Human
Heredity

Hum Hered 2013;76:86–113
DOI:10.1159/000360587

Published online: March 18, 2014

42nd European Mathematical Genetics Meeting (EMGM) 2014

April 1–2, 2014, Cologne, Germany

Abstracts

Guest Editors
Michael Nothnagel
Sabine Siegert

KARGER Basel · Freiburg · Paris · London · New York · Chennai · New Delhi ·
Bangkok · Beijing · Shanghai · Tokyo · Kuala Lumpur · Singapore · Sydney

Human Heredity

Abstracts

Oral Presentations

Keynote Lecture

Current Concepts in Genetic Epidemiology: A Critical Appraisal

Andreas Ziegler[1,2]

[1]Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany; [2]Center for Clinical Trials, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
Email: ziegler@imbs.uni-luebeck.de

Genetic epidemiology emerged as a marriage between the disciplines genetics and epidemiology, and it was primarily driven by technological developments in molecular genetics in the past three decades. In the first part of my presentation, I will provide a brief history of the definitions of genetic epidemiology. I will critically discuss whether these definitions are timely, and I will provide a modified definition of the term genetic epidemiology to cover modern needs. The aim of most genetic epidemiological studies is to identify an association between a disease and a chromosomal region. In population genetics one aim is to identify population stratification, and I will argue in the second part that the study of religious, cultural, and political differences between groups can be helpful to justify differences on the genetic level. The third part of the presentation will focus on study designs, and risk prediction models and Mendelian randomization studies will serve as examples. Specifically, I will review the assumptions which need to be fulfilled for drawing valid conclusions. In conclusion, the rigorous application of statistical design methodology is required in modern genetic epidemiological studies.

Invited Lecture

Kinship, Heritability, and Prediction

David Balding

Genetics Institute, University College London, London, UK
Email: d.balding@ucl.ac.uk

Measuring the extent to which the genetic similarity of pairs of individuals can explain their phenotypic similarity has been at the heart of quantitative genetics for over a century. Until recently there was in effect only one definition of kinship, based on expected genome sharing computed from known pedigrees. But pedigrees only specify expected genome sharing, whereas nowadays genome-wide SNPs allow genetic similarity between individuals to be measured directly. There are many SNP-based kinship measures available, allowing much greater flexibility for example in heritability analyses. At the same time some confusion has arisen as traditional notions of kinship and heritability cease to apply. The new methods allow the assessment of heritability of genomic regions and classes functional sites, providing new tools for dissecting the genomic architecture of complex traits. It also opens the way to new approaches to the prediction of phenotype from genome-wide SNPs. I review these ideas and illustrate them with analyses of several datasets. This is joint work with postdoc Doug Speed, who is funded by the UK Medical Research Council.

O-1

Testing the Spatial Correlation of Association Signals from Two Genome Scans

Julian Hecker[1], *Dmitry Prokopenko*[1], *Pedro Costa*[1], *Edwin Silverman*[2], *Melissa Naylor*[3], *Scott Weiss*[2], *Christoph Lange*[1,2,4,5], *Heide Loehlein Fier*[1]

[1]Institute of Genomic Mathematics, University of Bonn, Germany; [2]Channing Laboratory, Brigham and Women's Hospital, Boston, USA; [3]Pritzker School of Medicine, University of Chicago, USA; [4]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany; [5]Department of Biostatistics, Harvard School of Public Health, Boston, USA
Email: hecker.julian@gmail.com

It is widely recognized that complex diseases like i.e. Alzheimer's disease, Cardio-Vascular-Diseases, Diabetes, Asthma, psychiatric diseases, and many others are characterized by a interaction of genetic, environmental and individual factors (e.g. Hunter 2005, Manolio et al., 2006; Risch et al., 2009; Manolio et al., 2008; Goldstein 2009; Thomas 2010).

As a result usually multiple correlated traits exist that describe the syndromes of a complex disease. Pinpointing and quantifying genetic risk factors of a complex disease based on several correlated traits however implies a challenge, since different traits might be differently associated with the genetic profile of the analyzed subjects (e.g. Wang 2012, Yang and Wang 2012). In order to understand the genetic architecture of complex diseases it is therefore crucial to unravel the genetic dependencies between correlated disease traits (pleiotropic effects).

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/13/0762–0086$39.50/0

In this communication, we propose a method to estimate the spatial correlation of two GWAS association signals. For this, we use the large sample properties of test statistics and the LD information about the genomic region. Based on this knowledge, we can estimate the variance components of a linear mixed model. From this, we can compute the spatial correlation and asymptotic variance. We conduct a simulation study to assess how well our method is able to detect spatial correlation of GWAS signals and apply our method to real data.

O-2

A Comparison of Multivariate GWAS Methods

Heather Porter, Cathryn Lewis, Paul O'Reilly

MRC SGDP Centre, Institute of Psychiatry, King's College London, UK
Email: heather.porter@kcl.ac.uk

Genome-wide association studies (GWAS) traditionally adopt a univariate approach, focusing on a single phenotype of interest. In recent years numerous multivariate methods have been proposed; these model multiple phenotypes simultaneously to investigate their joint association with single-nucleotide polymorphisms (SNPs). These methods allow shared genetic architecture and correlation between phenotypes to be exploited, increasing the power to detect true genotype-phenotype associations. So far there has been limited method comparison, with no clear overall preference or guidance on when different multivariate methods should be implemented. We undertake a simulation study to conduct a thorough comparison of multivariate methods that test multiple phenotypes SNP-by-SNP, including TATES, PCA and MultiPhen. We compare their statistical power across a range of phenotype correlations and genetic effect sizes, and model direct and indirect relationships between genotypes and phenotypes. Our simulation framework has three strands. First, we employ a structured approach to modelling phenotypic correlations and genetic effect sizes, such as setting all to be equal or fixing one to be zero. Next we perform more general simulations, where effect sizes and phenotype correlations are sampled from the uniform distribution. Finally, we exploit real genotype and phenotype data to simulate realistic genetic effects and phenotypic correlations. We demonstrate why each method performs well in certain scenarios, and detail the power, computational speed and interpretability of the methods to aid user choice. We believe that this study provides the most comprehensive guide to the performance of multivariate GWAS methods, and their general utility, so far.

O-3

Detecting Genetic Interactions in Case-Control Data: Information Theory Based Methods

Tomasz Ignac[1], Alexander Skupin[1], Nikita Sakhanenko[2], David Galas[2]

[1]Luxembourg Centre for Systems Biomedicine, Luxembourg; [2]Pacific Northwest Diabetes Research Institute, USA
Email: tomasz.ignac@uni.lu

Phenotypic variations, including those which underlay health and disease in humans, result from multiple interactions among both genetic variations and environmental factors. While diseases caused by single gene variants can be identified by established association methods and family-based approaches, complex phenotypic traits resulting from multi-gene interactions remain very difficult to characterize. Recently, new information theory based methods that are aimed specifically at detecting complex, non-additive interactions have been proposed.

Here we present a set of tools based on information theory. We start from discussing naïve straightforward applications of conditional entropy and mutual information, and demonstrate that they become inefficient when applied to large data sets. Then, we present the Interaction Distance (ID) – a newly developed measure of information among three variables. It combines two concepts: the interaction information, which is a generalization of the mutual information to three variables, and the normalized information distance which measures informational sharing between two variables. The specific design of ID allows for a computationally efficient detection of pairwise effects on phenotype of genetic markers that have negligible effect while considered alone. We demonstrate how ID improves on previous approaches to identify genetic interactions. While ID aims at detecting pairwise interactions, it is now hypothesised that complex diseases are caused by complex and non-linear multi-gene interactions that determine phenotypic traits. Based on our work, we argue that information theory related methods could be extended to these cases.

O-4

Comparison of Variable Selection Methods in Random Forests for Genomic Data Sets

Silke Szymczak[1], James Malley[2], Andre Franke[1]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany; [2]Center for Information Technology, National Institutes of Health, USA
Email: s.szymczak@ikmb.uni-kiel.de

Random forests (RFs) are a promising approach for classification based on high dimensional genomic data sets. Variable importance measures allow variables to be ranked according to their predictive power.

We evaluated several variable selection procedures based on a breast cancer gene expression data set. RFs were used to estimate the probability of positive estrogen receptor status. Our comparison included a permutation-based approach (PERM), recursive feature elimination (RFE) and our new method based on recurrent relative variable importance measurements (r2VIM). The data set was repeatedly split into training and test sets and comparisons were based on the number of selected genes and mean squared error (MSE) of a RF built on selected genes only. We also permuted the phenotype to generate data sets under the null hypothesis.

MSE were similar but the median number of selected genes ranged from 33 for r2VIM to 183 for PERM. RFE was the only approach that used more genes under the null hypothesis compared to the original data. PERM selected a median of 145 genes whereas 50% of the splits resulted in less than four genes using r2VIM. Five genes were selected in more than 80% of the splits in each of the approaches. These genes are known to be associated with estrogen receptor status in breast cancer. Run time was comparable for RFE and r2VIM whereas PERM was 20 times slower.

In conclusion r2VIM is a sensible choice for variable selection in RF for application to high dimensional data sets.

O-5

A Fast Method Using Polygenic Scores to Estimate the Number of SNPs Contributing to a Trait and the Variance Explained by Genomewide SNP Panels

Luigi Palla, Frank Dudbridge

Department of Non-Communicable Disease Epidemiology,
London School of Hygiene and Tropical Medicine, UK
Email: luigi.palla@lshtm.ac.uk

Recent research has addressed the issue of estimating the variance explained by the entire set of SNPs in a genomewide panel, with the awareness that the number of SNPs that are individually significantly associated is limited by the sample size. In particular, a method based on polygenic scoring was proposed by Stahl et al (Nat Genet 2012) to estimate both variance explained and number of SNPs affecting the trait. Their approach uses whole genome simulations together with rejection sampling to obtain Bayesian estimates of the genetic model parameters.

We propose a fast analytic method that is based on the formula for the noncentrality parameter of the test of association of a polygenic score with the trait of interest (Dudbridge, PLoS Genet 2013). This can be expressed as a function of genetic model parameters and compared to the observed test statistic in a dataset. We show how model parameters can be estimated from the results of multiple polygenic score tests based on SNPs with P-values falling in different intervals. We give a fast method for constructing approximate confidence intervals for the estimated parameters, and a more exact method that requires whole genome simulations. We show that our methods give nearly unbiased estimates of the variance explained and number of SNPs affecting a trait, with appropriate coverage of the confidence intervals.

We compared various approaches for constructing polygenic scores for this approach, and found that estimates were more accurate when SNPs were selected into the polygenic scores according to disjoint intervals of P-values, rather than nested intervals. Furthermore we found that unweighted scores yielded more accurate estimates than scores in which SNPs were weighted by their effect sizes.

O-6

Application of GREML and Genomic Profile Risk Scoring to Study the Presence of Shared Risk Alleles Between Major Depression and Rheumatoid Arthritis

Jack Euesden, The RADIANT Depression Consortium, Cathryn Lewis

Institute of Psychiatry, King's College London, UK
Email: jack.euesden@kcl.ac.uk

Introduction: As well as association studies within a phenotype, genome-wide data can be used for investigating the presence of shared risk alleles across disorders. Major Depression (MDD) and Rheumatoid Arthritis (RA) co-occur in the same individuals more often than chance. We are interested in testing whether a common genetic diathesis may underlie these two heritable, complex disorders.

Methods: Using Genomic Profile Risk Scoring (GPRS) and Genomic-relatedness-matrix residual maximum likelihood (GREML), we assessed pooled effects of SNPs that, individually, capture modest effects and do not reach genome-wide significance. GPRS assumes that, when studying complex traits, the stringent genome-wide significance threshold commonly used can lead to the discarding of causal variants. Consequently we relax the alpha threshold progressively until we are modelling disease status using an optimal number of false-negative SNPs selected based on p-value. GREML, meanwhile, constructs kinship matrices using randomly selected genome-wide SNPs in order to fit linear mixed models for disease liability based on genetic similarity. We applied both methods to MDD and RA data from the RADIANT consortium and WTCCC respectively.

Results and Conclusions: In GPRS, minimum p-value for MDD predicting RA was 0.17, and in RA predicting MDD was 0.42. GREML estimate of bivariate heritability was non-significant (0.29, 95% CI: -0.05 – 0.63). We find no evidence for shared genetic risk between RA and MDD. This is validated using GREML and GPRS, methods that have converging estimates and whose assumptions we review.

O-7

Genomic Prediction for Complex Traits Following Feature Selection: Results from Bayes C and Genomic Best Linear Unbiased Prediction (G-BLUP)

Mairead Bermingham[1], Ricardo Pong-Wong[2], Athina Spiliopoulou[1], Caroline Hayward[1], Igor Rudan[3], Harry Campbell[3], Alan Wright[1], James Wilson[3], Felix Agakov[4], Pau Navarro[1], Chris Haley[1]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, UK; [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK; [3]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK; [4]Pharmatics Limited, UK
Email: mairead.bermingham@igmm.ed.ac.uk

Genomic prediction for complex traits following feature selection: results from Bayes C and genomic best linear unbiased prediction (G-BLUP).

Genome-wide association studies (GWAS) have identified thousands of SNPs associated with health-related traits, and thus provide a source of information about useful predictors for these traits. The best practices in the implementation of genomic prediction approaches using these high-dimensional GWAS data have yet to be determined. One important issue is feature selection (i.e. selection of SNPs exhibiting non-redundant information) which could reduce model complexity and computational requirements. In this study we investigated the effect of supervised feature selection on the performance of two widely used prediction methods: Bayes C and genomic best linear unbiased prediction (G-BLUP). We explored prediction of the complex traits height, high density lipoproteins (HDL) and body mass index (BMI) within 2,186 Croatian and into a replication population of 810 UK individuals (ORCADES). Using all 263,357 markers, Bayes C and G-BLUP had similar prediction accuracy across all traits within the Croatian data, and for the highly polygenic traits height and BMI when predicting into the ORCADES data. Although Bayes C outperformed G-BLUP in the prediction of HDL (which is influenced by fewer quantitative trait loci than BMI and height) into the ORCADES data, it was more than 3,000 times slower computationally than G-BLUP. However, the application of supervised feature selection allowed G-BLUP to achieve equivalent predictive performance to Bayes C with greatly reduced computational effort. Feature selection in the G-BLUP framework therefore provides a flexible and more efficient alternative to computationally expensive Bayes C for all considered traits in this study.

