Mathematisches Forschungsinstitut Oberwolfach

Report No. 52/2015

DOI: 10.4171/OWR/2015/52

Mini-Workshop: Recent Developments in Statistical Methods with Applications to Genetics and Genomics

Organised by
Iuliana Ionita-Laza, New York
Michael Krawczak, Kiel
Xihong Lin, Harvard
Michael Nothnagel, Köln

8 November – 14 November 2015

Abstract. Recent progress in high-throughput genomic technologies has revolutionized the field of human genetics and promises to lead to important scientific advances. With new improvements in massively parallel biotechnologies, it is becoming increasingly more efficient to generate vast amounts of information at the genomics, transcriptomics, proteomics, metabolomics etc. levels, opening up as yet unexplored opportunities in the search for the genetic causes of complex traits. Despite this tremendous progress in data generation, it remains very challenging to analyze, integrate and interpret these data. The resulting data are high-dimensional and very sparse, and efficient statistical methods are critical in order to extract the rich information contained in these data. The major focus of the mini-workshop, entitled *"Recent Developments in Statistical Methods with Applications to Genetics and Genomics"*, has been on integrative methods. Relevant research questions included the optimal study design for integrative genomic analyses; appropriate handling and pre-processing of different types of omics data; statistical methods for integration of multiple types of omics data; adjustment for confounding due to latent factors such as cell or tissue heterogeneity; the optimal use of omics data to enhance or make sense of results identified through genetic studies; and statistical and computational strategies for analysis of multiple types of high-dimensional data.

Mathematics Subject Classification (2010): 62H30, 62J12, 92D20, 92D30, 92B15.

Introduction by the Organisers

The mini-workshop *"Recent Developments in Statistical Methods with Applications to Genetics and Genomics"*, organized by Iuliana Ionita-Laza (New York), Michael Krawczak (Kiel), Xihong Lin (Harvard), Michael Nothnagel (Köln), was attended by 16 participants with broad geographic representation from North America and Europe. This workshop was interdisciplinary, and had a nice blend of junior and senior researchers with diverse backgrounds in theoretical/applied statistics, and genomics. The small scale and focused meeting has allowed for plenty of time for discussions and brainstorming new ideas, and has started several new collaborative projects. During the week, 15 lectures have been given by the participants. The lectures were accompanied by lively and interesting discussions. This report contains extended abstracts of all the talks.

The major focus of the mini-workshop has been on the efficient integration of different sources of data to gain a better understanding of the genetic mechanisms that lead to complex diseases. Multiple omics data (genome, epigenome, transcriptome, proteome, metabolome, phenome) can now be easily collected simultaneously on a genome-wide scale, yet remarkably little is known about how to integrate these different data types in a knowledge-based way. Integrative analysis of multiple omics data types can help the search for the underlying biological mechanisms in disease by discovering genomic features that tend to be dysregulated by multiple mechanisms. Because many of these technologies only recently became feasible on a genomic scale, the data are only now becoming available on a large scale, making the timing of this workshop ideal for sharing the emergent results, and for beginning to address the many challenges associated with these complex sources of data. New statistical methods have been discussed to make full use of the multi-source data for clustering, classification and prediction. Most integrative methods do not take into account known biological relations between different data sources. For example, there are well known regulatory relations between genomic data sets; e.g. gene expression levels can be regulated by both genetic aberrations and epigenetic factors. Integrating multiple data sets without accounting for their intrinsic relationships may unnecessarily increase the degrees of freedom in data and fail to contribute new information to existing variables. Therefore, new methods are needed to address these issues and others. Additional relevant research questions that have been addressed during the workshop include causal inference methods to identify causal mechanisms in disease, adjustment for confounding due to latent factors, the optimal use of omics data to enhance interpretation of results of genome-wide association studies (GWAS) and the integration of multiple GWAS datasets on different, correlated phenotypes.

Overall, the topics of the workshop are very important, and the talks and discussions have attempted to provide a survey of the current state of the field, and to explore new ideas and directions for future data integration approaches. The organizers would like to thank the Institute staff for providing such a great environment for our meeting.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows".

Mini-Workshop: Recent Developments in Statistical Methods with Applications to Genetics and Genomics

Table of Contents

Abstracts

The role of joint cumulants in genetic analysis

Stefan Böhringer
(joint work with Brunilda Balliu, Eleni Karasami)

1. Introduction

Linkage disequilibrium (LD) - the covariance between allele indicators at genetic markers - plays an important role in genetics. Joint cumulants of the multivariate Bernoulli distribution can be viewed as a natural generalization of LD to more than two loci. In this paper, we characterize properties of such joint cumulants and relate them to the genetic situation. Important findings include the relationship with the corresponding multinomial distribution, bounds of joint cumulants, and the partitioning of sets of Bernoulli variables. Some earlier work has considered generalization of LD to few loci [2, 4]. A generalization to K markers was proposed with different properties to the current paper [1].

