburg, Marburg, Germany*
[e] *Department of Neurology, Klinikum Augsburg, Augsburg, Germany*
[f] *Department of Neurology, University of Leipzig, Leipzig, Germany, present affiliation: Institute of Medical Microbiology, Otto-von-Guericke University, Magdeburg, Germany*
[g] *Department of Neurology, St. Josef Hospital, Ruhr-University Bochum, Bochum, Germany*
[h] *Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany*
[i] *Department of Neurology, University Medical Center Mainz, Mainz, Germany*
[j] *Department of Neurology, University Hospital Heidelberg, Heidelberg, Germany*
[k] *Department of Neurology, University of Rostock, Rostock, Germany*
[l] *Institute of Human Genetics, University of Bonn, Bonn, Germany*
[m] *Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany*
[n] *Division of Medical Genetics, University Hospital, Basel, Switzerland*
[o] *Human Genetics Research Group, Department of Biomedicine, University of Basel, Basel, Switzerland*
[p] *Institute of Clinical Neuroimmunology, Ludwigs-Maximilians-Universität, Munich, Germany*
[q] *Department of Neurology, Charité University Medicine Berlin, Berlin, Germany*
[r] *Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University Duisburg-Essen, Essen, Germany*
[s] *Department of Neurology, Heinrich Heine University, Düsseldorf, Germany*
[t] *Neurology, MPI of Psychiatry, Munich, Germany*
[u] *Institute of Human Genetics, Technische Universität München, Munich, Germany*
[v] *Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany*
[w] *Institute of Epidemiology and Biobank popgen, Christian-Albrechts-Universität Kiel, Kiel, Germany*
[x] *Department of Neurology, Hospital Köln-Merheim, Köln, Germany*
[y] *Department of Neurology, University Hospital Erlangen, Erlangen, Germany*
[z] *Department für Neurologie, Klinik für Allgemeine Neurologie, Münster, Germany*
[aa] *Department of Neurology, Philipps-University of Marburg, Marburg, Germany*
[ab] *Department of Neurology, Hannover Medical School, Hannover, Germany*
[ac] *Institute of Neuroimmunology and MS (INIMS) and Department of Neurologie, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany*
[ad] *Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany*
[ae] *Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany*
[af] *Department of Neurology and Translational Center for Regenerative Medicine, University of Leipzig, Leipzig, Germany*
[ag] *Department of Neurology, University of Ulm, Ulm, Germany*
[ah] *Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany*
[ai] *Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany*
[aj] *Department of Neurology, University Hospital, Eberhard-Karls-Universität Tübingen, Tübingen, Germany*
[ak] *Munich Cluster for Systems Neurology (SyNergy), Munich, Germany*
[al] *Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany*
[am] *School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa*

Introduction: Genome-wide association studies (GWAS) have successfully identified various chromosomal regions associated with multiple sclerosis (MS).
Objective: To replicate reported associations from GWAS and to identify novel associations using an exome array in a large German sample.
Methods: German MS cases (n=4476) and German controls (n=5714) were genotyped using the Illumina HumanExome v1-Chip. Genotype calling was performed with the Illumina Genome StudioTM Genotyping Module, followed by zCall.
Results: Seven regions outside the human leukocyte antigen (HLA) region showed genome-wide significant associations with MS and further 3 regions yielded p-values<10^-5. The effect of 9 SNPs in the HLA region remained (p<10^-5) after adjustment for other significant SNPs in the HLA region. All 10 regions outside the HLA region and all except one of the associations in the HLA region have been reported in previous GWAS or candidate gene studies. In addition, we observed one novel association (rs4248153 within MUC22) with MS (p=6.56·10^-16).
Conclusions: Findings from previous GWAS for MS were successfully replicated. The newly identified association between the HLA gene MUC22 and MS warrants further investigation.

A11 - Poster P5

Metabonomic signatures of CNVs in TSPAN8 exonic regions associated with Type 2 Diabetes

De T [a,b,c], Prokopenko I [b], Jarvelin M-R [c], Coin L [b,d]

[a] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

[b] Department of Genomics of Common Disease, School of Public Health, Imperial College London, London W12 0NN, UK

[c] Epidemiology and Biostatistics, Imperial College London, Norfolk Place, London, W2 1QR, UK

[d] Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia

A deletion copy number variation (CNV) in the 3rd exon of the TSPAN8 gene (CNVR5583.1) was reported to be strongly associated with Type 2 Diabetes (T2D) in the WTCCC CNV study. Here, we characterize CNVs in the coding regions of the TSPAN8 gene including CNVR5583.1 through the 1000 genomes (1KG) sequencing data and further analyse these CNVs through genotyping arrays in two Finnish cohorts, and test the association of the 1KG derived CNVs with high-throughput metabonomic (lipids) phenotypes. We discovered association with replication of two novel CNVs in the exonic region chr12:69809401-69812861 encompassing the 6th and 7th exon of TSPAN8 gene where we found a large proportion of metabonomic and lipid phenotypes to be significantly associated using univariate (Pval=7.64e-4) and multivariate (MultiPhen, O'Reilly et al, 2012) (P=0.1.33e-6) association approaches at chr12:69809401. These CNVs were further found to be nominally associated with T2D outcome (P=3.32e-7 at chr12: 69809401, using raw intensity in a T2D case-control cohort) and were also in linkage disequilibrium (LD) with CNVR5583.1 with r^2>0.5. Results from this study shows an important novel role of T2D associated CNVs in lipid regulation in humans and promises new leads in uncovering the hidden genetic and metabonomic basis for diseases like T2D.

A12 - Session 6 - Friday 17 April 11:30

Rare Variant Collapsing Test with Variable Binning

Drichel D [a], Lacour A [a], Herold C [a], Schueller V [a], Vaitsiakhovich T [a], Becker T [a,b]

[a] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[b] Institute for Medical Biometry, Informatics and Epidemiology, Bonn, Germany

Methodology of rare variant analysis underwent a rapid development since the application of collapsing tests to predefined bins was proposed (Li and Leal, 2008). The basic collapsing test is the 1 d.f. χ^2 test dichotomizing carriers and non-carriers of rare variants. An individual is considered a carrier if at least one minor allele is present in the analysis unit (bin).

Although a multitude of statistical tests exist by now, the majority relies on a twofold parameterization of the dataset: the choice of the threshold minor allele frequency (MAFT) to define “rare” and “common” variants, and the definition of analysis bins (“binning”). The choice of parameters is an educated guess at best and implies a certain Type II error rate due to uncertain signal-no-noise ratio in predefined analysis units.

The Variable-Threshold (VT) method (Price et al., 2010) eliminates the choice of a specific MAFT by analyzing the space of all possible thresholds in each bin, finding the optimal MAFT and using permutations for multiple testing correction. The COLL VT method is computationally feasible even for large genome-wide datasets with ~20,000-100,000 bins (Drichel et al., 2014).

We investigated the possibility of elimination of a priori binning in an analogous fashion. In this scenario, all possible contiguous bins, i.e. bins with all possible combinations of start and end positions are tested for association. Although the number of combinations is very large ($n(n+1)/2$ per chromosome with n rare variants), in the case of COLL, the number of bins with distinct sets of carriers is much smaller due to the characteristic “individual-wise” collapsing condition.

We applied the Variable-Binning method (VB) implemented in INTERSNP-RARE to an imputed case-control dataset containing 9832 controls and 3332 late-onset Alzheimer's disease (AD) cases, with 2,279,783 rare variants (MAF<0.05). It took <17h on a single processor and < 24Gb of memory for the complete analysis with 100 permutations. The VB method resulted in an analysis of 222,107,312 distinct bins, achieving a reduction of 99.85% compared to the full number of bins. An inflation factor of 1.060 was calculated from asymptotic p-values of distinct bins. As expected, the association with the known APOE locus was found.

Overall, the COLL VB method eliminates the choice of candidate bins, instead, the space of all possible bins is tested. This eliminates binning as a source of Type II error and is expected to improve power.

A13 - Poster P6

Impact of the tagging on the statistical power of association tests in Genome-wide association studies

Emily M [a,b]

[a] Agrocampus Ouest, Rennes, France

[b] IRMAR, Rennes, France

Genome-wide association studies (GWAS) aim at detecting correlation(s) between a phenotypic trait and a set of hundreds of thousands biological markers, called single nucleotide polymorphism or SNPs. As usual in genomics, data suffer from heterogeneity that generates dependency between SNPs. To account for spatial dependency due to linkage disequilibrium, the first step in GWAS, called tagging, consists in selecting the most informative markers to analyze a smaller set of markers. However, tagging is known to generate a decrease in the power of detection since causal variants are likely to be untagged. In that case, signal of association are tested indirectly via the corresponding tag-SNP.

In this work, we propose an explicit calculation of the power function for 3 main tests of association: the chi-squared test, the likelihood ratio test and the odds-ratio based test. Our derivations allow the power comparison of the three tests in various scenario of association, i.e. in the alternative hypothesis of tests. Our results quantify the impact of 3 main parameters: the allelic frequencies of the causal and the tag SNP, the proportion of cases in the study and the amount of correlation between the causal and the tag SNP.

In single association test, we demonstrate that odds-ratio based test is always less powerful than the Chi-squared. Furthermore, likelihood ratio test and Chi-squared are equivalent when minor allelic frequencies are reasonable (MAF>0.05). However, when variants are rare (MAF<0.05), the proportion of cases affects the power of the likelihood-ratio test and the chi-squared test in different ways.

When testing for SNPxSNP interaction, similar results are obtained in the comparison of statistical tests. However, we prove that the loss of power can be very important for all tests. In fact, if the two causal SNPs are tagged with low correlations, the signal is weaken.

The application of our results on the WTCCC 2007 data set demonstrates that the choice of the association test is crucial for improving the power of detection.

A14 - Session 7 - Friday 17 April 14:00

Two-component mixture modelling approach integrating genetic and clinical variables in analysis of time to first seizure in epilepsy.

Francis B [a], Jorgensen A [a], Morris A [a], Marson A [a], Johnson M [b], Sills G [a] on behalf of the EpiPGX consortium

[a] University of Liverpool, United Kingdom

[b] Imperial College London, United Kingdom

Time to the first seizure after randomisation to an antiepileptic drug (AED) is an important outcome in the treatment of epilepsy. Clinical factors, including gender and epilepsy type, have been associated with the first seizure outcome and potential pharmacogenetic factors are now being investigated.

A total of 964 patients from the Standard and New Antiepileptic Drug (SANAD) study, a randomised trial that compared treatments with various AEDs in patients with newly diagnosed epilepsy, were genotyped to investigate genetic biomarkers for time to first seizure, as well as other longitudinal phenotypes. Analysis was initially undertaken using a traditional one component survival model for time to first seizure, but no genome-wide significant associations with SNPs were found.