O-8

MultiBLUP: Improved Prediction for Complex Traits

Doug Speed, David Balding

Genetics Institute, University College London, UK
Email: doug.speed@ucl.ac.uk

The high heritability of many diseases means there is great potential to improve patient diagnosis and prediction of risk by incorporating genetic markers. While the low prevalence of most diseases makes prediction difficult on a population level, in many cases it is possible to identify subgroups of patients for whom prevalence is much higher, and therefore where prediction models will offer far more benefit: for example, the prevalence of Type 2 diabetes among obese individuals is almost 50%, while the chance of developing epilepsy following a single seizure is about a third (much higher than the population prevalences of 5% and 0.5%).

Although the success of a prediction model is limited by sample size, with meta-analysis consortia established for many major diseases, sample size is no longer the primary obstacle. Instead, we are let down by our tools for constructing prediction models. For example, prediction typically relies on either genetic risk scores or BLUP (Best Linear Unbiased Prediction), which although computationally fast, are poorly suited for complex genetic architectures. While many more sophisticated tools exist, such as Lasso, HyperLasso and BSLMM (Bayesian Sparse Linear Mixed Models), the computational demands of these methods generally prevent them being applied to genome-wide SNP data for thousands of individuals.

As an alternative, we have developed MultiBLUP, which adapts the BLUP model to make it more suitable for complex traits. It does this by increasing the number of random effect terms, to appreciate that complex traits are likely to have many susceptibility loci with varying influence on risk. We apply MultiBLUP to the seven WTCCC traits, as well as Celiac Disease and breast cancer, consistently achieving better risk prediction than rival prediction methods. Moreover, because MultiBLUP retains much of the computational efficiency of BLUP, it is many times faster than the next best method, BSLMM. For example, we can process genome-wide imputed SNP data for thousands of individuals in under an hour. MultiBLUP can also be used for predicting continuous traits (we demonstrate this for cholesterol), and for non-human data (we consider 139 mice phenotypes). The MultiBLUP software is freely available at www.ldak.org.

O-9

Performance of Clustering Using a Reference Panel for Fine-Scale Population Identification in a French Population, The 3 City Study

Aude Saint Pierre[1,2,3], Céline Bellenguez[4,5,6], Emmanuelle Génin[1,2,3]

[1]Inserm UMR1078 'Génétique, Génomique fonctionnelle et Biotechnologies', Brest, France; [2]Université de Bretagne Occidentale, Brest, France; [3]Centre Hospitalier Régional Universitaire de Brest, France; [4]Inserm, U744, Lille, 59000, France; [5]Université Lille 2, Lille, 59000, France; [6]Institut Pasteur de Lille, Lille, 59000, France
Email: aude.saint-pierre@inserm.fr

Population stratification is an important cause of false positive results in genome wide association studies (GWASs). With the development of high throughput technologies and the possibility to assess genetic variations at a finer scale, it has been shown that differences in allele frequency exist at all the geographic levels, and even between regions within a country.

Several methods have been developed to account for population stratification in GWASs. Due to his simplicity and his good performance, the most popular is the principal component analysis (PCA). PCA, and related spectral methods, rely on a pairwise similarity measure between all individuals in the sample to generate axes of variation maximizing the genetic variability. These axes of variation depict a summary of the data set that can be used for population identification via clustering.

The choice of the similarity measure might impact the performance of clustering-algorithm for population identification. For fine-scale population stratification, such as the one observed within a country, Lawson et al. (2012) showed that the use of haplotypic information to compute similarities can lead to some significant improvements in the clustering of individual compared to the standard method based on the Identity By State sharing that considers each marker independently. Moreover, the analysis of the stratification in a sample can either be done on the sample only or after merging the sample with some reference panels of individuals of different known population ancestry that can serve as donor population in the FineStructure method developed by Lawson et al. (2012).

In this study, we explore the impact of the choice of the similarity measure and the benefits of using a reference panel to identify clusters among individuals coming from different regions of France. We use the data from the Three City Study, a reference panel of French elderly individuals that served as controls in several GWAS conducted with French patients. This panel includes individuals sampled in three regions of France (around Bordeaux in the South West, Montpellier in the South and Dijon in the East of France). We show how well these three sample locations are recovered with the different methods.

The results of this study would be useful to guide the search of genetic disease risk variants of low frequency that are likely to show finer stratification patterns than more common variants.

O-10

Haplotype Information Combined with Iterative Pruning PCA (ipPCA) to Improve Population Clustering

Kridsadakorn Chaichoompu[1,2]*, Ramouna Fouladi*[1,2]*,*
Pongsakorn Wangkumhang[3]*, Alisa Wilantho*[3]*,*
Wanwisa Chareanchim[3]*, Anavaj Sakuntabhai*[4]*,*
Sissades Tongsima[3]*, Kristel Van Steen*[1]

[1]Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium; [2]Bioinformatics and Modeling, GIGA-R, University of Liege, Belgium; [3]Biostatistics and informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand; [4]Functional Genetics of Infectious Diseases Unit, Institut Pasteur, France
Email: kridsadakorn.chaichoompu@ulg.ac.be

Single Nucleotide Polymorphisms (SNPs) are commonly used to capture variations between populations. Often genome-wide SNP data are pruned based on linkage disequilibrium (LD) patterns or small subsets of SNPs are selected (e.g. PCA-correlated SNPs) to reproduce the genomic structure of the complete data set. Identifying and differentiating between subpopulations using such a reduced set can become challenging, especially when similar geographic regions are involved or when spurious patterns are likely to exist.

Although PCA-based methods can resolve structure, they cannot infer ancestry. On the other hand, the structure of haplotypes in unrelated individuals can reveal useful information about genetic ancestry. Notably, haplotype composition and the pattern of LD between markers may vary between larger populations but may also play a role within more confined geographic regions. In addition, iterative pruning principal component analysis (ipPCA) has been shown to be a powerful tool to cluster subpopulations based on SNP profiles.

Despite the complexities that are associated with haplotype inference, we argue that added value can be obtained when the LD structure between SNPs is exploited in the search for relevant population strata. In this work, we propose to combine an LD-based novel haplotype encoding scheme with the ipPCA machinery to retrieve fine population substructures. The approach is compared to state-of-the-art methods in the context of population substructure and admixture analysis.

O-11

Using Network Methodology to Infer Population Substructure

Dmitry Prokopenko[1]*, Christoph Lange*[1,2,3,4]*, Julian Hecker*[1]*,*
Pedro Costa[1]*, Edwin Silverman*[3]*, Heide Loehlein Fier*[1]

[1]Institute of Genomic Mathematics, University of Bonn, Bonn, Germany; [2]Department of Biostatistics, Harvard School of Public Health, Boston, USA; [3]Channing Laboratory, Brigham and Women's Hospital, Boston, USA; [4]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
Email: dmitry.prokopenko@uni-bonn.de

One of the main caveats of association studies is the possible affection by bias due to population stratification. Existing methods rely on model-based approaches like structure and ADMIXTURE or on principal component analysis like EIGENSTRAT.

Here we describe the problem of population substructure from a graph-theoretical point of view. We group the sequenced individuals into triads, which depict the relational structure, on the basis of a predefined pairwise similarity measure. We then merge the triads into a network and apply community detection algorithms in order to identify homogeneous subgroups or communities, which can further be incorporated as covariates into logistic regression. We apply our method to populations from different continents in the 1000 Genomes Project and evaluate the type 1 error in simulation studies. Our results suggest that the network approach provides a more precise information of population structure than existing methods.

O-12

Modeling Self-Perception of Ancestry in Latin America

Kaustubh Adhikari

University College London, UK
Email: k.adhikari@ucl.ac.uk

'Race' is a contentious word in today's world, and is becoming increasingly redundant today as mixed ethnicities are steadily on the rise. A prime example of this is Latin America with ubiquitous mixing of Native American, European and African ancestries since the time of Columbus. However, social perception of ethnicity is a major factor in these countries which show a noticeable inequality in socioeconomic positions, with European ancestry being privileged in many cases. Of course, external perception of ancestry is often far from the true genetic ancestry. We embarked on the quest to find out to what extent is perceived ancestry is an approximation to the genetic ancestry, and what are the major factors that cause a bias in self-perception. We find out that various human physical characteristics such as the colour of our skin, eyes or hair affect people's perception, which is not unexpected. We also observe that prior knowledge of a person's socioeconomic position has a significant effect on this perception – people with higher educational degrees think of themselves as more European, for example. There is also considerable variation in the attitude towards ethnicities in different countries, from pride to indifference, and that is duly reflected in their self-perception. Combining these observations, we provide a mathematical model of how people's physical characteristics and socioeconomic position shape their self-perception of ancestry over and above their true genetic ancestry in several Latin American Countries.

O-13

Which Hypotheses Does Gene Set Enrichment Analysis (GSEA) Test?

Birgit Debrabant

Epidemiology, Biostatistics and Biodemography, Institute of Public Health, University of Southern Denmark, Denmark
Email: bdebrabant@health.sdu.dk

A popular approach to gene set analysis is the GSEA method going back to Subramanian et al., 2005. A first adaption to GWAS has been put forward by Wang et al., 2007, and by today, many extensions exist. GSEA and related methods test a specific set G of genes, e.g. a biological pathway, for association with the trait of interest in relation to those genes outside of G.

All genes are ranked according to their individual association, and a running sum is calculated based on the ranks of the genes within G relative to those of the genes outside. The maximum is called enrichment score and measures the overrepresentation of genes with low ranks within G relative to its outside. Thereby, a weight parameter p enables to give greater weight to more outstanding genes in G in the calculation. Significance is assessed permuting sample labels.

As one is contrasting genes within G relative to those outside, the null hypothesis was until recently considered to be the competitive hypothesis (The genes in G show the same or lower magnitude of association with the trait as the genes in the complement of G), cp. e.g. Goeman/Bühlmann, 2007. Only 2013, Maciejewski stated that the sample permutation scheme determines the null hypothesis, which therefore amounts to H0: No gene, neither in G nor its complement, is associated with the trait. Concerning the alternative hypothesis, the same work indicates this to be the negation H0. However, the test statistic is originally tailored to competitive alternatives and it remains unclear how much power there is against more general alternatives.

This talk elaborates on the appropriate formulation of null and alternative hypotheses, and presents power simulation results for different types of alternatives. Especially, we discuss the role of the weight parameter p.

References:

- Jelle Goeman, Peter Bühlmann: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics (Oxford, England) 2007;23:980–987.
- Henryk Maciejewski: Gene set analysis methods: statistical models and methodological differences. Briefings in Bioinformatics 2013.
- Aravind Subramanian, et al: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005;102:15545–15550.
- Kai Wang, Mingyao Li, Maja Bucan: Pathway-based approaches for analysis of genomewide association studies. American Journal of Human Genetics 2007;81:1278–1283.

O-14

Rare Variant Extensions of the Transmission Disequilibrium Test: Application to Autism Exome Sequence Data

Zongxiao He[1], Brian O'Roak[2], Joshua Smith[2], Gao Wang[1], Stanley Hooker[1], Regie Santos-Cortez[1], Biao Li[1], Mengyuan Kan[1], Nik Krumm[2], Deborah Nickerson[2], Jay Shendure[2], Evan Eichler[2], Suzanne Leal[1]

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA; [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, USA
Email: suzannemleal@gmail.com

Many population-based rare variant (RV) association tests, that aggregate variants across a region, have been developed to analyze sequence data. A drawback of analyzing population-based data is that it is difficult to adequately control for population substructure and admixture, and spurious associations can occur. For RVs this problem can be substantial, because the spectrum of rare variation can differ greatly between populations. A solution is to analyze trio data, parents and a proband, using the transmission disequilibrium test (TDT), which is robust to population substructure and admixture. We extended the TDT to test for RV associations using four commonly used methods. We demonstrate that for all RV-TDT methods, using proper analysis strategies, type I error is well-controlled even when there are high levels of population substructure or admixture. For trio data, unlike for population-based data, RV allele-counting association methods will lead to inflated type I errors. However type I errors can be properly controlled by obtaining p-values empirically through haplotype permutation. The power of the RV-TDT methods was evaluated and compared to the analysis of case-control data using a number of genetic and disease models. The RV-TDT was also used to analyze exome data from 199 Simons Simplex Collection autism trios and an association was observed with variants in ABCA7. Given the problem of adequately controlling for population substructure and admixture in RV association studies and the growing number of sequence-based trio studies, the RV-TDT is extremely beneficial to elucidate the involvement of RVs in the etiology of complex traits.

O-15

Identification of Rare Causal Variants in Sequence-Based Studies: Methods and Applications to Gene VPS13B, Involved in Cohen Syndrome and Autism

Iuliana Ionita-Laza[1], *Marinela Capanu*[2], *Silvia De Rubeis*[3], *Kenneth McCallum*[1], *Joseph Buxbaum*[3]

[1]Columbia University, New York City, USA; [2]Memorial Sloan Kettering Cancer Center, New York City, USA; [3]Mount Sinai School of Medicine, New York City, USA
Email: ii2135@columbia.edu

Pinpointing the small number of causal variants among the abundant naturally occurring genetic variation is a difficult challenge, but a crucial one for understanding precise molecular mechanisms of disease and follow-up functional studies. We propose and investigate two complementary statistical approaches for identification of rare causal variants in sequencing studies: a backward elimination procedure based on groupwise association tests, and a hierarchical approach that can integrate sequencing data with diverse functional and evolutionary annotations for individual variants. Using simulations, we show that incorporation of multiple bioinformatic predictors of deleteriousness, such as PolyPhen-2, SIFT and GERP++ scores, can improve the power to discover truly causal variants. As proof of principle, we apply these two methods to VPS13B, a gene mutated in the rare neurodevelopmental disorder called Cohen syndrome, and recently reported with recessive variants in autism. We identify a small set of promising candidates for causal variants, including a rare, homozygous probably-damaging variant that could contribute to autism risk.

O-16

A Novel Integrated Framework for Rare Variant Analysis

Ramouna Fouladi[1,2], *Kyrylo Bessonov*[1,2], *François Van Lishout*[1,2], *Jason Moore*[3], *Kristel Van Steen*[1,2]

[1]Systems and Modeling Unit, Montefiore Institute, University of Liege, Liege, Belgium; [2]Bioinformatics and Modeling, GIGA-R, University of Liege, Liege, Belgium; [3]Dartmouth College, Lebanon, USA
Email: ramouna.fouladi@student.ulg.ac.be

Due to advances in whole-genome sequencing technologies, several hypotheses regarding the involvement of rare variants in human complex diseases can be tested. When handling genetic variants, most methods either adopt a single locus, multiple locus or a collapsing strategy and either only consider rare variants or incorporate in the analysis both rare and common variants. In the context of rare variant analysis, there is still room for improvement, so as to accommodate a wide variety of complex trait types and several study design configurations. Ideally, a generic tool is created that can deal with different granularities of omics information (i.e., different architectures of common and rare variants, epigenetic markers, gene expression).

Here, a novel omics association analysis technique is proposed that builds upon the Model-Based Multifactor Dimensionality Reduction (MB-MDR) framework. At the basis of the method lies a data organization step that involves clustering of individuals. In the first implementations of MB-MDR, these features were SNPs, and individuals were clustered according to their genotypes. In genomic MB-MDR, any feature (continuous or categorical) can be analyzed, and features mapped to genomic 'regions of interest' (ROIs) are submitted to a clustering algorithm to find groups of similar individuals on the basis of selected ROIs.

We propose to cluster individuals according to their similarities based on rare and common variants in pre-selected ROIs, after which classic MB-MDR is applied. The performance of several feature selection methods, similarity measures, and clustering algorithms in genomic MB-MDR is investigated using synthetic and real-life next-generation sequencing data.

O-17

Exploring Allele Specific Expression from RNA-Seq Data Using Logistic Mixed Effects Regression

Line Skotte[1], *Melissa Sayres*[2], *Rasmus Nielsen*[2]

[1]Bioinformatics, Department of Biology, University of Copenhagen, Denmark; [2]Departments of Statistics and Integrative Biology, University of California, Berkeley, USA
Email: line@binf.ku.dk

When the two alleles of a gene are transcribed in unequal proportions the gene is subject to allele specific expression (ASE). RNA-seq allows genome wide quantification of ASE in sufficiently expressed genes. Recent studies (Lappalainen et al., Nature 2013) points to ASE-based analysis as future identificator of low-frequency regulatory variants.

When reads produced from RNA-sequencing are mapped to the transcriptome, heterozygous sites allows us to estimate the allelic ratio. State-of-the-art methods for inferring ASE uses SNP-wise binomial testing, and the most appropriate methods incorporate the possibility of mapping biases. However, a number of sources of techincal noise are typically not included in the analyses, including the well-known overdispersion of read counts inherent in NGS data.

We develop a new method for estimating ASE based on combining the information from multiple SNPs within a transcribed unit to disentangle ASE from technical noise. Using a logistic regression model to explicitly model the effects of reference bias and SNP type, we can combine information from many loci in the genome to control for bias and noise specific to a particular data set. We use a mixed effects approach to combine information regarding ASE from many individuals in a population for each gene and to test hypotheses regarding ASE in individual genes and functional categories of genes.

We apply this framework to the open-access GEUVADIS mapped RNA-seq data combined with the 1000 Genomes Project phased variant call data.

O-18

Combining Distinct Classifiers for Assessing Proteomic Diagnosis and Prognosis

Alexia Kakourou

Leiden University Medical Center, The Netherlands
Email: a.a.kakourou@lumc.nl

We consider a proteomic mass spectrometry case-control study for the calibration of a diagnostic rule for patients' disease status allocation. We propose an approach for combining a collection of classifiers for the construction of a 'combined' classification rule in order to enhance calibration and prediction ability. In a first stage this is achieved by calibrating distinct classifiers separately, each-one using the entire proteomic data set. A double leave-one-out cross validatory approach is used to estimate the class predicted probabilities on which the combination method will be calibrated. The efficacy of the combination approach is examined both through a breast cancer proteomic data set and through simulation studies. Our experimental results indicate that in many circumstances gains in classification performance and predictive accuracy can be achieved.

The breast cancer data set we analyse is the same data set which was used in the International Competition on Proteomic Diagnosis (Mertens, 2008) and the key idea of combining classifiers was recommended by David Hand as 'an effective way to improve classification predictions' (Hand, 2008).

O-19

Genome-Wide Scan of Metabolomics Data Using Non-Additive Intra-Locus Models

Yakov Tsepilov[1,2], S. Shin[3,4], N. Soranzo[3], T. Spector[5], Konstantin Strauch[6], Christian Gieger[6], Yurii Aulchenko[1], Janina Ried[6]

[1]Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia; [2]Novosibirsk State University, Novosibirsk, Russia; [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; [4]MRC University of Bristol Integrative Epidemiology Unit (IEU), Bristol, UK; [5]Department of Twin Research and Genetic Epidemiology, King's College London School of Medicine, St Thomas' Hospital, London, UK; [6]Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
Email: drosophila.simulans@gmail.com

Genome-wide association studies (GWAS) are the tool of choice for identification of genetic loci involved in complex traits. With the availability of high-throughput technologies for metabolite measurements, GWAS identified multiple loci that affect different metabolites. In most GWAS, however, the effect of a SNP on the phenotype is assumed to be additive. Other genetic models such as recessive, dominant or over-dominant were not yet considered in metabolomics GWAS. At the same time, there are theories that emphasize the relevance of non-additive effects as a consequence of physiological mechanisms.

In this study we systematically analyzed non-additive effects on a large panel of serum metabolites and all possible ratios in a population based study (KORA F4, N = 1,785). We applied a two-step approach. The first step screened for SNP effects on metabolite concentrations and ratios (22,801 in total) using a genotypic 2 d.f. model, which can identify additive as well as non-additive effects. In the second step the best genetic model was determined by comparing 1 d.f. models that correspond to dominant, additive, recessive, or over-dominant modes of inheritance. Twenty loci were found to be genome-wide significantly associated (Bonferroni corrected p-value $\leq 2.19\times10^{-12}$) with at least one metabolite or ratio, some of which have not been found in GWAS on metabolites before; for five loci the best model was non-additive. We were able to replicate large part of our findings in an independent study (Twins-UK, N = 846).

Unexpectedly low portion of non-additive loci among found were discovered. We suggested some possible explanations why we can identify a lot of additive genes using the provided scheme of analysis. On the other hand this prejudices the existing theories about majority of non-additive genes (as metabolic fluxes theory) and possibly proves the opposite theories about additive control of complex traits.

O-20

Genome-Wide Environmental Sensitivity Analysis of Human Metabolomics Data

Sodbo Sharapov[1,2], Yakov Tsepilov[1,2], Janina Ried[3], Konstantin Strauch[3], Christian Gieger[3], Yurii Aulchenko[1]

[1]Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia; [2]Novosibirsk State University, Novosibirsk, Russia; [3]Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
Email: sharapovsodbo@gmail.com

Genome-wide association studies (GWAS) were successful for finding loci, which are significantly associated with human complex traits and common diseases. In most GWAS interactions between both genetic factors themselves and environmental and genetic factors were ignored. Complex genetic models, which include many interacting loci and environmental factors can help to identify new genes and improve our understanding of the genetic architecture of complex traits. A major challenge in finding interactions is the high computational complexity of direct tests e.g. all possible gene x gene interactions. An alternative approach to detect interactions is based on screening for the heterogeneity of the trait variance: under the hypothesis of interaction, the variance of the trait may vary between different genotypes. In other words, a significant difference in the trait variance between genotypes indicates that this locus is potentially involved in interactions with unknown factors. These factors may be both genetic and environmental, and they need not be available for this type of analysis. If this strategy is used as screening step, the computational complexity will be reduced essentially.

This approach has been implemented in the Squared residuals Value Linear Model (SVLM) method, allowing for an analysis of imputed genotypes. We applied SVLM to genome-wide data of the population based KORA F4 study (1,800 individuals, 2.6 million imputed SNPs). 151 concentrations of metabolites measured in human blood serum were analyzed. We found a number of loci, which are potentially involved in interactions in the control of the human metabolome. Subsequently factors, which are interacting with the identified loci, should be found. These results are the basis for a future discovery of genetic and environmental factors that are interacting with the genetic loci identified in our analysis.

O-21

Bayesian Multiple Regression Methods for Mapping Longitudinal Quantitative Traits

Zitong Li[1], *Mikko Sillanpää*[2]

[1]University of Helsinki, Finland; [2]University of Oulu, Finland
Email: zitong.li@helsinki.fi

Many quantitative traits such as height, weight and blood pressure change over developmental process of life. When this type of longitudinal data is measurable in quantitative trait loci (QTL) or association mapping studies, it is often beneficial to take the dependency structure among the repeated measurements over time into consideration. We propose two possible Bayesian models for analyzing longitudinal traits, a multilevel model and a linear mixed effects model. The multilevel model can be seen as a two step approach: in the first step we estimate the phenotypic temporal trend of each individual and consider the estimated parameters as the latent traits, and in the second step we map those latent traits to the genetic markers. While in a mixed effect model, the temporal trends and effects of markers are simultaneously estimated. In these methods, a slab and spike prior is assigned to the coefficient of each marker for variable selection, and an efficient Gibbs sampling method is used for computation. A Bayesian false discovery rate base decision rule is further used for formally judging QTLs. We evaluate and compare the performances of the two approaches on both real and simulation data sets.

O-22

Fast Mixed Models for GWAS with Longitudinal Data

Karolina Sikorska, Emmanuel Lesaffre, Patrick Groenen, Paul Eilers

Erasmus Medical Center, Rotterdam, The Netherlands
Email: k.sikorska@erasmusmc.nl

There is a growing interest in genome-wide association studies with longitudinal data, but they present a formidable computational challenge. Mixed models are the common choice for analyzing correlated data. Fitting such a model for thousands of individuals typically takes around 1 second, which for 1 million of SNPs sums up to 12 days. With the advent of 1000 Genomes Project, the number of analyzed polymorphisms increased to around 30 million, or one year of computing. It is vital to significantly improve the speed of the computations.

We propose an algorithm for fast mixed model fitting in the GWAS setting. We exploit two characteristics of the problem. First, we write mixed model equations as penalized least squares problem, where the penalty matrix is derived from the covariance matrix of the random effects. We obtain a system of equations comparable to Henderson equations. Second, we assume that the penalty matrix is known (from the model without a SNP) and does not need to be re-estimated for each SNP. This assumption is reasonable as genetic effects in GWAS are very small. Next, the system is solved exactly, using many computational tricks. We avoid operations on large matrices and instead exploit the structure of the equations to solve them in semi-symbolic way.

Our algorithm is implemented in pure R code using basic matrix operations. It reduces computation time by several orders of magnitude. Fitting 1 million of mixed models for 10 thousands individuals can be done within half an hour on a single personal computer. We obtain essentially exact estimated effects and very accurate p-values in the typical GWAS range, -log10(p)<8.

Lastly, we compare the performance of that algorithm with other approximate procedures, especially with proposed by us earlier conditional two-step approach.

O-23

Quantifying the Degree of Deviation from Hardy-Weinberg Equilibrium in Meta-Analyses: A Comparison of Effect Size Measures

Michael Preuß, Andreas Ziegler

Institute of Medical Biometry and Statistics, University Hospital Schleswig-Holstein – Campus Lübeck, Germany
Email: michael.preuss@imbs.uni-luebeck.de

Detecting deviation from Hardy-Weinberg equilibrium (HWE) is an important step in the quality control of genetic epidemiological studies. This also holds true for replication studies in the meta-analytic setting. In previous work, formal statistical tests were proposed for detecting deviation from HWE in meta-analyses. However, neither the overall degree of deviation from HWE nor the amount of heterogeneity between studies was quantified. As effect measures for quantification, the disequilibrium coefficient (D), the relative excess heterozygosity (REH), and the inbreeding coefficient (f) could be used. In a simulation study, we show under the assumption of HWE that D and f have biased point estimates and inflated type I error levels if allele frequencies vary between studies. In contrast, the REH has unbiased estimates and valid type I error levels. We visualize point and interval estimates for detecting outlier studies using forest plots, and illustrate its usefulness using published data. We recommend the REH for quantifying the degree of deviation from HWE in meta-analyses. Its calculation and the visualization of results can easily be done with standard statistical software.

O-24
Recombination Promotes Canalization Against Deleterious Mutations

Bengt Bengtsson

Lund University, Sweden
Email: bengt_olle.bengtsson@biol.lu.se

It is a necessary part of life that all organisms suffer deleterious mutations. Recurrent deleterious mutations at a locus are removed by selection, thereby causing a loss in mean population fitness equal to the mutation rate.

But what happens if at another locus a rare allele segregates that decreases – but does not abolish – the negative phenotypic effect of the mutations at the first locus? The new allele does not change the mean fitness of the population (the mutation load still equals the mutation rate), but is it favoured by some indirect selection?

The answer is yes, but recombination affects the evolutionary behaviour of this ameliorating allele in an interesting way. The loser the linkage between the two loci is, the easier it is for the modifying allele to spread.

Thus, recombination promotes canalization against recurrent deleterious mutations.