The one component model assumes the population under investigation is homogeneous, and therefore may lack power to detect SNP associations in this context. The presence of two sub-populations for time to first seizure is apparent; those who experience another seizure (susceptible patients) and those who do not experience another seizure at any point during follow-up (non-susceptible patients). To consider these sub-populations, a two component model is required for survival analysis, and mixture modelling with a cure fraction was considered to be the optimal implementation.

Initially, the two component model was applied to identify relevant clinical covariates for time to first seizure. For investigation of pharmacogenetic factors influencing time to first seizure, ten SNPs with the lowest p-values from the one component analysis were selected to be included alongside clinical covariates in the two component model. Genome-wide significant association ($p<5x10-8$) was attained for one SNP with time to first seizure and for five SNPs with susceptibility to experiencing a first seizure, highlighting potential improvements in power over the one component model.

The two component model is computationally expensive, so to test association genome-wide, two phenotypes will be derived: (i) the probability of susceptibility to further seizures; and (ii) the clinical covariate adjusted residuals from the two-component survival model. The approach of utilizing summary statistics and model residuals as phenotypes will be applied in a larger population of patients with newly-diagnosed epilepsy. This collaborative dataset is being assembled via the EpiPGX consortium (www.epipgx.eu), and includes the SANAD study cohort as well as other cohorts of patients being collected worldwide to investigate genetic biomarkers of epilepsy.

A15 - Session 3 - Thursday 16 April 15:00

Estimating inbreeding in admixed population: application to the final phase of 1000 Genomes project

Gazal S [a,b], Sahbatou M [c], Babron M-C [a,b], Génin E [d,e], Leutenegger A-L [a,b]

[a] Inserm UMR 946, Genetic variation and Human diseases, Paris, France

[b] University Paris-Diderot, Sorbonne Paris Cité, UMR946, Paris, France

[c] Fondation Jean Dausset - CEPH, Paris, France

[d] Inserm UMR1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest, France

[e] Centre Hospitalier Régional Universitaire de Brest, Brest, France

Estimating the genomic inbreeding coefficient f of an individual from his genotype data is an important issue in both population genetics, to help characterize the mating habits of a population, and medical genetics, to localize genes involved in diseases. For that purpose, we have recently developed the FSuite pipeline (Gazal et al., 2014) which is an integrative solution to exploit inbreeding.

The inbreeding coefficient is estimated from the genotype data at multiple linked markers by modeling the identity-by-descent process between the two chromosome copies of an individual through a hidden Markov model in which the emission probabilities depend on the allele frequencies in the population where individuals are sampled. The allele frequencies are usually estimated in the studied sample under the assumption that the individuals come from a homogeneous population. In the case of admixed populations, where individuals have ancestry from different populations, this assumption is violated, which could bias the f estimates. Indeed, it has been shown that single-point methods estimating kinship and inbreeding coefficients are biased in presence of admixture in the population (Moltke and Albrechtsen 2013, Thornton et al., 2012) although this is less clear for multi-point methods (Thompson and Kuhner, 2014).

In order to investigate the accuracy of FSuite f estimates in admixed populations, we used HapMap 3 haplotypes to simulate samples of individuals with different inbreeding coefficients and different levels of European and African ancestry admixture. We then compared the f estimates from FSuite and single-point methods to the true ones obtained by following the founder haplotype transmissions. We also investigated the benefits of using allele frequencies weighted by the genomic proportion of European and African ancestry. We observed that FSuite was robust to the presence of admixture in the population, while single-point methods were not. The use of weighted allele frequencies barely improved FSuite estimates while it significantly improved single-point estimates.

Finally, we applied FSuite method on the final phase of the 1000 Genomes project, which includes admixed populations (5 populations from the American continent). We found that nearly a quarter of the individuals in this panel were inbred and that around 4% of them had inbreeding coefficients similar or greater than the ones expected for 1st cousin offspring.

A16 - Session 5 - Friday 17 April 9:40

Allelic versus genotypic level tests for multivariate phenotypes

Ghosh S [a]*, Majumdar A* [a,b]

[a] Indian Statistical Institute, Kolkata, India

[b] University of California, San Francisco, USA

A complex end-point clinical trait is usually characterized by multiple quantitative precursors and hence, it has been argued that a simultaneous analysis of these correlated traits is likely to be more powerful compared to analyzing the binary end-point trait itself. Various genotype-level methods of association, such as Multiphen (O'Reilly et al., 2012) have been developed in order to identify genetic factors underlying a multivariate phenotype. On the other hand, allele-level tests are known to yield more power than genotype-level tests in case-control association analyses. Lee et al. (2013) proposed an allelic test in the context of a univariate quantitative trait and investigated its properties. In this study, we explore two allele-level tests of association for analyzing multivariate phenotypes: one based on a Binomial regression model in the framework of inverted regression of genotype on phenotype and the other based on the Mahalanobis distance between the two sample means of vectors of the multivariate phenotype corresponding to the two alleles at a SNP. Both the methods inherit the flexibility of incorporating both discrete as well as continuous traits in the multivariate phenotype vector. We study some desirable theoretical properties of the methods. Using extensive simulations, the potential of the methods in enhancing the power of detecting pleiotropic association is evaluated in comparison with MultiPhen, which is based on a genotype-level test. We find that the allelic tests yield marginally higher power compared to MultiPhen for multivariate phenotypes, and are substantially more powerful for binary traits, particularly under a recessive mode of inheritance. The advantage of the allelic tests is also demonstrated by analyzing data on three correlated phenotypes: homocysteine levels, Vitamin B12 levels and folate levels in a North-Indian study on Coronary Artery Disease.

A17 - Poster P7

General Regression Model: a powerful "model free" association test for quantitative traits allowing to determine the underlying genetic model of transmission

Gloaguen E [a,b], Dizier M-H [c], Julier C [a,b], Mathieu F [a,b]

[a] INSERM UMR-S958, Paris, France

[b] University Paris Diderot, Paris, France

[c] INSERM UMR-946, Paris, France

Introduction: Genome-wide association studies (GWAS) of quantitative traits are usually assessed on a large panel of single nucleotide polymorphisms (SNPs) using linear regression models (LRM). Without knowledge of the underlying genetic model, additive mode of transmission is currently assumed, which may result in significant loss of power in case of departure from additivity. Here, we propose a General Regression Model (GRM) allowing detection of association by testing simultaneously both additive effect and dominance deviation from additivity. Moreover GRM permits also to test for the underlying genetic model of transmission.

Aims: To compare the power of GRM to other classical regression association tests under a large panel of genetic models.

Method: We used a simulation approach to generate 100.000 replicates of a sample of 1.000 cases under several genetic models. A normally distributed phenotype was simulated using various SNP frequencies (between 0.1 and 0.4), various genetic effects (phenotypic variance explained by the causal SNP between 5 and 25 percent) and different genetic modes of transmission (recessive, dominant and additive). GRM and LRMs (LRM_ADD for additive, LRM_DOM for dominant and LRM_REC for recessive modes respectively) association tests were applied to all replicates to estimate their respective powers to detect association. In replicates showing association by the GRM, the genetic underlying mode of transmission was then tested.

Results: Although GRM has one supplementary degree of freedom, its power to detect association was close to the power obtained under the true (simulated) model. Moreover GRM had much higher power than LRM-ADD in some case of deviation from additivity: the gain of power observed for GRM versus LRM_ADD test using a threshold of 5.10-7, reached 19.8% and 71.5% when the true model was dominant or recessive respectively. When association was detected by the GRM, the correct model was determined by the GRM in most cases (reaching 82%, 99% and 100% for dominant, additive and recessive simulated modes of transmission respectively).

Conclusion: GRM test appears powerful to detect association, as much as or even more than the LRM_ADD test, especially in case of recessive inheritance. In addition it allows determining the true mode of transmission. This test is easily applied to GWAS and is not computationally more intensive than the commonly used additive model, and may be used to analyze or re-analyze new or existing GWAS datasets to identify new susceptibility loci involved in complex diseases.

A18 - Session 8 - Friday 17 April 16:00

Heritability estimation from summary statistics using generalized estimating equations

Hecker J [a,b], Prokopenko D [a,b], Lange C [a,c,d,e], Loehlein Fier H [a,b]

[a] Institute of Genomic Mathematics, University of Bonn, Germany

[b] Institute of Human Genetics, University of Bonn, Germany

[c] Channing Laboratory, Brigham and Women's Hospital, Boston, USA

[d] Department of Biostatistics, Harvard School of Public Health, Boston, USA

[e] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Under the assumption of a polygenic architecture, Yang showed for complex traits that the test statistics in genome-wide association studies are expected to be inflated, even in the absence of confounding biases like cryptic relatedness or population stratification (Yang et al., 2011).

In a recently published work, Bulik-Sullivan and Finucane (Bulik-Sullivan et al., 2015) provide a methodological approach to differentiate between an inflation resulting from a polygenic architecture and from cryptic relatedness by considering all test statistics simultaneously. This makes it also possible to estimate the narrow-sense heritability from summary statistics without requiring individual-level genotype data.

The approach of Bulik-Sullivan and Finucane (Bulik-Sullivan et al., 2015) estimates so-called LD Scores from a reference panel and utilizes these quantities as covariates in a weighted linear regression of the squared test statistics. Since the test statistics are not independent of each other, a bootstrap estimator is applied to obtain robust standard errors.

Building on the same mean model, our objective is to incorporate more useful extern information into the estimation in order to improve the efficiency of the estimation. In particular, we divide the genomic region into blocks of moderate size. For these blocks, the correlation structure between squared test statistics can be well approximated by LD information from a reference panel. Our estimation procedure is based on generalized estimating equations (GEE). We use the LD information to set up the working-correlation matrices for each block, whereas we do not require that nearby blocks are independent. We show that the GEE-related asymptotic results are still valid under reasonable assumptions. It is important to note that the working-correlation matrices are not required to be exactly the true correlation matrices in order to obtain consistent estimates and correct standard errors.

In conclusion, these results imply that our approach improves the heritability estimation framework.

A19 - Session 2 - Thursday 16 April 12:10

A Gene-Gene Interaction Meta-Analysis framework: Application to IGAP GWA studies

Herold C [a], van der Lee SJ [b], Yang Q [c], Ramirez A [d,e], van Duijn C [b], Seshadri S [c], The International Genomics of Alzheimer's Project (IGAP), Becker T [a]

[a] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[b] Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands

[c] Department of Neurology, Boston University School of Medicine, Boston, Massachusetts, USA

[d] Department of Psychiatry and Psychotherapy, University of Bonn, Bonn, Germany

[e] Institute of Human Genetics, University of Bonn, Bonn, Germany

Genetic interaction is suspected to play an important role in complex diseases. However, due to computational challenges and lack of power, convincing examples of gene-gene interaction are still missing. To enable powerful interaction studies, we developed an efficient framework that extends the standards of GWAS meta-analysis to pair-wise SNP-SNP analysis. The framework was applied to data from the International Genomics of Alzheimer's Disease Project (IGAP), comprising 14,195 late-onset Alzheimer's disease (LOAD) cases and 31,493 controls.