O-25
Population Genetics Interpretation of Higher Order Linkage Disequilibrium

Stefan Böhringer, Brunilda Balliu

Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands
Email: emgm@s-boehringer.org

Linkage disequilibrium (LD) parameter D is the covariance between indicator variables on alleles at two loci. This concept can be generalized to more than two loci and we define joint cumulants to describe such higher-order LD. Pairwise LD can be standardized in several ways. D' is a linear standardization with respect to maximal/minimal possible values of D for given marginals.

We develop an analogous standardization for joint cumulants and present an algorithm to compute standardized cumulants. D' equal to one has the interpretation that no recombinations have yet taken place between two loci. We show that a similar interpretation is true for standardized cumulants. Standardized cumulants can therefore be used to better describe the genetic architecture of a given set of markers as compared to D'. For a set of N markers 2^N-1 cumulants are derived and visualization is challenging. We present several visualizations of complexity N (one value per locus) that can be used to describe the genetic architecture. We use hapmap data as examples.

O-26
Estimating Trace-Suspect Match Probabilities in Forensics for Singleton Y-STR Haplotypes Using Coalescent Theory

Amke Caliebe

Institute of Medical Informatics and Statistics, University of Kiel, Germany
Email: caliebe@medinfo.uni-kiel.de

In forensics, Y chromosomal DNA plays a crucial role since it allows identifying the male-specific portion of DNA evidence. In this context the so-called trace-suspect match probability is of importance, i.e. the probability that a certain individual (e.g. the donor of a trace found at a crime scene) has the same DNA profile as another individual (usually a suspect) drawn randomly from the same population. Of special interest and increasingly frequent are singleton haplotypes, i.e. haplotypes observed only once in a reference database augmented by a suspect profile. We compared the performance of four estimators of singleton match probabilities for Y-STRs, namely the count estimator, both with and without Brenner's so-called kappa correction, the surveying estimator, and a previously proposed, but rarely used, coalescent-based approach implemented in the BATWING software. Extensive simulation with BATWING of the underlying population history, haplotype evolution and subsequent database sampling revealed that the coalescent-based approach and Brenner's estimator are characterized by lower bias and lower mean squared error than the other two estimators. Moreover, in contrast to the two count estimators, both the surveying and the coalescent-based approach exhibited a good correlation between the estimated and true match probabilities. However, although its overall performance is thus better than that of any other recognized method, the coalescent-based estimator is still very computation-intense. Its application in forensic practice therefore will have to be limited to small reference databases, or to isolated cases of particular interest, until more powerful algorithms for coalescent simulation have become available.

O-27
Inferring Human Population History and Gene Flow from Multiple Genome Sequences

Stephan Schiffels

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK
Email: stephan.schiffels@sanger.ac.uk

The availability of complete human genome sequences from populations across the world has given rise to new population genetic inference methods that explicitly model their ancestral relationship under recombination and mutation. So far, application of these methods to evolutionary history more recent than 20–30 thousand years ago and to population separations has been limited. Here we present a new method that overcomes both of these shortcomings. The Multiple Sequentially Markovian Coalescent (MSMC) analyses the observed pattern of mutations in multiple individuals, focusing on the first coalescence between any two individuals. Results from applying MSMC to genome sequences from nine populations across the world give information about human population history as recently as 2,000 years ago. They suggest that the genetic separation of non-African from African populations was a gradual process that began before 150,000 years ago and lasted for over 100,000 years, and give information about the separation of populations after the out-of-Africa event including the bottleneck in the peopling of the Americas, and within Africa, East Asia and Europe.

O-28

Non-Parametric Estimation of the Age-Dependent Penetrance of a Mutation Using Familial Data

Flora Alarcon[1], *Violaine Planté-Bordeneuve*[2], *Hervé Perdry*[3]

[1]Laboratoire MAP5 – UMR CNRS 8145 – Université Paris Descarte, France; [2]Département de Neurologie, Hôpital Universitaire Henri Mondor, Créteil & INSERM U955-E10 Université de Créteil, France; [3]Université Paris-Sud UMR-S 669 & Inserm U669, France
Email: herve.perdry@gmail.com

For mendelian diseases with late onset, estimating the penetrance of a mutation as a function of the age – and of known penetrance modifiers such as sex, parent-of-origin, etc – is essential for genetic counselling and prevention. Being able to test wether a candidate penetrance modifier has a significant effect is also primordial for improving our understanding of the disease.

The Proband Excluded Likelihood (PEL) has been proposed [1] to estimate the penetrance using pedigrees ascertained through index cases. The PEL relies on likelihood maximum; it corrects for ascertainment bias by using only the phenotypic information from the relatives of the index case, and not of the index case himself.

However the PEL is a parametric method, which models the penetrance with a Weibull cumulative distribution. This is a limitation of the method, as it can introduce a bias at some ages if the true penetrance function is not well approximated by a Weibull distribution. Moreover, nothing had been proposed to incorporate covariates in the analysis.

We extend the PEL to obtain a non-parametric (Kaplan-Meier) penetrance estimate. We also propose a Cox model to accommodate covariates such as sex, parent-of-origin, or the genotype at a putative modifier gene. The implementation of this non-parametric PEL relies on an EM algorithm, which itself makes use of the Elston-Stewart algorithm for computing probabilities in pedigrees.

The method is illustrated on simulated data, and on real data for the Transthyretin Amyloid Neuropathy, with French and Portuguese families, showing in particular an important parent-of-origin effect.

Reference:

1 Alarcon F, et al: Genet Epidemiol 2009.

O-29

Detecting Relatives from Genome-Wide Data

Meng Sun

Department of Health Science, University of Leicester, UK
Email: ms626@le.ac.uk

Relationship estimation from genetic markers is relevant to many areas, including genealogical research, genetic counselling, forensics, conservation genetics and genetic epidemiology. Traditionally unlinked genetic markers (microsatellites) are used but the problems which can be solved are limited, firstly because the number of unlinked genetic markers is limited, and secondly because certain relationships with the same proportion of IBD sharing can only be distinguished from each other with linked genetic markers. Here we exploit the increasing availability of dense genome-wide SNP data for estimating distant relationships.

For pairwise relationship estimation, method of moment (MoM) estimators can give a general estimate of the degree of relatedness without any prior information and are quite accurate for close relationships. MoM estimators are also robust to the effect of Linkage Disequilibrium (LD). Often there is additional information leading to a set of plausible alternative relationships. In this case, a likelihood approach is better at detecting more distant relatives. There is the added advantage over MoM estimators in that extra individuals can be considered jointly in the pedigree likelihood calculation but the approach is sensitive to LD.

It is clear that the increase in information obtained from large sets of linked markers substantially increases the number of problems that can be solved. For distant relatives, LD has to be accounted for. Our aim is to use both MoM and likelihood approaches together for pedigree reconstruction.

O-30

Guaranteed Maximum Likelihood Pedigree Reconstruction

Nuala Sheehan[1], *Mark Bartlett*[2], *James Cussens*[2]

[1]University of Leicester, UK; [2]University of York, UK
Email: nas11@le.ac.uk

There are many applications requiring identification of relatives amongst a set of individuals and all such estimation problems can be described in terms of reconstructing the relevant pedigree. In theory, estimating the pedigree for a given set of individuals from genetic marker data simply requires consideration of all possible relationships amongst them and then computing the likelihood for each. For any reasonably sized problem, brute force enumeration is clearly impractical. The reconstruction problem can be formulated as a problem of Bayesian network (BN) learning or, more generally, graphical structure estimation, and is known to be NP-hard. The desired graph has structural constraints so maximum likelihood pedigree reconstruction can be viewed as a constrained optimisation problem.

We propose an integer linear programming (ILP) approach which is adapted to find valid pedigrees by imposing appropriate constraints. Our method, unlike others, is not restricted to small pedigrees and is guaranteed to return a maximum likelihood pedigree. Due to the probabilistic nature of genetic inheritance, there is no guarantee that the most probable pedigree is the true pedigree. Failure to acknowledge such (inevitable) uncertainty could lead to spuriously confident inferences. With additional constraints, we can also search for multiple high probability pedigrees and thus account for this uncertainty. For accurate reconstruction, prior information on age, sex, specific relationships, population characteristics etc. should be incorporated when available. The method performs well in a straightforward situation and extensions to more complex problems seem feasible.

O-31

Statistical Challenges for the Analysis of Human Longevity Data in Families

Jeanine Houwing-Duistermaat, Mar Rodriguez Girondo

Dept of Medical Statistics and Bioinformatics, Leiden University Medical Centre, The Netherlands
Email: j.j.houwing@lumc.nl

Although there is evidence from several studies that longevity aggregates within families, identification of genetic factors has not been successful. Reasons for lack of progress might be heterogeneity across studies and the ad hoc definition of being older than a specific threshold (e.g. older than 90 years of age). We will consider survival models for the analysis of longevity in family studies. Challenges are to model the ascertainment of the families, to take into account correlation between family members and to deal with delayed entry. There is literature about dealing with delayed entry in independent individuals, but methods that can be applied to clustered data are limited.

This work is motivated by the Leiden Longevity study comprising 420 families with at least two nonagenarian siblings. Genome wide SNP arrays and ten years of follow up are available; 13% is still alive, maximum observed age is 107 years. To obtain parameter estimates for the genetic associations which can be combined with population based data for meta-analysis, we propose a Cox model with inverse probability weighting to account for the selection of the families. The weights will be based on the latent frailties in a proportional hazards model. The frailty distribution appears to depend on the selection of the siblings and delayed entry mechanisms, that are affected by factors such as family size. By means of simulation we will study the effect of ascertainment on the parameter estimates and frailty distribution, and we will assess the performance of the weights.

Poster Presentations

P-1

A Clustering Approach for Mapping Rare Variants Based in Mutual Association

Saurabh Ghosh[1], *Soudeep Deb*[2]

[1]Human Genetics Unit, Indian Statistical Institute, Kolkata, India;
[2]Department of Statistics, University of Chicago, USA
Email: saurabh@isical.ac.in

In spite of successful identification of a large number of common variants associated with various complex disorders, a substantial proportion of the total variation in a trait still remains unexplained. It is being argued that the 'Common Disease Common Variant' paradigm needs to be modified and the 'missing heritability' can possibly be explained by rare variants, which could not be identified using genome-wide association studies. Most existing methods are based on collapsing multiple variant sites using different statistical algorithms. Motivated by the combined multivariate and collapsing (CMC) algorithm, we propose a clustering of rare variant sites, but based on their mutual extent of association rather than similarity in allele frequencies as proposed in CMC, thereby reducing the possibility of combining functional and non-functional variants. The Fisher's exact test is performed to identify blocks of variant sites such that the initial site in each block is associated with all other sites in the block. The test for association is performed within each block by comparing the proportions of affected and unaffected individuals carrying at least one copy of a rare variant using a variance stabilizing sine transformation. We carry out extensive simulations under different rare variant models and compare the false positive rate and the power of our proposed method with some of the popular competing methods: CMC, adaptive SUM, WSS, TestRare, RareCover and Kernel-based Adaptive Clustering. We find that the proposed test procedure yields more power than the existing approaches, especially with increasing sample size, while maintaining the correct size.

P-2

Using Brain Specific Gene Expression Networks to Identify Causal Epilepsy Genes

Melanie Bahlo

The Walter and Eliza Hall Institute of Medical Research, Australia
Email: bahlo@wehi.edu.au

Massively parallel sequencing (MPS) has made a significant impact in the discovery of disease causing mutations for variants with large effect sizes. The ability to directly observe the putative variant within a sample or a cohort allows the possibility to overcome both locus and allelic heterogeneity and furthermore investigate non-segregating genetic models such as de novo and somatic disease causing variants. However, each MPS experiment itself still has considerable failure rate (>50%), whereby either no variant or too many variants remain at the end of an analysis. We have been using publicly available, brain specific, gene expression data to infer gene expression networks with sets of known disease causing genes for a variety of brain-related disorders to help identify likely causal genes. We show results comparing different measures of association and how ordering of correlation matrices and the usage of partial correlation matrices can give insights into how these genes interact with each other. Finally we show these findings can help to identify new candidate disease causing genes for some of these diseases.

P-3

New Insights to Genomic Control Correction

Valentina Escott-Price

MRC Centre for Neuropsychiatric Genetics & Genomics,
Cardiff University, UK
Email: EscottPriceV@cf.ac.uk

The Genomic Control (GC) value plays an important role in the estimation of the overall inflation of the test statistics in a genome scan. We theoretically show that in the presence of genotyping errors in the case and control samples, the test statistic asymptotically follows a scaled non-central chi-square distribution. As the scaling factor and non-centrality parameter depend on the allele frequencies, the correction of the statistic should be based on the individual parameters for each separate SNP; in particular, rare variants will generally require stronger correction. We suggest an alternative approach to correction of the test statistic for the individual SNP on the basis of the sample sizes, the allele frequencies and the allele misrepresentation rate. We show that due to its linear growth in total sample size, the influence of the non-centrality will dominate over the scaling factor, and hence the test statistic cannot be effectively corrected by a GC factor only. Thus, the standard GC seems inappropriate, especially where the distortion can be partly attributed to differential genotyping errors. While batch effects within the case and the control samples (when combined from separate chips/platforms) can be dealt with by adding the batch as a regression covariate, our correction will apply when there is no separate batch variable because cases and controls form different batches. We illustrate our results comparing samples of the same 165 individuals (CEU + TSI) for whom 1,278,013 SNPs which have been genotyped in both the HapMap3 and 1000 Genomes projects.