Analysis is performed in a two-stage design. In stage I, each of 18 GWAS studies conducted a Genome-wide interaction analysis (GWIA) of all SNP pairs with quick screening test. All SNP pairs reaching a p-value p<=5x10^-05 in at least one study were retained. In stage II, these candidate SNP pairs were redistributed to the individual studies and analyzed with logistic regression tests, now with adjustment for standard covariate parameters. We considered two tests for interaction, the allelic test with 1 degree of freedom (df) and the genotypic test with 4 df. Meta-analysis was performed with the synthesis of regression slopes method (Becker and Wu, 2007) implemented in METAINTER (Vaitsiakhovich et al., 2014). The method combines effect estimates taking into account their covariance matrix.

We identified an LD-supported SNP pair reaching experiment-wide significance (p<1x10^-12) and several gene pairs close to significance. Follow-up studies are ongoing.

A20 - Poster P8

Linkage and association analyses of carotid intima media thickness for common genomic variants: Results from the Bonn IMT Family Study and the Heinz Nixdorf Recall Study

Heßler N [a], Geisel MH [b], Coassin S [c], Moskau-Hartmann S [d], Nürnberg G [e], Hennig F [f], Bauer M [g], Möhlenkamp S [h], Mahabadi AA [g], Moebus S [b], Erbel R [g], Karl-Heinz J [b], Hoffmann B [f,i], Nürnberg P [e], Klockgether T [j], Kronenberg F [c], Scherag A[k], Ziegler A [a,l,m], on behalf of the Heinz Nixdorf Recall Study Investigative Group

[a] Institute of Medical Biometry and Statistics (IMBS), University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[b] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University of Duisburg-Essen, Essen, Germany
[c] Divison of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria
[d] Department of Epileptology, University of Bonn, Bonn, Germany
[e] Cologne Center of Genomics, University of Cologne, Cologne, Germany
[f] IUF Leibniz Institute for Environmental Medicine, Düsseldorf, Germany
[g] West German Heart Center, Department of Cardiology, University Hospital of Essen, Essen, Germany
[h] Department of Cardiology, Hospital Bethanien Moers, Moers, Germany
[i] Medical Faculty, Heinrich Heine University of Düsseldorf, Düsseldorf, Germany
[j] Department of Neurology, University Hospital Bonn, Bonn, Germany
[k] Clinical Epidemiology, Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany,
[l] Center for Clinical Trials, University of Lübeck, Lübeck, Germany,
[m] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Carotid intima media thickness (CIMT) is a widely used marker for subclinical atherosclerosis. A recent genome-wide association meta-analysis (GWAMA) in 31,000 Caucasians identified three genomic regions. Here, we report on a linkage and association analysis from the Bonn IMT Family Study (BN-IMTFS) with independent replication in the Heinz Nixdorf Recall Study (HNR).

In BN-IMTFS, we genotyped 546 individuals from 132 nuclear families of German origin using the Affymetrix GeneChip Human Mapping 250K StyI chip. The software Merlin, accounting for linkage disequilibrium, was used for linkage analysis additionally to the family-based association test (FBAT). Regions with a p-value <10-4 were used for replication in the HNR. HNR subjects were genotyped on four different Illumina chips. Imputation was performed using IMPUTEv2.3.1 and 2,821 individuals were used for association analysis. Further, we performed look-ups of the published GWAMA association results in both samples.

Three regions on chromosome 4, 6 and 9 were identified using FBAT (p-value<10-4). Linkage analysis revealed two additional interesting results on chromosome 4 and 17 (LOD>2). However, none of the family-based associations could be replicated in the HNR. Regarding the published findings from the GWAMA, we also found no evidence for a replication in both studies (p-values>>0.13).

We detected evidence for association with CIMT and potential linkage within the BN-IMTFS, which we could not be replicated in HNR. Similarly, the GWAMA could not be replicated in our family- and population-based studies. We discuss possible source of inconsistent findings.

A21 - Session 7 - Friday 17 April 14:40

Statistical methods to analyze repeated measurements of overdispersed categorical data: an application in longitudinal microbiome data.

Houwing-Duistermaat J [a], Martin I [a], Tsonaka R [a]

[a] Department of Medical Statistics and Bioinformatics, LUMC, Leiden, Netherlands

Statistical modelling of clustered categorical data can be challenging. Our work is motivated by a longitudinal study on human gut microbiome measurements where multivariate count data distributed over a number of bacterial categories are available. Typically dimension reduction is obtained by considering the composition at phylum level (e.g. 6 categories in our dataset). In the motivating study, bacterial data for subjects from helminth endemic areas at two time points are collected and the question of interest is to assess the effect of infection status on the bacteria composition. The multinomial regression model can be used for analysis. However, the presence of the correlation within a study subject (overdispersion) and the clustering over time need to be modeled properly. To account for overdispersion, the Dirichlet-multinomial model (DMM) is typically used. Since Dirichlet is a conjugate distribution of the multinomial, fitting this model is straightforward compared to the multinomial logistic mixed model (MLMM). However, the linear predictor models the Dirichlet parameter and not the mean which may hamper interpretation of the parameters. Other drawbacks are that, the correlations between the bacteria categories are forced to be all negative and extension to a multilevel data is not straightforward.

We propose to extend the MLMM and use the combined model (CM), in which we incorporate two sets of random effects: a Dirichlet distributed random effect at the mean level to accommodate the overdispersion and a set of normally distributed random effects included in the linear predictor to model the association between the repeated measurements. We study the correlation between the various bacteria categories at cross sectional setting for the three models (DM, MLMM, CM) and evaluate the performance of our new method (CM) via simulations. Finally, we exemplify our proposed method on the bacterial data and estimate the effect of infection status over time in longitudinal setting.

A22 - Session 1 - Thursday 16 April 10:10

Increased power for detection of parent-of-origin (imprinting) effects using haplotype estimation

Howey R [a], Cordell H [a]

[a] Institute of Genetic Medicine, Newcastle University, UK

Parent-of-origin (or imprinting) effects relate to the situation where traits are inherited from only one parent, with the allele from the other parent giving little or no effect. In most case/parent trios the parent-of-origin of each allele in the offspring can be deduced unambiguously, however this is not true when all three individuals are heterozygous. Most existing methods for investigating parent-of-origin effects operate on a SNP-by-SNP basis and either perform some sort of “averaging” over the possible parental transmissions or else discard these ambiguous cases. If the correct parent-of-origin could be determined then this would provide extra information and increase the power to detect imprinting effects. We propose using haplotype estimation, thus making use of the surrounding SNP information, as a means of estimating the parent-of-origin of alleles for each case/parent trio, case/mother duo and case/father duo. This extra information is then used in an extension to our multinomial modelling software PREMIM/EMIM to estimate parent-of-origin effects at SNPs across the genome. We show through the use of simulated data that our approach shows increased power over previous versions of PREMIM/EMIM, particularly when the data consists of only duos. We apply our method to two real data sets and find a general decrease in significance of p-values in genomic regions previously thought to be possibly associated with imprinting effects, thus weakening the evidence that these genomic regions harbour imprinting effects.

A23 - Session 4 - Thursday 16 April 17:00

Identification of disease risk genes in psychiatric diseases, using empirical Bayes models

Ionita-Laza I [a], McCallum K [a]

[a] Columbia University, New York, USA

Despite much progress the genetic basis of psychiatric disorders, such as autism spectrum disorders (ASD) and schizophrenia (SCZ), remains largely unknown. We focus here on copy-number variable regions (CNVRs) because many large CNV regions have been implicated in risk to psychiatric disorders such as ASD and SCZ. However the underlying disease genes in these regions are mostly unknown, because these CNVRs are large and contain many genes. Furthermore, these CNVRs have not been comprehensively investigated using the large whole-exome sequencing (WES) datasets that have become recently available for ASD and SCZ. I will discuss new empirical Bayes methods to identify risk genes within such large CNVRs, and I will show that the proposed empirical Bayes methodology can be substantially more powerful than existing methods especially so in the presence of many non-disease risk variants, and in situations when there is a mixture of risk and protective variants. I will show results from applications to ASD and SCZ WES studies.

A24 - Session 7 - Friday 17 April 15:00

Methodological considerations deriving from a PHeWAS and LaboWAS on UGT1A genotype

Jannot A-S [a,b], Loriot M-A [c,d], Narjoz C [c,d], Burgun A [a,b], Zohar S [a,b]

[a] Biomedical Informatics and Public Health Department, University Hospital HEGP, AP-HP, Paris, France

[b] INSERM UMR_S 1138 Team 22: Information Sciences to support Personalized Medicine, Université Paris Descartes, Sorbonne Paris Cité, Faculté de Médecine, Paris, France

[c] INSERM UMR-S 775, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

[d] Biochemistry, Pharmacogenetics and Molecular Oncology Unit, University Hospital HEGP, AP-HP, Paris, France

Context: Genetic variation in UGT1A1 enzyme is known to be directly related to the Gilbert's syndrome, in which the homozygous variant genotype UGT1A*28/*28 is responsible for a less efficient bilirrubin glucoronidation. More recently, it was demonstrated that genotyping of UGT1A1*28 could also serve to predict irinotecan-associated neutropenia. Liver toxicity related to this variant was also demonstrated for another treatment. In Georges Pompidou European Hospital, an 800-acute-bed academic hospital located in Paris, patients have been routinely genotyped for UGT1A before starting an irinotecan-based therapy since 2003 and results were stored in a clinical data warehouse, along with billing codes and laboratory values. This provides an opportunity to study all types of toxicity on this patient subset by performing a laboratory-wide association study (LaboWAS) and a phenome-wide association study (PheWAS).

Methods: We extracted 400 patients genotyped routinely for UGT1A between 2003 and 2014 before an irinotecan-based chemotherapy and all their biological values and ICD-10 billing codes from HEGP clinical data warehouse. We first performed a phenome-wide association study by testing the association between all ICD-10 codes and UGT1A genotype. We then performed a laboratory-wide association study by testing the association between 160 dichotomized biological value and UGT1A genotype. The threshold of dichotomization for each value was found by minimizing the p-value of the corresponding association test. Threshold of significance was estimated using a permutation-based procedure to take into account both multiple testing and dichotomization procedure.

Results: Using the PheWas approach, we showed that UGT1A genotype was associated to an increased risk of metastatic tumors, confirming that UGT1A was associated to response to chemotherapy. We did not find an association with Gilbert syndrome ICD-10 code, showing that PheWas analysis using billing code is not relevant for some conditions. We also show that PheWAS results were sensitive to the way ICD-10 codes are grouped for the analysis. For the LaboWAS, we confirmed the strong association with bilirubin level, but other associations were very sensitive to the number and the period during which biological values were taken into account.

Conclusion: PheWas and LaboWAS studies are a way to identify pleiotropic effects of a genetic variant but are very sensitive both to grouping strategy and inclusion criteria, as ICD-10 codes and biological values can be extracted from a clinical data warehouse using numerous strategies.

A25 - Session 5 - Friday 17 April 10:00

A novel method and software tool for genome-wide multi-phenotype analysis of rare variants

Kaakinen M [a], Mägi R [b], Fischer K [b], Järvelin M-R [c,d], Morris AP [e], Prokopenko I [a]

[a] Department of Genomics of Common Disease, Imperial College London, London, UK

[b] Estonian Genome Center, University of Tartu, Tartu, Estonia

[c] Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK

[d] Institute of Health Sciences and Biocenter Oulu, University of Oulu, Oulu, Finland

[e] Department of Biostatistics, University of Liverpool, Liverpool, UK

Genome-wide association studies (GWAS) of common variants (minor allele frequency, MAF>5%) have helped in identifying thousands of genetic loci associated with hundreds of complex human traits. However, since much of the estimated heritability of these traits remains unexplained, the focus has been expanded to the analysis of low-frequency and rare variants (MAF=<5%, both denoted by RVs) of modest effect. This endeavour has been enabled by large-scale sequencing and imputation efforts, as well as by methods development for RV association analysis. Improved power for variant detection is usually achieved through increased sample size, but could also be gained via multi-phenotype analysis in which the correlation between traits is accounted for. We have developed a software for genome-wide Multi-phenotype Analysis of RVs (MARV), combining features from both RV burden tests and multi-phenotype analyses. Specifically, the proportion of rare variants at which an individual carries minor alleles within a gene region is modelled on linear combinations of phenotypes in a regression framework. MARV also implements model selection via the Bayesian information criterion (BIC). We have applied the software to 4788 individuals from the Northern Finland Birth Cohort 1966 using measurements of three correlated metabolic traits: fasting insulin (FI), triglycerides (TG) and waist-to-hip ratio (WHR). Individual were genotyped on the Illumina370CNV array and imputed up to the 1000 Genomes Project all ancestries reference panel (March 2012). FI/TG/WHR were adjusted for body mass index and three principal components to control for population structure. Residuals were subsequently transformed: natural logarithm for FI and inverse normal for TG and WHR. We identified RV associations, at genome-wide significance ($p<1.7x10-6$, Bonferroni correction for 30,000 genes) in ZNF259, which maps to a common variant GWAS locus for TG and coronary heart disease. Based on BIC, the model with TG and FI provided the best fit (Pmodel=3.1x10-9; PTG=6.6x10-10; PFI=0.001), and stronger association than in univariate analyses (PTG=6.7x10-8; PFI=0.13). The runtime of MARV with three phenotypes was not substantially changed over a univariate burden test. Using our novel method and software MARV, we demonstrate its ability to identify RV multivariate phenotype associations with greater statistical significance than in univariate analyses.

A26 - Poster P9

Identifying outbreaks of two common degenerative diseases as a strategy to identify genetic rare variants

Karakachoff M [a,b,d,e], Persyn E [a,d,e], Simonet F [a,d,e], Le Marec H [a,d,e,f], Probst V [a,d,e,f], Le Tourneau T [a,d,e,f], Tallec A [c], Redon R [a,d,e,f], Schott J-J [a,d,e,f], Molinaro S [b], Dina C [a,d,e,f]

[a] INSERM UMR 1087, Nantes, France

[b] Institute of Clinical Physiology, CNR, Pisa, Italy

[c] ORS Pays de la Loire, Nantes, France

[d] CNRS UMR 6291, Nantes, France

[e] Université de Nantes, Nantes, France

[f] CHU Nantes, Institut du thorax, Nantes, France

The involvement of genetic common variants in human pathologies has been confirmed by a large number of GWAS. However, it appears that a significant portion of heritability may be due to the presence of rare variants. A large fraction of these rare variants is recent and thus tends to be concentrated in particular geographical areas, mostly shared between distantly related individuals, as a consequence of specific historical and cultural conditions. When these rare alleles increase the individual risk of being affected by the disease, we can expect increased outbreaks of the disease. This results into an increased statistical power when establishing an association between rare variant and phenotype in these regions.

The study of fine geographical distribution of diseases could be an important prerequisite for the selection of areas to target in association studies, as their frequency would be significantly increased. This is the reason why we have been working to apply classical statistical spatial methods in order to identify fine-scale geographic areas of Western France, showing high prevalence of a common degenerative disease: calcific aortic valve stenosis (CAVS).

Hospital records and mortality databases from health national systems were used to generate disease maps and to identify local clusters. One of the clusters were consistent with those of a previous study that had demonstrated a familial aggregation of CAVS in neighboring areas.

In order to test the hypothesis of correlation between diseases distribution and genetic rare variants distribution, we simulated genetic data following the characteristics of Western France population. We chose two different computational programs for simulating sequence-level genetic data: QuantiNemo (Neuenschwander, 2008) and COSI (Schaffner, 2005). The definition of a complex scenario, where demography, spatial and genetics are defined following a specific model of evolution, allowed us to simulate local populations and their genomes. Applying later different scenarios with different relative risk magnitudes (RR=2, 4, 8) we observed that some clusters could be detected under a higher risk (RR=8).

By simulating genetics and demographical processes we want to verify the assumption of correlation between diseases outbreaks and genetic distribution and to verify the applicability of this model to our regions of Western France. Assessing whether our identified pattern of disease prevalence fits with rare allele dispersal model will provide the opportunity to test the general strategy implemented to study the local distribution of rare variants in Western France and to identify rare variants causing the diseases.

A27 - Poster P10

Exome sequencing in seven families and gene-based association studies support genetic heterogeneity and suggest possible candidates for fibromuscular dysplasia

Kiando SR [a,b], Barlassina M-C [c,d], Cusi D [c,d], Galan P [e], Lathrop M [f], Plouin P-F [b,g], Jeunemaitre X [a,b,h], Bouatia-Naji N [a,b]

[a] INSERM UMR970 Paris Cardiovascular Research Center (PARCC), Paris, France

[b] Université Paris-Descartes, PRES Sorbonne Paris Cité, Paris, France

[c] Department of Health Sciences, Genomic and Bioinformatics Unit, Viale Ortles 22/4, Milano, Italy

[d] Chair and Graduate School of Nephrology, University of Milano, Division of Nephrology, San Paolo Hospital, Milano, Italy

[e] Université Paris 13, Equipe de Recherche en Epidémiologie Nutritionnelle (EREN), Centre d'Epidémiologie et Statistiques Sorbonne Paris Cité, Inserm (U1153), Inra (U1125), Cnam, COMUE Sorbonne Paris Cité, Bobigny, France

[f] Centre National de Génotypage, Evry, France

[g] AP-HP, Department of Hypertension, Hôpital Européen Georges Pompidou, Paris, France

[h] AP-HP, Refferal Center for Rare Vascular Diseases, Hôpital Européen Georges Pompidou, Paris, France

Background- Fibromuscular dysplasia (FMD) is a group of nonatherosclerotic and noninflammatory vascular disease leading to stenosis, aneurysm and dissection of medium-sized arteries, mainly renal arteries and carotids. FMD occurs predominantly in females with a prevalence of ~4/1000 for clinical forms that cause hypertension, renal ischemia or stroke. The pathogenesis of FMD is unknown and a genetic origin is suspected given its demonstrated familial aggregation. Our study objective is to identify genetic variants involved in FMD aetiology.

Methods and Results- We performed whole exome sequencing (WES) in 16 FMD cases from 7 families (5 sibpairs and 2 sibtrios). Coding variants in 3,971 genes confidently called (read depth>20X) were prioritized on their frequency (allele frequency<0.01) and in silico predicted functionality. No gene harbored variants that were shared among all affected members of at least 3 out of 7 families. Rare coding variants from 16 known causative genes of vascular and connective tissue syndromes (e.g. FBN1, TGFB2 and COL3A1) were excluded as causative in these families. Genes with at least 4 rare coding variants identified in the 16 patients were followed-up using genotyping data by exome chip (Illumina HumanExome-12v1_A Beadchip) from 249 FMD unrelated cases and 689 controls. Gene-based association of rare variants using SKAT-O showed nominal significant (P<0.05) association with multifocal FMD (N=164) for OBSCN encoding a sarcomeric protein (P=0.003), DYNC2H1 encoding a cytoplasmic dynein (P=0.02) and RNF213 previously associated with Moyamoya disease (P=0.01).

Conclusion- Our study reports data from the first WES investigation conducted for familial forms of FMD. It supports strong genetic heterogeneity for FMD and excludes the implication of several known vascular diseases causative genes in familial FMD etiology. We provide some evidence of association with multifocal FMD for OBSCN, DYNC2H1 and RNF213, though these findings need to be confirmed in independent cohorts. More powerful WES and association studies (e.g. GWAS) will better decipher the genetic basis of FMD.

A28 - Session 4 - Thursday 16 April 17:20

A predominantly European Haplotype Reference Panel of over 32,000 individuals

Kretzschmar W [a] *on behalf of the Haplotype Reference Consortium*

[a] Wellcome Trust Centre for Human Genetics, Oxford, UK

Genotype imputation is now a key tool in the analysis of human genetic studies, enabling array-based genetic association studies to examine the millions of variants that are being discovered by advances in whole genome sequencing. Examining these variants increases power and resolution of genetic association studies and makes it easier to compare the results of studies conducted using different arrays. Genotype imputation improves in accuracy with increasing numbers of sequenced samples, particularly for low frequency variants. The goal of the Haplotype Reference Consortium is to combine haplotype information from ongoing whole genome sequencing studies to create a large imputation resource. To date, we have collected information on 32,874 sequenced whole genomes, aggregated over 20 studies of predominantly European ancestry, to create a very large reference panel of human haplotypes of ~45M genetic variants. These haplotypes can be used to guide genotype imputation and haplotype estimation. We will describe the methods we have used to combine data across these cohorts and new, computionally efficient methods for phasing such a large number of samples sequenced at low-coverage. We will illustrate the substantial increases in accuracy relative to the 1000 Genomes Project Phase 1 reference panel and other smaller panels, particularly for variants with frequency <1%. Our full resource will be available to the community through imputation servers that enable scientists to impute missing variants in any study and respect the privacy of subjects contributing to the studies that constitute the Haplotype Reference Consortium. The majority of haplotypes will also be deposited in the European Genotype Archive.