P-4

Bayesian Inference for Learning Between-Pathway Network: A New Tool for Studying Drug-Disease Interactions

Naruemon Pratanwanich, Pietro Lio

University of Cambridge, UK
Email: np394@cam.ac.uk

The interactions between genes within pathways have been functionally annotated, which provides useful information how genes works for a specific mechanism. Biological processes responsive to drug treatments or disease perturbations often result from certain sets of pathways. We therefore assume that pathways do not work in isolation. The simplest reason is that genes may regulate others in different pathways or a gene may be active in multiple pathways. Nevertheless, pathways can interact with each other through other molecular levels. Our between-pathway network approach can address this problem by allowing complex relationships between many actors, such as genes, proteins and metabolites in the network. In order to determine the interactions between pathways, we have developed a Gaussian Markov Random Field (GMRF) under a Bayesian matrix factorization framework given gene expression data and known gene-pathway memberships. Assuming a Gaussian distribution of pathway responsiveness to drug treatments or disease perturbations allowed us to infer the correlations between pathways. Out of gene expression data of 1,169 drugs together with 236 known pathways, 66 of which were disease-related pathways, our model yielded a significantly higher average precision than the existing methods for identifying pathway responsiveness to drugs that affected multiple pathways. This confirms our assumption that pathways are not independent, an aspect that has been commonly overlooked. We also demonstrate three case studies illustrating that the between-pathway network provides insights into disease comorbidity, drug repositioning, and tissue-specific gene expression analysis.

P-5

Polygenic Scoring used to Investigate Pleiotropy between Complex Traits – Application to Suicide Phenotypes and their Genetic Relationship with Psychiatric Disorders

Niamh Mullins, Cathryn Lewis

King's College London, UK
Email: Niamh.mullins@kcl.ac.uk

Background: Depression and suicidality are two heritable and often comorbid disorders. Genome-wide association studies (GWAS) on these complex traits have failed to identify specific genetic variants, due to lack of power to detect small effect sizes. Polygenic score analysis simultaneously investigates thousands of SNPs, which individually do not reach statistical significance. This can be used to assess pleiotropy between different phenotypes.

Methods: Subsets of SNPs were selected from a discovery GWAS on depression, according to their P value. A polygenic score consisting of all these alleles, weighted by their log odds ratios, was created for each individual in an independent depression validation sample. A logistic regression was used to test the ability of the scores to predict suicide attempt or ideation status in the validation dataset, to investigate the genetic relationship between depression and these associated traits.

Results: Polygenic scores for depression showed significant predictive ability for suicidal ideation in an independent validation sample, accounting for 1% of the variance in phenotype (Nagelkerke R2 = 0.01, P = 0.015). Polygenic scores for depression were also predictive of suicide attempt (Nagelkerke R2 = 0.003, P = 0.013). The amount of variance explained increased as more SNPs were added to the polygenic score, at increasingly liberal P value thresholds.

Conclusion: Using polygenic scores as predictors in independent datasets is an appropriate approach to investigating the genetics of highly polygenic traits and is a useful way to demonstrate the genetic relationship between comorbid phenotypes, such as depression and suicidality.

P-6

Mendel's Laws and Stochastic Processes

David Almorza[1], *Mariana Kandus*[2], *Juan Salerno*[2]

[1]University of Cádiz, Spain; [2]Genetic Institute Ewald Favret, Argentina
Email: david.almorza@uca.es

Introduction: In this work we are going to prove that from Mendel's Law we can obtain an example of the Hardy-Weinberg Equilibrium. In the demonstration we are going to use stochastic processes theory and Markov Chains to develop the transition matrix.

Keywords: Mendel's laws; Hardy-Weinberg equilibrium; Markov chains.

Mendel Matrix and Markov Process.

From first and second Mendel's Laws, law of segregation and law of independent assortment, and if we consider that the capital 'P' represents the dominant factor and lowercase 'p' represents the recessive, ordered (PP, Pp, pp), it can be defined a transition matrix A as follows:

```
$$A=\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} & 0 \\\noalign{\medskip}
\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\\noalign{\medskip}
0 &\frac{1}{2}& \frac{1}{2}
\end{pmatrix}$$
```

The distribution over states can be written as a stochastic row vector x with the relation `$x_n=x_0A^n$`.

Let us consider the initial position PP, that can be written as `$x_0=(1\ 1\ 0)$` . In this situation,

```
$$\lim_{n\to\infty} (1\ 0\ 0)A^n=(\frac{1}{4}\ \frac{1}{2}\ \frac{1}{4})$$
```

Otherwise, by taking in consideration the initial position Pp, that can be written as `$x_0=(0\ 1\ 0)$`, then:

```
$$\lim_{n\to\infty} (0\ 1\ 0)A^n=(\frac{1}{4}\ \frac{1}{2}\ \frac{1}{4})$$,
```

and also, from `$x_0=(0\ 0\ 1)$` we can obtain:

```
$$\lim_{n\to\infty} (0\ 1\ 0)A^n=(\frac{1}{4}\ \frac{1}{2}\ \frac{1}{4})$$.
```

So we can conclude that from Mendel's laws we can obtain an example for the Hardy-Weinberg equilibrium, what was expected as the frequency of alleles does not change from generation to generation (in other words, the population does not evolve) and after one generation of random mating, offspring genotype frequencies can be predicted from the parent allele frequencies.

References:

Brzezniak Z, Zastawniak T: Basic Stochastic Processes. Springer-Verlag (London) 2000.

Falconer DS: Introduction to Quantitative Genetics, Ed. 2. Longmans Green, London, New York, 1981.

Klug WS, Cummings MR: Essential of Genetics. Prentice Hall (New Jersey), 2002.

Mendel G: Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen, 1866, pp 3–47.

P-7

Integration Analysis of 'OMICS' Data Using Penalized Regression Methods: An Application to Bladder Cancer

Silvia Pineda[1,2], *Nuria Malats*[2], *Kristel Van Steen*[1]

[1]University of Liege, Belgium; [2]Spanish National Cancer Research Centre, Spain
Email: spineda@cnio.es

There is a growing interest in combining different 'omics' datasets to further dissect the mechanisms of human complex disease traits. The simplest form of data integration involves two different data types (for instance, GWAS and expression data, as in eQTL analyses). The availability of more than 2 omics data types derived from the same set of individuals is rare. And when these exist, several technical and statistical hurdles need to be taken to ensure optimal power and reliable biological relevant relationships. In this work, we rely on variable selection methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) approach and the Elastic Net method. Although these are promising statistical methods presenting good properties in the context of high-throughput data, they do not provide p-values to assess statistical significance of relationships, or give a formal assessment of the overall goodness-of-fit. Therefore, we adopt a permutation-based strategy to assess significance of discovered relationships, building upon the concepts of deviance and the Mean Squared Error (MSE). Validity and utility of these methods is shown on synthetic data as well as real data from the pilot Spanish Bladder Cancer/EPICURO study (bladder cancer cases recruited in 2 hospitals in Spain in 1997–1998), while integrating gene expression, DNA methylation and genome-wide SNP data from tumor samples.

P-8

Estimation of Cell-to-Cell Regulatory Heterogeneities from Cell Populations

Christiane Fuchs

Helmholtz Zentrum München, Germany
Email: christiane.fuchs@helmholtz-muenchen.de

Even when appearing perfectly homogeneous on a morphological basis, tissues can be substantially heterogeneous in single-cell molecular expression. Such heterogeneities might govern the regulation of cell fate, and hence one is interested in their quantification in a given tissue. This is typically done given single-cell gene expression data. However, such data is expensive and subject to high technical noise. Hence, in this project we refrain from single-cell data; instead, small numbers of cells are randomly selected, and the subpopulation average expression level is measured. I will discuss how heterogeneities can be detected from such data by application of statistical methods, and how the proportions, mean values and standard deviations of the groups of different cells can be estimated. The estimation techniques are applied to measurements of human breast epithelial cells, showing remarkable agreement with experimental results. Importantly, population-level inference turns out to be much more accurate with pooled samples than with one-cell samples when the extent of sampling is limited.

P-9

Pharmacogenetic GWAS Meta-Analysis of LDL Cholesterol Response to Statins

Iris Postmus[1], Stella Trompet[1,2], Helen Warren[3,4], J. Wouter Jukema[2,5,6], Mark Caulfield[3,4]

[1]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; [2]Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands; [3]William Harvey Research Institute, Barts, UK; [4]The London School of Medicine, Queen Mary University of London, UK; [5]Durrer Center for Cardiogenetic Research, Amsterdam, The Netherlands; [6]Interuniversity Cardiology Institute of the Netherlands, Utrecht, The Netherlands
Email: h.r.warren@qmul.ac.uk

Purpose: Statins are widely prescribed for the prevention and treatment of cardiovascular disease with great proven effectiveness of 20–30%. However, it is also established that there is inter-individual variability in LDL-C response to statins, which may be partly due to pharmacogenetic variation. The only genetic variants consistently reported from previous studies are located within the APOE and LPA gene regions. To determine whether additional loci may influence LDL-C response to statins, we formed the Genomic Investigation of Statin Therapy (GIST) consortium and conducted a pharmacogenetic meta-analysis of genome-wide association studies of LDL-C response to statins, which is the largest, most comprehensive study of its kind conducted to date.

Methods: The meta-analysis comprises GWAS datasets from both randomized controlled trials (RCTs) and observational studies, including approximately 40,000 statin-treated subjects overall, divided into a first discovery stage and a second validation stage.

The response variable analysed is the difference between the natural log transformed LDL-C levels on- and off-treatment, and is adjusted for baseline LDL-C, in order to eliminate the confounded effect of association between the genetic variant and baseline LDL-C levels. Further adjustment was made for the type and dose of statins used, in order to combine several different types of statins across the contributing trials and within the observational studies.

Results: Overall, at genome-wide significant level, we have identified two new loci: SORT1/CELSR2/PSRC1 (rs646776, $\beta = -0.013$, SE = 0.002, $P = 1.05 \times 10^{-9}$) and SLCO1B1 (rs2900478, $\beta = 0.016$, SE = 0.003, $P = 1.22 \times 10^{-9}$), which have not been previously identified in GWAS. Furthermore we have successfully confirmed the known associations with APOE (rs445925, $\beta = -0.043$, SE = 0.005, $P = 1.58 \times 10^{-18}$) and LPA (rs10455872, $\beta = -0.059$, SE = 0.006, $P = 1.95 \times 10^{-11}$).

Our results were further investigated and validated with additional functional analyses, such as conditional, eQTL and pathway analyses. For example, the genome-wide conditional analysis highlighted 14 independent SNPs explaining 5% of the variance, of which 6 SNPs reached genome-wide significance in our combined meta-analysis. Collectively our functional and pathway analysis confirmed a strong biological and functional role in statin response for several strongly associated gene loci, including APOE, and SORT1/CELSR2/PSRC1. Furthermore, a genetic risk score including our 4 lead SNPs was significantly associated with coronary disease risk in theCARDIoGRAM and C4D consortium.

Conclusions: Our findings advance the understanding of the pharmacogenetic architecture of statin response, and illustrate that SNPs with modest effect on LDL response to these widely used drugs can influence coronary artery disease risk.

P-10

Exploiting Linkage Disequilibrium to Detect G×E Interactions in Case-Only Studies

Pankaj Yadav, Sandra Freitag-Wolf, Michael Krawczak

Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany
Email: yadav@medinfo.uni-kiel.de

Gene-environment (G×E) interactions have been invoked to explain the gap between the known heritability of common human diseases and the genetic component hitherto explained by disease-associated variants. Chronic inflammatory bowel disease (IBD) is an example of a condition where many genetic risk factors have been identified in the past, but where the possible interplay between genetic and environmental risk factors is notoriously understudied. One possible reason for this shortcoming may be an insufficient power of previous case-control or cohort studies. A different and more efficient strategy to analyze G×E interactions in epidemiological studies is the use of a case-only design. Here, genotype and exposure information from cases alone is used to estimate the level of interaction. In the past, case-only studies usually followed a candidate (or single) gene approach, and their genome-wide utility still has to be explored. For instance, the influence of linkage disequilibrium (LD) on the chance to detect GxE interaction has not been studied in much detail before. We therefore systematically assessed the power to indirectly detect GxE interactions by way of exploiting LD, following a case-only approach.

P-11

Breast and Ovarian Cancer Risk Assessment: Ascertainment Bias

Tinhinan Belaribi[1], *Antoine De Pauw*[2], *Flora Alarcon*[3], *Nadine Andrieu*[2], *Dominique Stoppa-Lyonnet*[2], *Gregory Nuel*

[1]Pierre et Marie Curie University (Paris 6), France; [2]Institut Curie, France; [3]Paris Descartes University, France
Email: belaribi.nina@gmail.com

Nowadays, genetics is essential in cancer risk assessment to establish the decision thresholds needed by clinicians to define prevention strategies. Germline mutations of the BRCA1 and BRCA2 genes are responsible for monogenic form of breast and ovarian cancer susceptibility. Genetic testing for these two genes is now a standard procedure for patients with severe family history (FH). However, BRCA1/2 mutations do not explain all cases. The BOADICEA model (Antoniou et al., 2002, 2004), used at Institute Curie, allows calculating the lifelong risk for an unaffected woman to be diagnosed with breast or ovarian cancer given her FH. Unfortunately, a retrospective study from Institute Curie's family data (De Pauw, 2012) has shown that BOADICEA has a poor predictive power, which could be partially explained by the complex problem of ascertainment bias.

Our initial work aimed to study, through simulations, the bias introduced by the ascertainment on the predictive risk calculation.

Results had shown an overestimation of the risk in absence of correction for the ascertainment process and an underestimation with the BOADICEA correction. Indeed, in this correction the likelihood is conditioned only by phenotypes while recruitment criterion for Institut Curie's family data may also depend on genotypes. Hence, we propose a Prospective Likelihood correction (Alarcon, 2009), which consists in conditioning the likelihood by the ascertainment event.

The unbiased estimation obtained by the Prospective method suggests the importance of good ascertainment criteria modelling and thus, the need to include clinical investigation in order to know the exact nature of ascertainment process and improve the model's reliability.

P-12

Discrimination Between Correlated SNPs in Genetic Association Studies: Comparison Between Case-Control and Familial Studies

Claire Dandine-Roulland[1], *Françoise Clerget-Darpoux*[2], *Hervé Perdry*[1]

[1]Université Paris-Sud UMR-S 669 &Inserm U669, France; [2]Inserm U781, France
Email: claire.dandine-roulland@inserm.fr

Association studies identify loci associated with a disease. However, the association between a Single Nucleotide Polymorphism (SNP) and the disease can result from a causal variant in linkage disequilibrium (LD) with the considered SNP. Even assuming that the true causal variant is among the observed SNPs, the power to discriminate between this causal variant and other SNPs in LD can be low [1].