A29 - Session 3 - Thursday 16 April 14:40

Relatedness distribution estimation between individuals in multi-population panels

Laporte F [a], Charcosset A [a], Mary-Huard T [a,b]

[a] INRA, UMR 0320 / UMR 8120 Génétique Quantitative et Évolution – Le Moulon, Gif-sur-Yvette, France

[b] AgroParisTech UMR518, Paris 5e, France

The relatedness between two individuals is a distribution of probabilities related to the number of alleles inherited from one or several common ancestors. It is an important concept in populations genetics, quantitative genetics and plant and animal breeding. For many years relatedness has been inferred from the pedigree information, but nowadays, thanks to genotyping techniques, relatedness between individuals of a panel can be inferred with incomplete pedigree or without pedigree.

The determination of relatedness between two or more individuals based on genotypic information is not a recent problem (Crow and Kimura, 1970). It has many applications in genetics, including genome-wide association studies (Yu, 2006) or forensic genetics. More recently, the use of the relatedness matrix has also been used in genomic prediction.

In this work, we focus on the estimation strategy proposed by Milligan who models the relatedness distribution inference problem between two individuals as a mixture model. In this model, each observation corresponds to a marker, for which the observed variables are the 4 alleles (2 alleles for each individual), which define the Identity by State (IBS) mode of the marker. The hidden variable are the ancestral origins of the alleles, which define Identity by Descent (IBD) mode of the marker. The objective is then to estimate the proportions of the different IBD modes over all available markers. Maximum likelihood inference can be performed to estimate these proportions using the EM algorithm.

We propose several extensions of the work of Milligan. Milligan assumed markers to be multiallelic (more than 3 alleles). Considering that nowadays SNP biallelic markers are widely used, and that marker information may be phased (i.e. the gametic origin of alleles is known) thanks to haplotypic reconstruction (in human genetics) or haplotypic knowledge (in plant genetics, where individuals are lines or crossing between lines), we extended the approach to biallelic markers and phased haplotypes. Moreover, it is now of common practice to consider multi-population panels that simulateously include individuals with different origins. In such cases, allelic frequencies are specific to each population, something that is not handled by most relatedness inference methods. This is explicitly taken into account within our methodology. The use of biallelic markers has important consequences on the identifiability of the mixture model. We investigate this aspect and exhibit the identifiable relatedness coefficient combinations. The extended method is available in an R package. Computational time and performances are illustrated on simulated data.

A30 - Session 1 - Thursday 16 April 9:30

The collapsed haplotype pattern method for linkage analysis of next-generation sequencing data

Leal SM [a], Wang GT [a], Zhang D [a], Li B [a], Dai H [a]

[a] Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

Traditionally, linkage analysis was used to map Mendelian diseases and genes within the linked regions were sequenced to identify the causal variants. Recent advances in next generation sequencing (NGS) make it possible to directly sequence genomes and exomes of individuals with Mendelian diseases and identify causal mutations by filtering variants in an affected individual(s) family member(s), removing those variants with higher allele frequency, e.g. >0.1% in variant databases. Linkage analysis of SNP data are sometimes used in conjunction with NGS to increase the success of identifying the causal variant. With the reduction in cost of NGS, DNA samples from entire families can be sequenced and linkage analysis can be performed directly using NGS data. Inspired by “burden” tests which are used for complex trait rare variant association studies, we developed the collapsed haplotype pattern (CHP) method to generate markers from sequence data for linkage analysis. To demonstrate the power of the CHP method compared to analyzing individual variants, we analyzed and performed empirical power calculations using the allelic architecture for several known non-syndromic hearing loss genes, i.e. GJB2, SLC26A4, MYO7A & MYH6. Power analysis demonstrated that the CHP method is substantially more powerful than analyzing individual SNVs in the presence of inter-allelic familial heterogeneity, i.e. families have different pathological variants within a gene or intra-familial heterogeneity e.g. compound heterozygotes. Specifically for an autosomal recessive model with allelic heterogeneity and locus heterogeneity of 50%, it requires 12 families for the CHP method to achieve a power of 90% for the SLC26A4 gene, while analyzing individual SNVs requires >50 families to achieve the same power at a genome-wide significance level of α=0.05. Unlike the commonly practiced filtering approaches used for NGS data, the CHP method provides statistical evidence of the involvement of a gene in Mendelian disease etiology. Additionally because it incorporates inheritance information and penetrance models it is less likely than filtering to exclude causal variants in the presents of phenocopies and/or reduced penetrance. We recommend the use of the CHP method in parallel to filtering methods to take full advantage of the power of NGS in families. The CHP method is incorporated in the SEQLinkage software which is freely available http://www.bioinformatics.org/seqlink/.

A31 - Session 6 - Friday 16 April 12:10

Statistical method for Next Generation Sequencing pipeline comparisons

Leblay N [a,b,c,d], Elsensohn M-H [a,b,c,d], Dimassi S [e,f], Campan-Fournier A [e,f], Labalme A [e], Sanlaville D [e,f], Lesca G [e,f], Bardel C [a,b,c,d], Roy P [a,b,c,d]

[a] Service de Biostatistique, Hospices Civils de Lyon, Lyon, France

[b] Université de Lyon, Lyon, France

[c] Université Lyon 1, Villeurbanne, France

[d] CNRS UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France

[e] Service de Génétique, Hospices Civils de Lyon, Lyon, France

[f] Centre de Recherche en Neurosciences, INSERM U1028-CNRS, UMR5292, Lyon, France; Université Lyon 1, Lyon, France

Sanger sequencing was the most widely used method over the past 25 years. It is now rapidly substituted by Next Generation Sequencing (NGS) which cuts drastically the time and expenses of genome sequencing. The analysis of NGS data involves many steps for which several bioinformatic tools, academic or commercial, have been developed. These tools may be combined into NGS data analysis pipelines. In this work, we developed a statistical method to compare NGS pipelines. This method was applied to the comparison of an academic pipeline (BWA-GATK) with a commercial pipeline (TMAP-NextGENe®) in two conditions: with and without reference to a gold standard (here, Sanger sequencing).

The statistical unit was the DNA base and each patient was taken as a separate study. A “meta-analysis” approach combining all the patients was used to compare first the positions then the positions and the natures of the variants found by the pipelines. The sequencing results (number of variants) were displayed in 2x2 contingency tables, one per patient, to which log-linear models for pairwise agreements between pipelines were fitted. To determine whether the margins and the ORs for agreement were heterogeneous, three log-linear models were used: a full model, a homogeneous-margin model, and a model with single odds ratio (OR) for pipeline agreements shared by all patients. Then a log-linear mixed model was fitted taking the biological variability into account through a random effect.

Our results showed that, among the 390,339 base-pairs sequenced, TMAP-NextGENe® found 1% of variants (substitutions and indels) whereas BWA-GATK found 0.8% of variants. Considering only variant positions, the mean OR for agreement was very large (OR=267.87). Using Sanger sequencing as gold standard, we found that the specificities of BWA-GATK and TMAP-NextGENe® were relatively close but significantly different (0.9920 vs. 0.9898; $p<2.2e-16$). The sensitivities were not significantly different when only variant positions were considered (0.6435 vs. 0.6174; $p=0.9994$). A lack of power due to the small number of variants found by Sanger sequencing may explain these results. Same-trend results were obtained when only substitutions were considered as variants; higher OR, specificities and sensitivities (3929.97; 0.9988 vs. 0.9990 and 0.7087 vs. 0.6796).

In conclusion, our results showed that, taking Sanger sequencing into account or not, the performances of the two pipelines were close but still improvable in terms of sensitivity.

A32 - Session 5 - Friday 17 April 9:20

Robust penalized regression for association mapping of multiple quantitative traits

Li Z [a,b]

[a] Biocenter Oulu, Oulu, Finland

[b] Department of Mathematical Sciences, and Department of Biology, University of Oulu, Oulu, Finland

Classical linear regression based association mapping methods for analyzing quantitative traits assume the normality of residuals. In reality, many quantitative traits may follow some non-normal distributions, possibly because the data include some outlying phenotypic observations, or the data are collected from multiple sources. In such case, Gaussian linear regression based methods may not be efficient and it may lose power to correctly detect some important quantitative trait loci. For improvement, one may apply some transformation methods such as log transformation to make the phenotypes more normally distributed. However, in practice, it is often difficult to find an optimal transformation method. Alternatively, robust regression methods by assuming the residual errors of a linear model to follow some more skewed distribution than a normal distribution might be appropriate and efficient for analyzing skewedly distributed data. We consider a least absolute deviation (LAD) regression based approach, which assumes the residual errors to follow a Laplace distribution. The LAD regression can be combined with the popular LASSO or l1 norm penalty, and provides shrinkage estimation and variable selection on multiple SNPs. We further generalize the LAD-LASSO penalized regression to multivariate case, so that it can be used for analyzing multiple correlated traits (Möttönen and Sillanpää et al., 2014; Li et al., 2014). Penalized regression can be used together with stability selection or multiple-split test as QTL decision rules. By evaluating our robust penalized regression approach on both simulated and real data sets, we show that the method perform as well as standard penalized if the trait is normally distributed, but show greater power for QTL detection if the trait having outliers or following some skewed distributions.

A33 - Session 3 - Thursday 16 April 15:20

Genetic Ancestry and Mortality in Chile

Lorenzo Bermejo J [a], González Silos R [a], Peil B [a], Marcelain K [b], Fuentes M [b,c], Rothhammer F [b,c]

[a] Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

[b] Program of Human Genetics, Institute of Biomedical Sciences, Medical Faculty, University of Chile, Chile

[c] Instituto de Alta Investigación, Tarapacá University, Chile, and Centro de Investigaciones del Hombre en el Desierto, Arica, Chile

Several publications have recently examined the genetic admixture in North and Latin America. We have conducted an aggregate-data study to investigate the relationship between ancestry components and mortality rates in Chile. Our study relied on genome-wide single nucleotide polymorphism data from 1376 Chileans, and on 639,789 deaths caused by 500 different diseases registered between 2005 and 2011. The ADMIXTURE program was used for a supervised estimation of individual ancestry components. After fitting a linear regression model to estimate the expected regional ancestries adjusting for age, gender, educational level, socioeconomic status and salary, Poisson regression was used to quantify the association between regional disease-specific mortality rates and the expected ancestry taking age, gender and calendar year into account.

The average Native American, European and African ancestry components in Chile amounted 48%, 49% and 3%, respectively. The Native American proportion was increased in the north (Arica y Parinacota and Tarapacá regions) and in the south (La Araucanía, Los Ríos, Los Lagos and Aisén regions) of Chile. Several associations were identified. For example, a 1% increase in European ancestry translated into a 2.5% increase in the mortality risk due to hypertensive hearth disease (95% confidence interval 1.5% to 3.6%, Pval=4e-06), and a 2.6% increased mortality risk due to chronic ischaemic heart disease (95% confidence interval 1.3% to 3.9%, Pval=e-04). Among all investigated neoplasms, only the mortality due to gallbladder cancer was associated with Native American ancestry (0.6% increase in mortality per 1% increase in Native American ancestry, 95% confidence interval 0.3% to 0.9%, Pval=3e-05). Native American ancestry was positively associated with the mortality due to asthma, and negatively associated with the mortality risk due to pulmonary edema. Details on the study design, associations with mortality attributable to other diseases, and the independent validation of results will be presented during the meeting.

A34 - Session 4 - Thursday 16 April 16:40

Discovery and fine-mapping of EGFR susceptibility loci through trans-ethnic meta-analysis

Mahajan A [a], Haessler J [b], Okada Y [c], Stilp A [d], Laurie C [d], Franceschini N [e], Morris A [a,f]

[a] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

[b] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, United States

[c] Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan

[d] Department of Biostatistics, University of Washington, Seattle, United States

[e] University of North Carolina, Chapel Hill, United States

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

Genome-wide association studies (GWAS) have been successful in identifying loci for estimated glomerular filtration rate (eGFR). However, these loci are typically characterised by common lead SNPs with association signals extending over large genomic intervals containing multiple transcripts. As a result, limited progress has been made in identifying causal variants and understanding the downstream disease pathogenesis. To address these drawbacks, we performed trans-ethnic meta-analysis to: (i) discover novel eGFR loci; and (ii) fine-map known eGFR loci by leveraging differences in linkage disequilibrium between diverse populations.

We considered eight GWAS comprising of 59,880 individuals of European, African American, Hispanic, and East Asian ancestry, each supplemented by imputation up to the 1000 Genomes Project reference panel (March 2012 release). Within each study, association with eGFR (MDRD equation) was tested under an additive model. We then combined association summary statistics across studies: (i) using trans-ethnic fixed effects meta-analysis, for discovery; and (ii) with MANTRA, 1Mb up and down of the lead SNP at eGFR loci, and constructed “credible sets” of SNPs that encompass 99% of the posterior probability of being causal.

We identified six novel eGFR loci at genome-wide significance ($p<5.0x10-8$), the strongest association signals mapping near LRP2 ($p=8.8x10-9$) and NFATC1 ($p=1.3x10-8$). We resolved fine-mapping of potential causal variants to less than ten variants at seven known eGFR loci: UMOD/PDILT (1 SNP, 1bp), GCKR (3 SNPs, 11.7kb), RGS14 (3 SNPs, 6.8kb), MPPED2 (3 SNPs, 19.6kb), BCAS3 (4 SNPs, 16.1kb), SHROOM3 (6 SNPs, 28.2kb), and UNCX (6 SNPs, 5.8kb). At GCKR, the credible set covers three SNPs including GCKR P446L, a predicted functional variant at this locus. Potential causal variants at the remaining six loci, map to introns or overlap regulatory elements from ENCODE, thereby highlighting potentials mechanism for the action of these loci on eGFR.

Our findings provide evidence that trans-ethnic GWAS can be used for discovery of novel loci and to fine-map potentially causal variants that can be taken forward for experimental validation and could help to further our understanding of the biological mechanisms underlying disease.

A35 - Session 6 - Friday 17 April 11:50

A statistical association test for the identification of clustered disease risk variants

Persyn E [a], Karakachoff M [a], Simonet F [a], Schott J-J [a], Redon R [a], Bellanger L [b], Dina C [a]

[a] Institut du thorax, Inserm UMR 1087 / CNRS UMR 6291, CHU de Nantes, France

[b] Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, Université de Nantes, France

Genome-wide association studies have identified numerous common variants associated with a wide variety of complex diseases. However these variations only explain a small proportion of the heritability. A hypothesis is that rare variants may play an important role in disease risk. Testing association between rare variants and diseases represents a challenge due to their very low frequency in the population. Statistical methods have been developed testing the association with group of rare variants, since CAST (Cohort Allelic Sum Test) described in 2007 by Morgenthaler and Thilly.

Few of these tests (Fier et al., 2012; Ionita-Laza et al., 2012; Lin, 2014; Schaid et al., 2013) take into account spatial genetic information. However, it is has been shown that disease variants may cluster in important functional domains of genes. For instance, pathogenic mutations are localized in the gene FLNA, causing congenital malformations (Robertson et al., 2003). Ionita-Laza et al. also identified in 2012, clusters of rare disease variants in the gene LRP2, associated with autism spectrum disorders.

We developed a statistical test, DoEstRare, whose aim is detecting clusters of disease rare variants while preserving sufficient power when there is no cluster but still enrichment of rare alleles (overall the sequence). The DoEstRare statistics consists in comparing the mutation distributions, estimated by kernel method, between cases and controls.

We compared the power and the type I error of our method to several published association tests for rare variants. Power and type I error computations are based on simulations conducted under two main genetic scenarios: absence or presence of one cluster of causal variants, varying also the proportion of causal variants. We are also now simulating with the software COSI according to the model developed in Schaffner et al. (2005) in order to mimic the demographical history of the European population.

We observed the change in power according to the introduction of different statistical components. This test is thus adapted to non-cluster situations with the use of a burden component. A weighting scheme is also adopted in order to better discriminate between causal and neutral variants.

We show consistent increase in power in both scenarios (a 19.4-39.7% increase compared to SKAT-O). While this may be specific of the simulations carried out here, we think that DoEstRare represents a convenient and powerful alternative to test rare allele variants effects when there is no prior hypothesis of the real distribution of causative alleles.

A36 - Session 3 - Thursday 16 April 14:00

Utilizing the Jaccard index to reveal population stratification in sequencing data: A simulation study and an application to the 1000 Genomes Project

Prokopenko D [a,e], Hecker J [a,e], Silverman E [b], Pagano M [c], Noethen M [e], Lange C [a,b,c,d], Loehlein Fier H [a,e]

[a] Institute of Genomic Mathematics, University of Bonn, Germany

[b] Channing Laboratory Brigham and Women's Hospital, Boston, USA

[c] Department of Biostatistics, Harvard School of Public Health, Boston, USA

[d] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[e] Institute of Human Genetics, University of Bonn, Germany

Population stratification is one of the major sources of confounding in genetic association studies, potentially causing false-positive and false-negative results. The effectiveness of existing adjustment approaches which are mostly built on the estimation of the genetic variance/covariance matrix is unclear for rare variants, since those variants are genetically much “younger” and might represent a different pattern of population structure. Here, we present a novel approach for the identification of population substructure in high density-genotyping data/next generation sequencing data. The approach exploits the co-appearances of rare genetic variants in individuals. The method can be applied to all available genetic loci, does not require linkage disequilibrium (LD) pruning, and is computationally fast. Using sequencing data from the 1000 Genomes Project, the features of the approach are illustrated and compared to existing methodology (i.e. EIGENSTRAT). We find that our approach works particularly well for genetic loci with very small minor allele frequencies. The results suggest that the inclusion of rare-variant data / sequencing data in our approach provides a much higher resolution-picture of population-substructure than it can be obtained with existing methodology. Furthermore, we performed extensive simulation studies based on the minor allele frequencies of the European populations. We find scenarios where our method was able to control the type 1 error more precisely and showed higher power.

A37 - Poster P11

Comparison of three calling algorithms for genotyping of rare variants

Riveros Mckay Aguilera F [a], Marenne G [a], Mistry V [b], Wheeler E [a], Farooqi S [b], Barroso I [a]

[a] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

[b] Wellcome Trust-MRC Institute of Metabolic Science, University of Cambridge, Cambridge CB2 3DY, UK

Array-based genotyping platforms capturing common variants have been extensively used for genome-wide association analyses (GWAS) and genotype-calling algorithms are well established, and accurate, in this setting. More recently, newer genotyping platforms capturing lower frequency variants have been developed and extensively deployed to extend the range of GWAS. However, the genotype-calling algorithms for rare variants are not as well characterised and their accuracy has been less extensively studied.

In this work, we took advantage of 415 samples from the Severe Childhood Onset Obesity Project (SCOOP) that were genotyped on the Illumina HumanCoreExome beadchip (genotyping done using a total of 3657 samples from 3 different cohorts) and were whole-exome sequenced (WES – Agilent HumanAllExon 50 Mb – 56x – multi-sample calling using 5233 samples from the UK10K project). We considered the WES data as the gold standard and compared them to the SNP genotypes generated using three commonly used genotype-calling algorithms: Illuminus, GenCall and GenCall followed by ZCall (ZCall). Quality control (QC) based on Hardy-Weinberg p-values and call rates were applied for each algorithm. The proportions of SNPs that survived QC were 92.61%, 98.05% and 99.27% for Illuminus, GenCall and ZCall, respectively. In order to describe the results, we defined four minor allele frequency (MAF) bins for variants detected in the 415 samples: monomorphic, rare ($0\%<MAF<1\%$), low-frequency ($1\%\leq MAF<5\%$) and common variants ($MAF\geq 5\%$). In terms of sensitivity to detect minor alleles, ZCall outperformed the two other algorithms for all ranges of frequencies. The sensitivity was lower for rare variants: 92.22%, 97.68% and 99.68% for Illuminus, GenCall and ZCall respectively, and it was higher for common variants (>99.79% for all three algorithms). In terms of false positive calls (containing the minor allele while the genotype is homozygous for the major allele according to the WES data), their proportion decreased as the allele frequency increased, and GenCall consistently outperformed Illuminus and ZCall for this criterion. GenCall false-positive rates were 2.38%, 1.12%, 0.14% and 0.08% for monomorphic, rare, low-frequent and common variants respectively.

In conclusion, our results suggested that: 1) Illuminus performed poorly compared to GenCall and ZCall; 2) Additional calls made by ZCall improved sensitivity but also increased the rate of false-positive calls. This comparison supports the use of GenCall genotype calls, including for rare variants at least for genotyping batch sizes comparable to the one we used in our experiment (N>3500).

A38 - Session 2 - Thursday 16 April 11:30

Extension of the One-Degree-Of-Freedom Test for Supra-Multiplicativity of SNP Effects in Logistic Regression Models

Schüller V [a], Drichel D [a], Lacour A [a], Vaitsiakhovich T [a], Becker T [a,b], Herold C [a]

[a] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[b] Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany

A possible reason for the phenomenon of missing heritability encountered in the genetics of complex diseases is the omission of interaction effects in genome-wide association studies (GWAS). Even pairwise SNP-SNP-interaction may be a too simple assumption in the presence of polygenic influences.

In order to address multi marker models, we recently developed the test for supra-multiplicativity (SMT) that is applicable to SNP sets of up to 1,000. The SMT requires only one degree of freedom and is designed to detect amplifying interaction as implied by liability threshold models. In general, the SMT allows deviation from multiplicativity. A drawback of the original approach is the need to use a Bonferroni-correction to adjustment for multiple testing.

Here, we present an extended SMT (ESMT), an improved version of the original SMT. By implementation of a generalized regression equation, the ESMT avoids multiple testing and provides more sophisticated modeling of amplifying interaction. In a simulation study, we show that the ESMT uniformly has higher power than its predecessor. In addition, computing time is improved substantially. The ESMT is implemented in the interaction analysis tool INTERSNP.