The method developed in [1] relies on the comparison between the Armitage statistics at the considered SNPs. We adapt it to Affected Sib-Pair (ASP) and control data, using the test statistic proposed in [2] instead of the Armitage statistic. This statistic relies on the genotypes of control cases for the ASP, and on the number of alleles shared Identical-By-Descent by the affected sibs. We compare the discrimination power of the procedures using case-control and familial data, showing the advantage of the latter.

We show moreover that the familial information allows to estimate the LD between the observed SNPs and a putative unobserved causal variant, and to test for complete LD. In contrast, case-control information doesn't allow such an estimation.

We illustrate the proposed procedures on Multiple Sclerosis data from [3], consisting in 26 SNPs in IL2RA for 522 controls and 82 ASP.

References:

1 Udler MS, Tyrer J, Easton DF: Evaluating the Power to Discriminate Between Highly Correlated SNPs in Genetic Association Studies, Genetic Epidemiology 2010;34:463–468.

2 Perdry H, Müller-Myhsok B, Clerget-Darpoux F: Using Affected Sib-Pairs to Uncover Rare Disease Variants. Human Heredity 2012;74:129–141.

3 Babron MC, Perdry H, et al: Determination of the real effect of genes identified in GWAS: the example of IL2RA in multiple sclerosis. EJHG 2011;20:321–325.

P-13

Leveraging Different Sources of Signal in Genomic Predictions of Complex Traits

Athina Spiliopoulou[1], *Reka Nagy*[1], *Jennifer Huffman*[1], *Mairead Bermingham*[1], *Caroline Hayward*[1], *Igor Rudan*[2], *Harry Campbell*[2], *Alan Wright*[1], *Jim Wilson*[2], *Ricardo Pong-Wong*[3], *Chris Haley*[1], *Felix Agakov*[4], *Pau Navarro*[1]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, UK; [2]Centre for Population Health Sciences, University of Edinburgh, UK; [3]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, UK; [4]Pharmatics Limited, UK
Email: a.spiliopoulou@ed.ac.uk

Recent years have seen great advances in the collection of human genomic data and its association with complex phenotypes. Our goal is to assess the advantage of using previously published results to inform genomic predictors of complex traits and the plausibility of transferring predictive models across populations. We explore whether we can increase prediction accuracy by building meta-models that combine simpler predictors, each capturing a different type of predictive signal. We evaluate predictive performance within and across populations and examine how to maximise model transference and achieve good generalisation performance when the target individuals come from a different population to the training samples.

Here we focus on regularised linear models and evaluate their performance on predicting height, body mass index and high density lipoproteins in two population cohorts, originating in Croatia and Scotland. We examine models with different sparsity levels learned from data using lasso or elastic nets, or determined a priori, by considering markers and their corresponding effects from large meta-analysis association studies. We present results from within and across-population experiments, and demonstrate that transference is possible if we use samples from the target population for model selection, subject to sample size and trait architecture. We show that a meta-model combining penalised regression with meta-analysis-based polygenic scores is always better than either model on its own, suggesting that finding rigorous ways to incorporate the plethora of available data, biological annotations and previous results into statistical models is a promising research direction for complex trait prediction.

P-14

A Novel Tree-Based Procedure for Deciphering the Genomic Spectrum of Clinical Disease Entities

Cyprien Mbogning[1], *Hervé Perdry*[1,2], *Philippe Broët*[1,2]

[1]INSERM U669, France; [2]University Paris-Sud, France
Email: cyprien.mbogning@inserm.fr

Background: Dissecting the genomic spectrum of clinical disease entities is a challenging task. Recursive partitioning (or classification trees) methods provide powerful tools for exploring complex interplay among genomic factors, with respect to a main factor, that can reveal hidden genomic patterns. To take counfounding variables into account, a tree-based regression procedure has been proposed by Chen et al. [1]. It combines regression models and tree-based methodology.

Methods: We developed a novel procedure that represents an alternative to Chen's et al. procedure, using different selection criteria. A simulation study with different scenarios has been performed to compare the performances of the proposed procedure to the original PLTR strategy.

Results: The proposed procedure with a BIC criterion achieved good performances to detect the hidden structure as compared to Chen's et al. procedure. The novel procedure was used for analysing patterns of copy-number alterations in lung adenocarcinomas, with respect to KRAS mutation status, while controlling for a cohort effect. Results highlight two subgroups of pure of nearly pure wild-type KRAS tumors with particular copy-number alteration patterns.

Reference:

1 Chen J, Yu K, Hsing A, Therneau TM: A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genetic Epidemiology 2007;31:238–251.

P-15

Host Genomics in Sepsis: An Update on the Available Evidence and the Ongoing Methodological Challenges

Franziska Schöneweck, Miriam Kesselmeier, André Scherag

Clinical Epidemiology, Integrated Research and Treatment Center, Center for Sepsis Control and Care, Jena University Hospital, Jena, Germany
Email: andre.scherag@med.uni-jena.de

Angus and van der Poll (2013) state in a recent review on severe sepsis and septic shock that '…[t]here is considerable interest in the contribution of host genetic characteristics to the incidence and outcome of sepsis, in part because of strong evidence of inherited risk factors. …'. This interest is driven by heritability estimates derived from classic/formal genetics (Sorensen et al., 1988). Consequently many molecular genetic association studies have been performed to elucidate such host genomic factors (e.g. Namath and Patterson, 2011): Large inconsistencies have been reported for candidate gene association studies (e.g. Clark and Baudouin, 2006) and only one genome-wide association study has been published by now (Man et al., 2013).

To address the discrepancy between formal genetic expectations and molecular genetic findings we summarize the evidence from more recent formal genetic investigations and performed a systematic review on systematic reviews for candidate gene association studies related to sepsis. Finally, we discuss the apparent discrepancy between formal and molecular genetic results referring to statistical model assumptions, biology including the role of the genetic architecture and special challenges in the field of sepsis research.

References:

Angus DC, van der Poll T: Severe sepsis and septic shock. N Engl J Med 2013;369:840–851.

Sorensen TI, Nielsen GG, Andersen PK, Teasdale TW: Genetic and environmental influences on premature death in adult adoptees. N Engl J Med 1988;318:727–732.

Namath A, Patterson AJ: Genetic polymorphisms in sepsis. Crit Care Nurs Clin North Am 2011;23:181–202.

Clark MF, Baudouin SV: A systematic review of the quality of genetic association studies in human sepsis. Intensive Care Med 2006;32:1706–1712.

Man M, Close SL, Shaw AD, Bernard GR, Douglas IS, Kaner RJ, Payen D, Vincent JL, Fossceco S, Janes JM, Leishman AG, O'Brien L, Williams MD, Garcia JG: Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis. Pharmacogenomics J 2013;13:218–226.

P-16

Improvement of Genotype Imputation Accuracy Through Integration of Sequence Data from a Subset of the Study Population

Barbara Peil[1], *Maria Kabisch*[2], *Christine Fischer*[3], *Ute Hamann*[2], *Justo Lorenzo Bermejo*[1]

[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Germany; [2]Molecular Genetics of Breast Cancer (B072), German Cancer Research Center (DFKZ), Heidelberg, Germany; [3]Institute of Human Genetics, University Hospital Heidelberg, Germany
Email: peil@imbi.uni-heidelberg.de

In addition to genotypes from external data repositories (e.g. HapMap), the sequences of a subset of individuals from the study population have been shown to improve imputation accuracy. We examined different strategies to select this subset of individuals.

We assumed that the study population, which was genotyped using a commercial array, consisted of the two European subpopulations CEU and TSI (n = 204) from HapMap phase III. The remaining individuals from HapMap constituted the external reference panel (n = 798).

Five alternative strategies were examined to select the subset of sequenced individuals added to the external reference panel. The strategy 'none' incorporated no additional sequence to the external reference panel. The strategy 'random' incorporated the sequences of a random subset of 10% individuals in the study population. The strategies 'univariate depth', 'bivariate depth' and 'trivariate depth' relied on a genome-wide principal component analysis based on array data, followed by the identification of 10% of individuals with the largest statistical depth based on the first one, two and three principal components in the study population.

Genotypes were imputed using IMPUTE2. In order to assess imputation accuracy, HapMap genotypes of study individuals not selected for sequencing which were not represented in the commercial array were masked and subsequently imputed using different reference panels. Imputation accuracy was measured by the mean imputation quality score (IQS) between true and imputed genotypes.

The inclusion of additional sequences from the own study population outperforms an approach relying on an external reference only. Additional results will be provided at the conference.

P-17

A SNP Genotyping Study Reveals an Association of mt-ND4 11719 A/G Polymorphism with Ulcerative Colitis in Two German Cohorts

Theresa Holste[1], *Torsten Schröder*[2], *Steffen Möller*[3], *Xinhua Yu*[4,5], *David Ellinghaus*[6], *Florian Bär*[2], *Bandik Föh*[2], *Saujanya R. Ventrapragada*[2], *Kilian von Medem*[2], *Jürgen Büning*[2], *Klaus Fellermann*[2], *Hendrik Lehnert*[2], *Stefan Schreiber*[6], *Andre Franke*[6], *Christian Sina*[2], *Saleh M. Ibrahim*[3], *Inke R. König*[1]

[1]Institute of Medical Biometry and Statistics, University Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Germany; [2]Department of Medicine I, University Hospital Schleswig-Holstein, Campus Lübeck, Germany; [3]Department of Dermatology, University Hospital Schleswig-Holstein, Campus Lübeck, Germany; [4]Department of Immunology and Cell Biology, Research Center Borstel, Germany; [5]Laboratory of Autoimmunity, The Medical College of Xiamen University, China; [6]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany
Email: inke.koenig@imbs.uni-luebeck.de

Ulcerative colitis (UC) is a chronic inflammatory disorder of still unknown pathogenesis. Increasing evidence points to a pathophysiological role of altered mitochondrial respiration and thus ATP production. This might contribute to mucosal energy depletion and thereby impaired intestinal barrier function, a pathophysiological hallmark in UC. In general, genetic alterations of mitochondrial genes are a common cause of mitochondrial dysfunction. Therefore, we investigated whether mitochondrial gene variants exist in association with UC.

Here, we present data from a mitochondrial genome-wide association study of two large German cohorts including a total of 2,687 cases and 6,605 controls. As the salient finding we identified a genetic variation of the mt-ND4 gene (11719 A/G) to be associated with UC in both independent cohorts. Mt-ND4 encodes for a subunit of the mitochondrial electron transport chain complex I, which is pivotal for cellular ATP production. Therefore, genetic alterations of the mt-ND4 gene might alter the mitochondrial respiratory chain capacity and contribute to mucosal energy deficiency and UC.

P-18

Hampel's Function in Robust Logistic Regression Applied to Genetic Association Analysis

Miriam Kesselmeier[1,2], *Justo Lorenzo Bermejo*[1]

[1]Institute of Medical Biometry and Informatics, University Hospital Heidelberg, Germany; [2]Clinical Epidemiology, Center for Sepsis Control and Care, University Hospital Jena, Germany
Email: miriam.kesselmeier@med.uni-jena.de

Observations which differ from the majority of the data (outliers) can seriously influence statistical results. In high-dimensional data, the presence of outliers is expected and their handling is particularly challenging. The aim of robust statistics is to integrate outliers in the analysis by controlling their influence. Cantoni and Ronchetti proposed a robust framework for generalized linear models which has been implemented in R (package robustbase) relying on Huber's function to down-weight the influence of outliers [1, 2, 3]. Several other weighting functions are possible, such as Hampel's re-descending function. We adapted the framework to use Hampel's function to account for outliers and investigated different parameterizations of Huber's and Hampel's function. Data for a logistic regression model for disease status with age and genotype as predictors was simulated considering plausible genotyping error rates. We evaluated the false positive rate, the power, and the mean square error of parameter estimates relying on standard and robust logistic regression. We will present details of the method development, simulated data, results, and practical recommendations during the conference.

References:

1 Cantoni E, Ronchetti E: Robust inference for generalized linear models. JASA 2001;96:1022–1030.

2 R Core Team (2012). R: A language and environment for statistical computing.

3 Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, et al: robustbase. Basic Robust Statistics 2012.

P-19

A New Gene-Based Test of Association Using Rasch Models

Wenjia Wang

Pharnext, France
Email: wwang@pharnext.com

In GWAS analysis, gene-based tests of association have become an interesting alternative to the traditional single-marker association analysis as gene is the unit of the functional mechanisms. However, several statistical issues limited the performance of gene-based tests when assessed to real data. Therefore we introduce a new test to provide a p-value for a gene by using Rasch Models. Rasch Model is a mathematical framework for analyzing categories data to measure latent trait. It can provide a score for each individual in GWAS by aggregating genotypes of SNPs within a gene and the weight of each SNP. Then a p-value for each gene is generated by comparing the scores of the groups of control and case. In a series of simulations with different scenarios (number of DSL, relative risk, LD structure of genes) we compared the Rasch Models to 6 other gene-based association tests: minP, Margin test, Goeman test, GATES, SKAT, Fisher's method. The Rasch models maintain correct false positive rate in every situations. The power of Rasch Model was the highest compared to other methods when SNPs are independent, and remains in good position when SNPs are structured in blocks of Linkage Disequilibrium (LD) within a gene.

P-20

Statistical Approaches for Gene-Based Analysis: A Comprehensive Comparison Using Monte-Carlo Simulations

Carmen Dering, Inke König, Andreas Ziegler

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany
Email: ziegler@imbs.uni-luebeck.de

In recent years several studies detected associations between groups of rare variants and common diseases. These findings resulted in the development of the 'rare variant-common disease' (RVCD) hypothesis, stating that multiple rare variants together may be causal for a common disease. Therefore, many statistical tests, the collapsing methods, were developed which are the topic of this work.