A39 - Poster P12

Shannon equivocation for forensic Y-STR marker selection

Siegert S [a], Roewer L [b], Nothnagel M [a]

[a] Department of Statistical Genetics and Bioinformatics, Cologne Center for Genomics, University of Cologne, 50931 Cologne, Germany

[b] Department Forensic Genetics, Institute of Legal Medicine and Forensic Sciences, Charité-Universitätsmedizin Berlin, 13353 Berlin, Germany

Short tandem repeat (STR) markers are widely and continuously used in forensic applications. However, past research has demonstrated substantial allelic correlation between STR markers on both autosomes and the X chromosome, leading to partially redundant information that these markers can provide. Here, we quantify the allelic correlation between Y-chromosomal STR markers that are part of established forensic panels, separately for three different continental groups. We further propose a sequential marker selection procedure that is based on Shannon's equivocation and that accounts for allelic correlation between STR markers, leading to a maximal gain in independent information. In application to three real-world data sets, we demonstrate the procedure's superior performance when compared to single-locus diversity selection strategies, resulting in the optimal marker set for a given data set in the majority of marker subsets. Noting the inferior performance of the established Y-STR marker panels in a retrospective investigation, we suggest that future forensic marker selection should be guided, besides by other technical selection criteria, by an equivocation-based approach to obtain maximally discriminatory marker sets at minimal cost.

A40 - Poster P13

Fine-scale Genetic Structure in Western France

Simonet F [a,b,c], Karakachoff M [a,b,c,d], Persyn E [a,b,c], Violleau J [a,b,c,e], Gros F [a,b,c,e], Schott J-J [a,b,c,e], Redon R [a,b,c,e], Dina C [a,b,c,e]

[a] Institut du Thorax, Nantes, France

[b] CNRS, UMR 6291, Nantes, France

[c] Université de Nantes, Nantes, France

[d] Institute of Clinical Physiology, National Research Council, Pisa, Italy

[e] CHU Nantes, Institut du thorax, Service de Cardiologie, Nantes, France

The Common Variant Common Disease hypothesis was only partly verified through effective discovery of statistical associations. This empirical observation led the research community to reconsider the importance of rare variants. Rare alleles of recent origin are likely to cluster geographically in communities with limited migration rates, like French rural populations before 20th century. Where these rare alleles strongly increase the risk of disease, we'd expect outbreaks of disease prevalence and enhanced power to establish the variant-phenotype relationship. Recently, we demonstrated that genetic structure existed at the level of Brittany and Anjou (Karakachoff et al., 2014). The present work aims to assess whether the same demographical history lends itself to a finer-scale population structure.

We genotyped (Affymetrix Axiom chip) 190 individuals which have their four grand-parents born within a distance of less than 15 kilometers. The grand-parents'birth places are restricted in a small region spanning 200 kilometers long and 200 kilometers large in Western France.

We first performed a multi-dimensional scaling analysis based on the average Identity by State matrix generated by the Plink software.

Two principal components revealed important correlation with latitude ($p<1e-06$). Correlation with longitude is also observed ($p=0.01$).

These results show that population stratification can be observed even at a very fine geographical level in populations usually considered as not isolated. This suggests that recent rare alleles are likely to cluster geographically even in these populations. Therefore, identification of rare variants inducing disease susceptibility can benefit from a strategy focusing on small geographic units.

Applying spatially explicit methods like spatial-PCA (Jombard et al., 2009) allowed identification of potential genetic clusters at the level of groups of villages.

Of note, part of the correlation between genetics and geography is possibly attributable the Loire river acting as a barrier.

96 additional individuals born in the neighboring Morbihan, using the same ascertainment, are being genotyped and will allow an estimation of genetic clines and clusters along the Atlantic coast.

A41 - Poster P14

A comparison of Cox and logistic regression for use in genome-wide association studies

Staley J [a], Jones E [a], Kaptoge S [a], Sweeting M [a], Wood A [a], Howson J [a]

[a] University of Cambridge, UK

In contrast to traditional epidemiology, logistic regression is often used instead of Cox regression to analyse genome-wide association studies (GWAS) of single nucleotide polymorphisms (SNPs) and disease outcomes with cohort and case-cohort designs. GWAS typically identify SNPs with small disease associations using thousands of samples. In this context, the original work evaluating logistic regression in cohort studies may not be generalisable. The performance of logistic regression in case-cohort studies has not been previously assessed. Here, we evaluated Cox and logistic regression applied to cohort and case-cohort genetic association studies using simulated data and SNP data from the cardiovascular disease (CVD) component of the European Prospective Investigation into Cancer and Nutrition study (EPIC-CVD). In the cohort setting, the improvement in power to detect SNP-disease associations using Cox regression compared to logistic regression tended to be modest (<2%), but increased as the disease incidence increased. Whereas, logistic regression had more power than Cox regression in the case-cohort setting. Logistic regression produced biased estimates of effect (assuming that the hazard ratio is the underlying measure of association) for both study designs, especially for SNPs with greater effect on disease. However, logistic regression is substantially more computationally efficient than Cox regression in both settings. We therefore propose a two-step approach to GWAS in cohort and case-cohort studies. Firstly, to analyse all SNPs with logistic regression to identify associated variants below a pre-defined p-value threshold and secondly, to fit Cox regression to those identified SNPs to avoid bias in their estimated association with disease.

A42 - Poster P15

The kernel-based score test for functional linear models in family-based samples

Svishcheva G [a,b]*, Belonogova N* [a]*, Axenovich T* [a,c]

[a] ICG SB RAN, Novosibirsk, Russia

[b] VIGG RAN, Moscow Russia

[c] NSU, Novosibirsk, Russia

Within the framework of functional data analysis approach, we developed new kernel-based method to test associations between quantitative traits and genetic (rare or/and common) variants in a genome region using samples of relatives. Here, the regional genome of an individual is considered as a continuous stochastic function that contains information about both linkage and linkage disequilibrium between the genetic variants. The new kernel-based method is built on functional linear models with random effects of interest and uses the variance-component score statistic to test the null hypothesis of no associations. This method takes into account covariates, relationships of individuals and locations of variants in region. We compare statistical properties of the new method and two methods recently developed. The first method (FFBSKAT) tests the multiple linear regression model using the variance-component score statistic, and the second one (famFLM) tests the functional linear model using the F-statistic. The comparative analysis is carried out on Genetic Analysis Workshop 17 mini-exome family data and a wide range of simulation scenarios where the causal variants are (1) all rare, (2) both rare and common, and (3) all common.

A43 - Session 7 - Friday 17 April 14:20

Powerful methodology for the analysis of "time-to-event" data in pharmacogenetic studies

Syed H [a], Jorgensen A [a], Morris A [a,b]

[a] Department of Biostatistics, University of Liverpool, Liverpool, UK

[b] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Methodology for the analysis of traditional genome-wide association studies has focused on binary (case/control) phenotypes and quantitative traits. However, in pharmacogenetic studies, the outcome of interest is often "time-to-event" data, for example time to death, or remission, or the occurrence of an adverse event, after treatment intervention. One approach is to dichotomise these survival times at some fixed time-point, and to treat the outcome as binary, but this would be expected to result in a loss of power to detect association with genetic variants. We have undertaken a simulation study to compare the power of alternative methods for the analysis of time-to-event data in genetic association studies, including a Cox proportional hazards model and a logistic regression model of a survival indicator variable at the conclusion of the study. We considered a range of scenarios including variable censoring and SNP-treatment interaction effects. For each scenario, we generated 1000 replicates of genotype data and survival times for a sample of 1000 individuals. All simulations and analyses were performed in R 3.1.2. The Cox proportional hazards model was demonstrated to be uniformly more powerful than logistic regression analysis of dichotomised survival. However, the difference in power between the two approaches was highly dependent on the rate of censoring. These findings have important implications for the development of analytical protocols in the analysis of time-to-event data in pharmacogenetic studies. The simulation machinery has been implemented in user friendly software to allow for power calculations for time-to-event outcomes in pharmacogenetic studies, allowing for variable censoring options and inclusion of SNP-treatment interaction effects.

A44 - Session 1 - Thursday 16 April 10:30

Secondary phenotype analysis in ascertained family designs: Application to the Leiden Longevity Study

Tissier R [a], Tsonaka R [a], Houwing-Duistermaat J [a]

[a] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

Genome-wide association studies (GWAS) typically adopt a case-control design to test associations between the case-control status and genetic variants. In addition to this primary phenotype a number of additional traits, also known as secondary phenotypes, are routinely recorded such as lipidomics, glycomics, metabolomics and clinical phenotypes. Studying associations between genetic variants and these secondary traits is of great interest. However, when analyzing secondary phenotypes in case-control studies failure to properly adjust for the sampling design may lead to biased genetic effect estimates, especially when the marker tested is associated with the primary phenotypes and when the primary and the secondary phenotypes tested are correlated. Several methods have been proposed in the literature to study the association between the secondary phenotype and genetic markers but they are limited to case-control studies and not directly applicable to more complex designs, such as the multiple-cases family studies. A proper secondary phenotype analysis in this case is complicated by the often complex sampling design at the family level, the within-family correlations and the mixed type of outcomes recorded i.e. binary primary phenotype and continuous or categorical secondary phenotypes.

We propose a novel approach to accommodate the ascertainment process while explicitly modelling the familial relationships. Our approach pairs existing methods for mixed-effects models with the retrospective likelihood framework which corrects for ascertainment of the families and uses a multivariate probit model to capture the association between the mixed type primary and secondary phenotypes. To examine the efficiency and bias of the estimates we performed simulations under several scenarios for the association between the primary phenotype, secondary phenotype and genetic markers. Simulations showed that not taking into account the sampling mechanism can lead to biased inferences, false positive association and severe underestimation of the heritability estimates for secondary phenotypes. We will illustrate the method by analyzing secondary phenotypes (triglycerides and glucose) and genetic markers from the Leiden Longevity study, a multiple-cases family study that investigates healthy ageing.

A45 - Poster P16

Meta-analysis of multiple regression models in Genome-wide association studies

Vaitsiakhovich T [a], *Becker T* [a,b]

[a] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[b] Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany

Meta-analysis refers to the statistical synthesis of the results from a series of related studies. Meta-analysis of summary statistics from Genome-wide association studies (GWAS) may lead to discovery of new susceptibility loci without the need to exchange genotype data.

Well-known p-value combination methods can be applied to an arbitrary association test. However, these methods do not provide summary effects and are known to be underpowered for high-dimensional models. The broadly used methods for meta-analysis of individual GWAS results address mostly one parameter models. Results of multiple regression analysis applied, for instance, to genome-wide interaction analysis have rarely been used due to the complexities underlying the process of their synthesis.

For high-dimensional models the gain of power can be achieved when meta-analysis methods incorporate information on study-specific properties of parameter estimates, their effect directions, standard errors and covariance structure. In this context, a method for the synthesis of regression slopes (MSRS) represents a perspective approach of meta-analysis for advanced GWAS. MSRS provides meta-analysis p-values and common parameter estimates of multiple regression models for an arbitrary number of parameters, and can be used to test the homogeneity of studies results.

We introduce an efficient, powerful and freely available software tool METAINTER, which implements MSRS and three further meta-analysis methods: Fisher's method, Stouffer's method with weights and Stouffer's method with weights and effect directions. METAINTER enables meta-analysis of tests for and under gene-gene interaction, single-SNP association tests with several degrees of freedom, global haplotype tests etc. Simulation study shows that MSRS has correct type I error and its power comes close to that of the joint sample analysis. We have conducted a real data analysis of six GWAS of type 2 Diabetes. For each study, a genome-wide interaction analysis of all SNP pairs has been performed by logistic regression tests. The results have then been meta-analysed with METAINTER.

A46 - Poster P17

Utilising machine-learning algorithms to uncover complex genetic interactions in schizophrenia

Vivian-Griffiths T [a], Escott-Price V [a], Walters J [a], Moran J [b], McCarroll S [c], O'Donovan M [a], Owen M [a], Pocklington A [a]

[a] Institute of Psychological Medicine and Clinical Neurosciences - Cardiff University

[b] Stanley Center for Psychiatric Research - Broad Institute

[c] Harvard Medical School

Studies have shown that there are potentially thousands of common genetic variations implicated in schizophrenia, all contributing a small effect to the disorder (Sullivan et al., 2003; Ripke et al., 2013). These variants have been used in predictive models by calculating a polygenic-score (PS) for each individual. PS is a weighted average of the minor allele counts at the genetic loci, weighted by the log-odds-ratio of the respective loci and alleles, calculated from a Genome Wide Association Study (GWAS). This score can then be used in a logistic regression to predict the case/control status in an independent sample. This is a linear combination of the genetic variants which does not take into account complex interactions between them. To include these explicitly into a regression model is not feasible due to such an enormous number of possible combinations of n-way interactions, and the concomitant problem of correcting for the results for multiple comparisons.

Here we investigate the ability of Support Vector Machines (SVMs) algorithms to discriminate between schizophrenia cases and controls using GWAS data. SVMs can account for interactions in the data via the use of Kernel functions, which increase the number of dimensions of the predictors. We investigated the Polynomial Kernel function, which can calculate n-way interactions based on the degree (n) of the polynomial and Radial-Basis-Function Kernels, which are capable of increasing the number of dimensions infinitely and thus include all possible interactions.

We used the datasets drawn from the CLOZUK study (Hamshere et al., 2014), which was also included in the Psychiatric Genetics Consortium GWAS (Ripke et al., 2013). The first set consisted of 125 weighted allele counts defined by the index genome-wide significant SNPs, The second dataset used all relatively ($r2<0.2$) independent variants which were also significant at the $p<.05$ level ($n=31,166$). We tested whether the predictive model based upon these individual scores predicts the case/control status better than the PS and compare its performance with the logistic regression. The K-fold cross validation procedure was employed to validate the predictive model built with SVMs.

The initial findings have shown that the performance of the SVM algorithms did not improve the prediction as compared to logistic regression analysis. Future work will include the use of decision tree and random forest algorithms and performance assessment of all these methods.

A47 - Session 5 - Friday 17 April 10:20

MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies

Vuckovic D [a], Gandin I [a], Nicola P [a], Soranzo N [b,c], Iotchkova V [b,d]

[a] Department of Medical, Surgical and Health Sciences, University of Trieste, Italy

[b] Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, UK

[c] Department of Haematology, University of Cambridge, UK

[d] European Bioinformatics Institute, Wellcome Trust Genome Campus, UK

While typically Genome Wide Association Studies (GWAS) test one phenotype at a time, many biological features are better described by a combination of several variables. Thus addressing multiple phenotypes can give great increase in power, by taking into account the underlying correlations between variables. Multivariate regression is a straightforward generalization of standard GWAS. In particular a linear multivariate mixed model is suitable to control for population stratification and relatedness and thus has been recently implemented in MTMM software (Korte et al., 2012) and GEMMA (Zhou & Stephens, 2014).

Here we describe a novel statistically efficient method to perform meta-analysis in a multivariate setting. It is an inverse-variance based method that allows different weights for each cohort in order to take into account the accuracy of each effect estimate, similar to the one implemented in METAL for univariate GWAS (Willer et al., 2010). It is part of the R package MultiMeta together with plotting functions for results visualization, already available on CRAN.

In order to test our software, we analyzed UK10k data (2 cohorts, total N~3000 samples) on 6 lipid traits: apolipoprotein A, apolipoprotein B, High-Density Lipids, Low-Density Lipids, Triglycerides and Total Cholesterol. Despite limited sample size, 6 previously reported associations were confirmed with highly significant p-value. Interestingly some of these were not detected with univariate GWAS meta-analysis on the same sample, although the reverse was also true: multivariate analysis did not detect all of the loci associated with univariate ones. Finally two possibly new loci reached genome-wide significance on chromosome 6, for which we are currently searching for replication in independent cohorts.

In conclusion, combining results from different cohorts is particularly important for GWAS, where large sample sizes are required to reliably detect alleles with small effects. The R package MultiMeta provides a flexible approach to meta-analyzing multivariate GWAS and easily visualizing results.

A48 - Session 8 - Friday 17 April 16:40

A Multi-Marker Genetic Association Test Based on the Rasch Model Provides New Insights into Genetics of Alzheimer's Disease

Wang W [a,b], Commenges D [b], Guedj M [a]

[a] Pharnext, France

[b] Université de Bordeaux, France

Results from Genome-Wide Association Studies (GWAS) have shown that the genetic basis of complex traits often include many genetic variants with small to moderate effects whose identification remains a challenging problem. In this context multi-marker analysis at the gene and pathway level can complement traditional point-wise approaches that treat the genetic markers individually.

In this paper we propose a novel statistical approach for multi-marker analysis based on the Rasch model. The method summarizes the categorical genotypes of SNPs by a generalized logistic function into a genetic score that can be used for association analysis.

Through different sets of simulations, the false-positive rate and power of the proposed approach are compared to a set of existing methods, and shows good performances. The application of the Rasch model to the Alzheimer's disease (AD) ADNI GWAS data allows a coherent interpretation of the results. Our analysis supports the idea that APOE is a major susceptibility gene for AD. Most of the other top genes could also be functionally linked to pathological processes associated with AD. In particular, a pathway analysis of these genes also highlights the metabolism of cholesterol, known to play a key role in AD pathogenesis. Interestingly, these findings can be integrated in a hypothetic signalling network.

A49 - Poster P18

Pharmacogenetic Meta-Analysis of Response to Antihypertensive Drugs within ASCOT

Warren H [a,b], Sever P [c], Poulter N [c], Stanton A [d], Caulfield M [a,b], Munroe P [a,b]

[a] Department of Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK

[b] NIHR Barts Cardiovascular Biomedical Research Unit, Queen Mary University of London, London, UK

[c] International Centre for Circulatory Health, Imperial College, London UK

[d] Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin, Ireland

Hypertension currently affects approximately one third of adults worldwide, contributing to half of all cardiovascular deaths annually. Antihypertensive drugs are widely prescribed, however there is significant inter-individual variation in treatment response, which is believed to be partly due to genetic variation. By identifying genetic variants associated with a differential response to antihypertensives, it is hoped that pharmacogenetics may aid in the selection of optimal treatments.

So far, only three genetic loci have been validated from pharmacogenetic genome-wide association studies (GWAS), only for BP response to diuretics. It has been hypothesised that the loci associated with BP would also modify the effects of response to antihypertensives. However the most recent study considered the then 39 BP-associated single-nucleotide polymorphisms (SNPs) and none attained significant pharmacogenetic associations.

We have derived novel methods to take advantage of data from the Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT). A total of 19,342 European hypertensive patients were randomized to either a beta-blocker (BB) or a calcium-channel-blocker (CCB). Genetic data is available for 3,804 subjects from UK/Ireland (ASCOT-UK) and 2,468 from Scandinavia (ASCOT-SC). A pharmacogenetic GWAS has been performed within each dataset, followed by a combined meta-analysis.

Subjects were restricted to those on monotherapy treatment, comparing BB vs. CCB patients within a linear regression drug-gene interaction analysis. Three BP-related phenotypes were considered: systolic BP (SBP), diastolic BP (DBP) and heart rate (HR). The response was defined as the mean phenotype measure over all longitudinal visits whilst on monotherapy, and was adjusted for several covariates, including the baseline phenotype measure, to eliminate the confounded effect of association between SNP and baseline BP. Any subjects taking BB (or CCB) drugs at baseline were excluded, resulting in a total sample size of up to 1,659 UK and 961 Scandinavian subjects.

We have identified one novel genetic locus on chromosome 11, reaching genome-wide significance for HR response. Furthermore we performed an updated look-up for 79 SNPs published for BP and identified two SNPs with significant pharmacogenetic association after Bonferroni correction: one for HR response on chromosome 11; one for SBP response on chromosome 4, each of which has potential relevance and functional support. The latter variant has been previously associated within a BP risk score for pharmacogenetic response, but this is the first time either SNP has been identified individually.

Due to the novel methodology we will also discuss many methodological challenges and modelling options surrounding pharmacogenetic analyses.

A50 - Poster P19

Genetic testing for autism spectrum disorder is lacking evidence of cost-effectiveness: Systematic Review

Ziegler A [a,b,c], Vonthein R [a,b], Rudolph-Rothfeld W [a]

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

[b] Zentrum für Klinische Studien Lübeck, Universität zu Lübeck, Lübeck, Germany

[c] Deutsches Zentrum für Herz-Kreislauf Forschung, Lübeck, Germany

Background: Autism Spectrum Disorder (ASD) is a highly heritable neural development disorder characterized by social impairment. The earlier the diagnosis is made, the higher are the chances of obtaining relief of symptoms. A very early diagnosis uses molecular genetic tests, which are offered by commercial companies and universities.

Objective: Evaluation of the economic impact of genetic tests in ASD.

Data sources: We performed a systematic search of databases Pubmed, Medline, Cochrane (DARE, EED), Econlit, NHS Center for Reviews and Dissemination (i.e., York, DARE, NHSEED, HTA) for articles in English and German from 2000 to 2013.

Study selection: Original articles published in peer-reviewed journals and describing economic evaluations of genetic tests for ASD.

Results: We found 8 articles on economic evaluations (EE) and 13 articles on diagnostic tests for ASD. We identified 149 economic evaluations of genetic tests for various diseases. However, not a single economic evaluation of genetic tests has been found for ASD.

Conclusion: There is no evidence for cost-effectiveness (CE) of any genetic diagnostic test for ASD, although they are available commercially.