We compared fourteen statistical approaches in a gene-based analysis of simulated case-control data of the Genetic Analysis Workshop (GAW) 17 in various collapsing scenarios and 200 replicates. Scenarios differed in minor allele frequency (MAF) threshold and functionality of corresponding collapsed rare variants.

Almost all of the investigated approaches showed an increased type-I-error. Furthermore, none of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data.

Irrespective of the statistic test used, collapsing methods seem to be generally useless in small case-control studies. Recent work indicates that large sample sizes and a substantial proportion of causing rare variants in the gene-based analysis can yield greater power. However, many of the investigated approaches use permutation which means high computationalcost, especially when applying a genome-wide significance level. Overcoming the issue of lowpower in small case-control studies is a challenging task for the near future.

P-21

The 'Isolation with Initial Migration' (IIM) Model: Theoretical Foundations and Computer Implementation

Rui Costa, Hilde Wilkinson-Herbots

University College London, UK
Email: ucakrjb@live.ucl.ac.uk

The 'isolation with migration' (IM) model is a common tool to detect gene flow during speciation and model the speciation process in general. Recent papers have questioned the reliability of IM model estimates, especially due to its assumption of constant gene flow until the present. In this paper, we deal with an extension to the IM model, the 'isolation with initial migration' (IIM) model. Our IIM model allows for one parent species, two descendant species, an initial period of (potentially asymmetric) gene flow between the descendant species, and a more recent period of complete isolation. We derive and describe a fast method of fitting this IIM model, or any of its nested models (including the IM model and the complete isolation model) to real data. This is a maximum likelihood method, applicable to observations on the number of segregating sites between pairs of DNA sequences from a large number of independent loci. To derive the likelihood, we define the coalescent process of a pair of sequences and solve it using eigenvectors and eigenvalues. In addition to obtaining parameter estimates, our method can also be used to distinguish between alternative models representing different evolutionary scenarios, by means of likelihood ratio tests or AIC scores. We illustrate the procedure on pairs of Drosophila sequences from approximately 30,000 loci. The fitting time for the most complex model version, using this data set, does not exceed a couple of minutes.

P-22

Detection of Copy Number Variations from Targeted Sequencing Data

Ilaria Gandin, Dragana Vuckovic

Department of Medical Sciences, University of Trieste, Italy
Email: ilaria.gandin@trieste.burlo.it

Copy Number Variations (CNVs) were proved to be involved into many genetic disorders, thus there is a huge interest in finding and validating new disease associations. Many bioinformatics tools were developed to detect CNVs from whole exome sequencing (WES) data implementing different algorithms.

Based on recently published reviews, we selected a set of software for CNV detection (including very popular ones like CONIFER, ExomeDepth and cn.MOPS) that proved to give good results on WES data. Since it is not clear if such tools perform well when applied to targeted sequencing (TS) data, we carried out a comparison analysis. We considered Ion Torrent TS data of two panels: 96 genes involved in hearing loss (HL) and 71 genes related to non-syndromic intellectual disability (ID). We processed 21 and 47 subjects respectively. Results were very different depending on the software used. The number of total CNVs called ranged from 41 to 166 for the HL panel and from 13 to 256 for the ID panel. Surprisingly only one overlap was detected. A possible explanation for this result is that tools behave differently for small target panels with respect to WES, although we cannot exclude an absence of true CNVs.

Given the increasing importance of CNVs for diagnostic purposes and the recent popularity of target sequencing panels, further investigation on this issue is needed. We plan to proceed by increasing our sample size and including subjects with CNVs diagnosed by MLPA molecular analysis in order to provide a useful guidance for this scenario.

P-23

Using HLA Imputation to Dissect the Genetic Component of Immune-Mediated Disorders: Application to Rheumatoid Arthritis

Jelmar Quist[1], Ian Scott[1,2], Sarah Spain[1], Rachael Tan[2], Sophia Steer[3], Andrew Cope[2], Cathryn Lewis[1]

[1]Department of Medical and Molecular Genetics, King's College London, UK; [2]Academic Department of Rheumatology, Centre for Molecular and Cellular Biology of Inflammation, King's College London, UK; [3]Department of Rheumatology, King's College Hospital, London, UK
Email: jelmar.quist@kcl.ac.uk

Variation in HLA genes on chromosome 6 plays a major role in susceptibility to immune-mediated disorders like rheumatoid arthritis (RA) and type 1 diabetes. Although SNP associations are highly significant in this region, classical HLA alleles and amino acid polymorphisms provide information at a functional level. HLA alleles can be imputed from SNP alleles using SNP2HLA, HLA*IMP or HIBAG. SNP2HLA also imputes amino acid changes.

RA is a heterogeneous autoimmune disease characterised by an increased autoantibody production with bone erosions as a key severity outcome measure. The HLA region plays an important role in RA susceptibility, specifically 5 amino acid positions in HLA-DRβ1. In this study, we tested the role of HLA in disease severity indexed by development of bone erosions within two years of diagnosis using 421 early active RA cases from the CARDERA trial. HLA alleles and amino acid polymorphisms were imputed using SNP2HLA and tested for association with erosion status. SNP2HLA outputs the presence/absence status of each HLA amino acid; these were converted to polymorphic amino acid-level markers and analysed using UNPHASED.

Association testing using HLA amino acid polymorphisms revealed that residue 10 in HLA-DRβ1 was significantly associated with new erosion status ($p = 0.0001$), passing the empirical p-value threshold from permutations. The presence of tyrosine at this position reduced the risk of new erosions (OR 0.44, 95% CI 0.29 – 0.71, p-value = 0.0007). No significant results were obtained from analysis of HLA alleles, indicating different HLA contributions to RA susceptibility and severity.

P-24

Resolving Variants of Unknown Significance Through Large-Scale RNA-seq

Daria Zhernakova, Patrick Deelen, Marijke van der Sijde, Joeri van der Velde, Mark de Haan, Kristin Abbott, Cisca Wijmenga, Richard Sinke, Morris Swertz, Jingyuan Fu, Lude Franke

University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
Email: dashazhernakova@gmail.com

In recent years, exome sequencing has emerged as a very effective strategy for genome diagnostics. However, the functional significance is unclear for many of the identified variants, hindering clinical interpretation. To improve upon this, we systematically assessed whether these rare variants might exert effects on gene expression levels. We hypothesized that if a variant of unknown significance is affecting gene expression, it is more likely to be pathogenic, similar to what we have observed before for common disease-associated variants (Westra et al., Nature Genetics 2013).

We therefore analyzed public RNA-seq data from over 3,000 samples. We first developed methodology to QC and harmonize the RNA-seq data and to account for differences in sequencing strategy, tissue differences and experimental perturbations. We subsequently called and imputed SNP genotypes, resulting in the availability of both genotype and expression data for each sample.

This enabled us not only to identify effects of common variants on the gene expression levels of thousands of genes (cis-eQTLs), but also to accurately estimate allele frequencies of many rare variants and identify their effects on gene expression by assessing allele specific expression (ASE). We observed that many of these rare variants known to be pathogenic indeed strongly affect gene expression levels.

Since the amount of RNA-seq data that is available in public repositories is growing exponentially, we expect our strategy will become even more powerful over time. This enables identification of downstream functional consequences for many rare variants, and will likely resolve many variants of unknown significance.

P-25

Exploiting Network Methodology for Rare Variant Association Analysis

Heide Loehlein Fier[1], *Julian Hecker*[1], *Dmitry Prokopenko*[1], *Christoph Lange*[1,2,3]

[1]University of Bonn, Working group of Genomic Mathematics, Germany; [2]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany; [3]Department of Biostatistics, Harvard School of Public Health, Boston, USA
Email: heide.fier@gmail.com

We propose a new analysis approach for rare variant analysis that exploits network methodology. The approach is designed for large genomic regions, varying effect directions, and enables the analysis of the relationships between the variants. First, we construct networks based on the allele counts of rare variants in cases and in controls. Subsequently, we introduce a test statistic that compares the derived networks of the cases and of the controls. In simulation studies, we assess the power of the methodology and compare it with standard rare variant approaches for association analysis. For scenarios with 100 or 500 loci, our approach achieves up 100% more power than standard approaches. We illustrate the practical aspect of the approach by an application to a sequencing study for nonsyndromic cleft lip with or without cleft palate.

P-26

Distinguishing Between Population Growth and Skewed Offspring Distribution Using the Site Frequency Spectrum

Bjarki Eldon[1], *Matthias Birkner*[2], *Jochen Blath*[1]

[1]Institute for Mathematics, TU Berlin, Germany; [2]Institute for Mathematics, JGU Mainz, Germany
Email: eldon@math.tu-berlin.de

We obtain recursions for the expected values and covariances of the site-frequency spectrum (SFS) associated with Lambda-coalescents. Lambda-coalescent are obtained from population models of high fecundity. The recursions for the expected values allow us to estimate parameters associated with Lambda-coalescents, and to form statistics to distinguish between the usual Kingman and Lambda-coalescents. The statistical power of simple statistics to distinguish between the Kingman coalescent and special subclasses of Lambda-coalescents is estimated using simulations. In addition we obtain recursions for expected values and covariances of the SFS associated with variable population size. To assess how well we can distinguish between population growth and Lambda-coalescents we estimate the statistical power of simple statistics based on the SFS. The results are promising provided the mutation rate is high enough.

P-27

Phenotype-Orientated Pedigree Simulation

Alexandra Gillett[1], *Ammar Al-Chalabi*[2], *Cathryn Lewis*[1,3]

[1]MRC SGDP, Institute of Psychiatry, King's College London, UK; [2]Department of Clinical Neuroscience, Institute of Psychiatry, King's College London, UK; [3]Department of Medical and Molecular Genetics, King's College London, UK
Email: alexandra.a.gillett@kcl.ac.uk

Pedigree simulation packages allow users to generate disease and linked markers under varying evolutionary pressures, allowing comparisons between competing linkage and association analysis methods. When exploring the genetic architecture of disease by comparing observed and simulated phenotypic summary measures, e.g. recurrence risk ratios (RRRs), such packages are computationally inefficient due to the unnecessary generation of non-causal markers. Here we present a software package for simulating families ascertained on a single affected proband. Family structure is simulated with the transmission of a single disease locus. Multiple common genetic loci, contributing to disease additively, can also be simulated using a polygenic threshold model. When modelling a single risk locus, the user specifies the penetrance, mode of inheritance, and risk allele frequency. For polygenic simulation, users specify a narrow-sense heritability and prevalence. The package includes a forwards algorithm, iterating from the founding generation forwards-in-time to the proband generation, useful for complex models of disease, and a backwards-in-time algorithm for computational efficiency. Summary functions are available to compute RRRs for various degrees of relatives under the pre-specified disease model.

The utility of this simulation tool is demonstrated via application to Amyotrophic Lateral Sclerosis (ALS), where dominant inheritance of incomplete penetrance mutations account for much of the genetic burden of disease. Using polygenic simulation we investigate whether the observed penetrance is incomplete due to an unobserved polygenic disease liability.

This package, and its future extensions, will be used to explore the types of genetic architecture that are consistent with observed familial patterns of ALS.

P-28

METAINTER: Meta-Analysis Tool for Multiple Regression Models

Tatsiana Vaitsiakhovich[1], *Christine Herold*[2], *Tim Becker*[2]

[1]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany; [2]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
Email: tankvait@yahoo.com

Meta-analysis of summary statistics is an essential approach to guarantee the success of Genome-wide association studies. Application of the fixed effects model to SNP-by-SNP association tests with a single parameter estimate is currently a standard practice. More complex models involving multiple parameters have been used seldom. This could be explained by the lack of a respective meta-analysis pipeline in addition to the power issue. Fisher's combination test, Stouffer's method with study-specific weights are p-values based meta-analysis methods that can be applied to any association test. However, in order to be powerful, meta-analysis methods for high-dimensional models should incorporate study-specific properties of parameter estimates, their dependence, as well as across study assessment of the consistency of effect directions. In this context, a method for the synthesis of linear regression slopes has been recently proposed in the educational science and adapted for linear regression gene-environment tests with two degrees of freedom. We elaborate this method for logistic regression and introduce an analysis tool METAINTER which implements the method for an arbitrary number of model parameters. METAINTER will enable meta-analysis of e.g. the 2-df single-SNP association test, global haplotype tests and tests for and under gene-gene interaction. The synthesis of regression slopes relies on availability of the covariance matrix of the model parameters. Since it is not always provided by genetic analysis tools, we plan to support the analysis framework presented here with an update of our own genetic interaction analysis tool INTERSNP to avoid a potential unavailability of the covariance matrix.

P-29

Analysis Pipeline for Detecting Rare Exonic Variants Associated with Adverse Drug Reactions: Application to 5FU Cohort

Jemma Walker, Monica Arenas Hernandez, Aathavan Loganayagam, Stephen Newhouse, Amos Folarin, Hamel Patel, Jeremy Sanderson, Anthony Marinaki, Paul Ross, Cathryn Lewis

King's College London, UK
Email: jemma.walker@kcl.ac.uk

Pharmacogenetic studies have strong potential for immediate translation to clinical practice, particularly for rare variants with high risk of adverse drug reactions. Exonic variants are good candidates. Here we describe an analysis pipeline for the exome array including quality control and both SNP and gene-based analysis that can identify pharmacogenetic associations.

Manual inspection of SNP cluster plots for rare variants is essential for preservation of these within the data. GWAS QC was then performed for SNPs and patients, and we used ancestry-informative principal components to correct for ancestry in our ethnically diverse cohorts. Correction for this stratification was assessed through genomic control. SNP-based association testing for common variants used logistic regression to for SNPs of frequency >0.05. Gene-level analyses for rare variants (MAF < 0.05) was then performed using an optimal sequence kernel association test (SKAT-O).

We applied this strategy to a cohort of cancer patients treated with Fluoropyrimidine 5-fluorouracil (5FU) and the prodrug-capecitabine. Adverse drug reactions (ADRs) can result in treatment discontinuation and lengthy hospitalization. Such side effects include diarrhoea, mucositis, neutropenia (DMN) and hand–foot syndrome (HFS). We identified 24 carriers of rare DPYD sequence variants (previously identified), all of which had severe DMN, and omitted these from further analysis. In the remaining 375 patients, 77 had DMN ADR. No common SNP reached Bonferroni corrected p-value for association with DMN or HFS. In the gene-level analysis, one gene reached significance (2.5e-06). It is associated with severe DMN (P = 6e-07) and with HFS in the capecitabine subgroup (P = 5.75e-05).

P-30

Novel Approaches for Detection of Splicing QTLs

Renee Menezes[1,2], *Marianne Jonker*[1], *Mark van de Wiel*[1], *Peter-Bram 't Hoen*[2,3]

[1]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands; [2]BIOS (Biobank-based Integrative Omics Studies) consortium, a consortium funded by BBMRI-NL, the Biobanking and Biomolecular Research Infrastructure in The Netherlands; [3]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
Email: r.menezes@vumc.nl

It is well-known that alternative splicing takes place, a process by which a gene may be expressed by transcripts formed with different subsets of its exons (ref). It is likely that alternative splicing events depend on SNP genotypes located within the genes exons. Here we propose a mixed-effects model to look for alternative splicing as a function of SNP genotypes. We include all SNPs mapping to the gene's exons simultaneously in the model, which has advantages compared to when individual SNPs are considered individually. Firstly, it yields more power to find subtle changes, as all available information is considered together. In addition, interaction effects between SNPs will be considered naturally by the model. The model takes the different exons corresponding to a single gene as part of a dependent variable with a multinomial distribution, which means that the model is robust to SNP effects on all exons simultaneously. This makes our model robust to eQTL effects. In addition, we correct exon expression for exon length effect with a genome-wide mixed-effects model, to avoid exon length confounding the analysis. We illustrate our results using exon sequencing data from the GEUVADIS project, and corresponding SNP genotype profiles produced by the 1000 Genomes project. We find evidence for splicing eQTLs based upon this data for a large set of genes.

P-31

PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R

Bastian Pfeifer[1], *Ulrich Wittelsbürger*[1], *Sebastian Ramos Onsins*[2], *Martin Lercher*[1,3]

[1]Institute for Computer Science, Heinrich Heine University Düsseldorf, Germany, Germany; [2]Centre for Research in Agricultural Genomics, Bellaterra, Spain; [3]Cluster of Excellence on Plant Sciences, Düsseldorf, Germany
Email: Bastian.Pfeifer@uni-duesseldorf.de

While many computer programs can perform population genetics calculations, they are typically limited in the analyses and data input formats they offer; few applications can process the large datasets produced by whole-genome resequencing projects. Furthermore, there is no coherent framework for the easy integration of new statistics into existing pipelines, hindering the development and application of new population genetics and genomics approaches.

PopGenome is a population genomics package for the R software environment (a de-facto standard for statistical analyses). PopGenome can efficiently process genome-scale data as well as large sets of individual loci. It reads DNA alignments and SNP datasets in most common formats, including those used by the HapMap, 1000 human genomes, and 1001 Arabidopsis genomes projects. PopGenome also reads associated annotation files in GFF format, enabling users to easily define regions or classify SNPs based on their annotation; all analyses can also be applied to sliding windows. PopGenome offers a wide range of diverse population genetics analyses, including neutrality tests as well as statistics for population differentiation, linkage disequilibrium, and recombination. PopGenome is linked to Hudson's MS and Ewing's MSMS programs to assess statistical significance based on coalescent simulations. PopGenome is freely available from CRAN (http://cran.r-project.org/) for all major operating systems under the GNU General Public License.

P-32

New Genetic Matching Methods for Handling Population Stratification in Association Studies

André Lacour

German Center for Neurodegenerative Deseases (DZNE), Bonn, Germany
Email: andre.lacour@dzne.de

A usually confronted problem in association studies is the occurrence of population stratification. In this talk, we propose a decisive extension to the Cochran-Armitage Trend test in order explicitly take into account structures obtained from matchings and clusterings. Based on earlier work, we employ pairwise and groupwise optimal case-control matchings and present an agglomerative hierarchical clustering, both based on a genetic similarity score matrix. By simulations of genotype data under the null hypothesis we assess our framework, in order to affirm that it correctly controls for the type-1 error rate. By a power study we ascertain, that structured association testing using our framework displays reasonable power. We compare the results from our methods with those obtained from a logistic regression model with principal component covariates. We also highlight and discuss a possible false-positive association to Alzheimer's disease using the principal components approaches, which is neither reproduced by our new methods nor by the results of a most recent large meta analysis.

P-33

Power Considerations in Pharmacogenomic Studies

Seth Seegobin, Cathryn Lewis

King's College London, UK
Email: seth.seegobin@kcl.ac.uk

Pharmacogenomics is the study of how genetic variation among individuals contributes to inter-patient variation in efficacy and safety response to drug therapy. In contrast to complex disease studies, genetic variants within pharmacogenomics tend towards large effect sizes with respect to drug response. However as most data are sourced from clinical trials, obtaining an adequate number of cases is often a challenge. Consequently pharmacogenomic studies tend to be smaller as compared with complex disease association studies and depending on the effect size the loss or gain of a few samples can have a large impact on the power to detect associations. Starting from known allele frequencies in publicly available control datasets researchers can a priori determine the number of cases needed to reliably detect associations at a given significance threshold. In this study we explored the impact that effect size and the proportion of cases vs. controls within a modest sample size have upon the power to detect associations at α. We compared the power estimates and derivation from two common genotype-based analysis techniques: the Cochran-Armitage trend test and binary logistic regression with an ordinal predictor. Further we investigated the minimum number of subjects required to obtain a power of 1 – β with a significance level of α across a range of control risk allele frequencies and case-control proportions.

P-34

Estimating Global Individual Ancestry Using Principal Components for Family Data

Mariza Andrade[1], *Debashree Ray*[2], *Julia Soler*[3]

[1]Mayo Clinic, Rochester, USA; [2]University of Minnesota, Minneapolis, USA; [3]University of Sao Paulo, Sao Paulo, Brasil
Email: mandrade@mayo.edu

Studies of human complex diseases and traits associated with candidate genes are potentially vulnerable to bias (confounding) due to population stratification and inbreeding, especially in admixture population. In genome-wide association studies (GWAS) the Principal Components (PCs) method provides a global ancestry value per subject, allowing corrections for population stratification. However, these coefficients are typically estimated assuming unrelated individuals and if family structure is present and it is ignored, such sub-structure may induce artifactual PCs. Extensions of the PCs method have been proposed by Konishi and Rao (1992) taking into account only siblings relatedness and by Oualkacha et al. (2012) taking into account large pedigrees and high dimensional data. In this work we applied these methods to estimate the global individual ancestry including PCs extracted from different variance components matrix estimators. We applied these methods to the GENOA sibship data consisting of European and African American subjects and the Baependi Heart Study consisting of 80 extended families collected from the highly admixture Brazilian population, both data with genotyping data from Affymetrix 6.0 chip. Our results showed that the family structure plays an important role in the estimation of the global individual ancestry for extended pedigrees but not for sibships. All the analyses were performed using R package.

P-35

Quantifying the High-Dimensional Information of Population Structure

Omri Tal

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
Email: omrit1248@gmail.com

I describe a framework for deriving measures of 'informativeness' from multi-locus genetic data. Specifically, given a collection of genotypes sampled from known multiple populations I would like to quantify the potential for correct classification of genotypes of unknown origin, or alternatively, provide a measure of data 'clusteredness'. Motivated by Shannon's axiomatic approach in deriving a unique information measure for communication, I identify a set of intuitively justifiable criteria that any such quantitative information measure should satisfy, where the notion of communication noise can be made analogous to sampling noise. I will show that standard information-theoretic measures such as mutual information or relative entropy cannot satisfactorily account for this sense of information, necessitating methods from statistical-learning. I will also review very recent empirical work of biologists to assess the 'population signal' from genetic samples.

P-36

The Effect of Phenotype Transformations on Statistical Power in Genome-Wide Association Studies

Pimphen Charoen

London School of Hygiene and Tropical Medicine, UK
Email: pimphen.charoen@lshtm.ac.uk

Statistical power in genome-wide association studies (GWAS) has increased substantially in recent years owing to the use of increasingly large sample sizes. However, there has been little investigation into which phenotype transformations maximise statistical power when searching for genetic variants affecting continuous traits. Here we use both simulated and real data to investigate how the log transformation (LT), inverse normal transformation (INT) and no transformation (NT) affect power in the subsequent linear regression. For the small effect sizes common in GWAS, we observe that the INT tends to lead to greater power than the LT and has markedly higher power for phenotypes with low skew and high coefficient of variation (CV = standard deviation/mean) and vice versa. Using real data from the Northern Finland Birth Cohort 1966 (NFBC1966), we identify twice as many reported genotype-phenotype associations using the INT than the LT among phenotypes with low CV and high skew (such as BMI, Insulin, HOMA-B). We also investigate the effect of these transformations when used within a meta-analysis framework. We found that while the LT can lead to a substantial loss in power in the presence of between-cohort heterogeneity of CV when using an inverse variance meta-analysis, the INT is robust to this heterogeneity regardless of the meta-analysis method used. Our findings can be exploited to determine which transformation should be applied given the distribution of the phenotype under investigation in order to maximise discovery in GWAS.

P-37

New Software and Developments in the GenABEL Project

Lennart Karssen, on behalf of the GenABEL team

Email: l.c.karssen@gmail.com

In the last year the GenABEL project has seen a considerable number of improvements. These improvements do not only consist of updates to the packages in the GenABEL suite, but we also improved the development process and the way packages are made available to our users.

We will demonstrate newly implemented features in the GenABEL suite packages, as well as introduce OmicABEL, a package for rapid mixed-model based genome-wide association analysis of multiple traits (for example, metabolomics, glycomics, etc.).

Recently we started using the Jenkins Continuous Integration server to help us release software of higher quality. Jenkins is a framework that automatically runs several tests (e.g. static code analysis, checks for memory leaks) after a new commit to our version control system. This allows us to detect problems in the code at an early stage, before they bug the user.

After GenABEL, ProbABEL is the second package that is available as a Debian package, allowing users of upcoming Debian releases to install and upgrade ProbABEL with a single command. This also benefits users of Linux distributions derived from Debian, like Ubuntu.

In the coming year more packages are expected to be added the GenABEL suite as well as continued efforts to improve the existing ones. Moreover, we plan to increase both the ease of installation as well as the visibility of the GenABEL suite by adding more packages into both the Debian and Red Hat Linux repositories (as well as derivatives like CentOS and Scientific Linux).

P-38

Genetic Determinants of mRNA Processing

Peter 't Hoen[1], Daria Zhernakova[2], Eleonora de Klerk[1], Martijn Vermaat[1], Renee de Menezes[3], Tuuli Lappalainen[4], Lude Franke[2]

[1]Department of Human Genetics, Leiden University Medical Center, The Netherlands; [2]Department of Genetics, University of Groningen, University Medical Center Groningen, The Netherlands; [3]Department of Epidemiology and Biostatistics, VU University Medical Center Amsterdam, The Netherlands; [4]New York Genome Center and Department of Systems Biology, Columbia University, New York, USA
Email: p.a.c.hoen@lumc.nl

Recent large-scale studies of expression quantitative trait loci (eQTL) have shown that the expression level of the majority of human genes is influenced by common genetic variants present in the population. The introduction of RNA-seq technologies enabled the study of expression at a more fine grained resolution than previously used microarrays. Genetic variants appear to have an equally strong impact on the relative expression of transcripts from the same gene. However, the methods for the identification of QTLs that influence splicing, polyadenylation and other RNA processing events leading to the formation of different transcripts, are still under development. Here, we will review the currently available methods and demonstrate our recent improvements on pipelines for the detection of the QTLs for RNA processing. Insight into the genetic determinants of RNA processing will further increase our understanding of the post-transcriptional control of gene expression.

The authors would like to acknowledge funding by BBMRI-NL, the Biobanking and Biomolecular Research Infrastructure in The Netherlands. PACtH, DVZ, MV, RXdeM and LF are part of the BIOS (Biobank-based Integrative Omics Studies) consortium, funded by BBMRI-NL.

P-39

A Novel Method to Prioritise Genetic Loci Identified in GWAS for Sequencing

Paul O'Reilly[1], Clive Hoggart[2]

[1]MRC SGDP Centre, Institute of Psychiatry, King's College London, UK; [2]Dept. of Genomics of Common Disease, Imperial College London, UK
Email: paul.oreilly@kcl.ac.uk

Genome-wide association studies (GWAS) have identified thousands of genotype phenotype associations, but the usual assumption is that these highlight genetic regions harbouring causal variants rather than revealing the causal variants themselves. Sequencing such susceptibility loci makes detecting the causal variants feasible, crucial for translating the findings from GWAS into medical application. However, most genetic loci identified by GWAS have not been interrogated for the causal variants via sequencing, largely due to the cost involved, and when they have it has tended to be on an 'ad hoc' basis according to the resources of individual studies or the significance of association signals at certain loci. Here, we introduce a novel method that estimates the posterior probability of the allele frequency of putative causal variants in a genetic region, which can be used to determine whether subsequent sequencing is likely to reveal the causal variant or whether it is more likely already present among the genotyped or imputed SNPs. The method exploits the SNP GWAS association P values, their minor allele frequencies and local Linkage Disequilibrium (LD), to calculate the likelihood that the data are consistent with a causal variant(s) of a given effect size and frequency. While the probability computed is dependent on certain modelling assumptions, we show how our method can be used to prioritise the sequencing of genetic loci known to influence human traits or diseases based on GWAS results.

Human
Heredity

Author Index

Numbers refer to page numbers

KARGER

E-Mail karger@karger.com
www.karger.com/hhe