2. Reparametrization

We assume all random variables (RVs) to be Bernoulli throughout. Let $\mathcal{N} = \{1, ..., N\}$. We then have:

Definition: $X = X_{\mathcal{N}} = (X_1, ..., X_N)$ multivariate RV. The joint cumulant of X, $\delta_X = \delta_{\{1,...,N\}}$ is defined as $\delta_{\mathcal{N}} := \sum_\pi (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{B\in\pi} E\left(\prod_{i\in B} X_i\right)$, where π iterates all partitions of $\mathcal{N}$.

If we denote with $\eta_S := E\left(\prod_{i\in S} X_i\right)$, cumulants can be expressed as: $\delta_{1,...,N} = \sum_\pi (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{B\in\pi} \eta_B$. Here, η_B can be seen as the marginal haplotype frequency of the haplotype composed of alleles 1 at loci in B. If θ_i denotes $P((X_1, ..., X_n) = i_2)$, where i_2 is the binary representation of i, the following lemma can be established.

Lemma: Reparametrization. Let $\theta = (\theta_0, ..., \theta_{2^N-1}) \in H = \{(0, 1)^N | \sum \theta_i = 1\}$ be such that $P((X_1, ..., X_n) = i_2) = \theta_i$, then the mapping $\phi : H \to \phi(H), \theta \to \delta = (\delta_{S_1}, ..., \delta_{S_{2^N}})$ is bijective, δ_{S_i} is the cumulant of RVs $\emptyset \neq S_i \subset \mathcal{N}$.

θ is interpreted as multinomial frequencies for the different outcomes of X. In contrast to θ, δ contains a structure, namely the number of variables/loci involved in the respective cumulant. This structure can be exploited by formulating null hypotheses in the space of cumulants instead of haplotype frequencies. By limiting the comparison of cumulants to those involving only few loci (say one or two), degrees of freedom can be reduced and interpretability can be improved as compared to the comparison of the full distributions. Yet the likelihood theory is easy to establish using the reparametrization. We explore several testing strategies in simulations and a data application that show that power is increased substantially in certain situations.

3. Standardization

In the pairwise case LD that is standardized to minimal and maximal bounds, has a genetic interpretation, namely that a possible haplotype is missing from the distribution if the bounds are hit. This in turn relates to the occurrence of genetic recombinations. We can establish sharp bounds for multivariate cumulants.

Lemma: (Missing haplotypes vs cumulants) For $S \subset \mathcal{N}$, let δ_S be cumulant of $X_S = (X_{i_1}, ..., X_{i_s})$ and δ'_S be the corresponding standardized cumulant $\in [0, 1]$. We then have

$\exists S : \delta'_S \in \{0, 1\} \Rightarrow \exists t : \theta_t \in \{0, 1\}$,

which generalizes the pair-wise result to general join cumulants. θ_t again denotes corresponding multinomial frequency of haplotype t. In the multivariate setting much richer missingness patterns are possible as compared to the pairwise situation. The above lemma guarantees at least one missing haplotype when the joint cumulant hits the bounds, however there might be many missing haplotypes.

We investigate missingness patterns by enumerating sequences of genetic events composed of mutation (flipping the state of a marginal variable) and recombination (exchanging subsets of variables between pairs of joint Bernoullis). We compute a corresponding haplotype distribution by iteratively starting with a single - ancestral - haplotype and applying all possible sequences. This produces existence patterns of haplotypes for which a uniform distribution is assumed corresponding to recombination equilibria. This distribution is expressed in terms of standardized cumulants. After removing allele frequencies (cumulants of marginal variables) we call this vector a cumulant signature and associate it with the possible histories producing this signature. These signatures can be used to visualize closeness of actual data to genetic histories. Potential applications lie in the analysis of population stratification and a descriptive analysis of the sampling process. Methods are illustrated using HapMap data.

4. Genome Partitioning

Partitioning the genome into sets of independent markers plays an important role in genetic applications. For example, makers used in linkage analysis, or population stratification analysis based on principle component based or explicit models assume sets of independent markers. Also multiple testing correction in the context of genome wide association studies would be facilitated by such a partitioning.

It is well known that $X_{S_1} \perp X_{S_2}$ with $S = S_1 \dot{\cup} S_2$ implies $\delta_S = 0$ [3]. The converse, however, is untrue in general and it is straightforward to construct a counterexample. A sufficient criterion for independence based on joint cumulants is given by the following lemma.

Lemma: (Sufficient criterion for independence)
$X_{\mathcal{N}}$ MV Bernoulli; $S = \mathcal{N} = S_1 \dot{\cup} S_2; S_1, S_2 \neq \emptyset$.
$X_{S_1} \perp X_{S_2} \Longleftrightarrow (\forall T : T \cap S_1 \neq \emptyset \neq T \cap S_2 \Rightarrow \delta_T = 0)$

In words, whenever a subset of RVs is considered that contains at least one variable from both sets S_1 and S_2, respectively, the joint cumulant has to be zero.

This lemma suggests a family of test statistics that allow to test for indepenence in the framework of equivalence testing.

$$T = \min_{S_1 \dot{\cup} S_2 = S} \sum_{S | S \cap S_1 \neq \emptyset \neq S_2 \cap S_1} \lambda_S \delta_S^2.$$

The hypotheses are given by

$$H0 : T \leq \epsilon \qquad \text{vs.} \qquad H1 : T < \epsilon$$

Here, λ_S are arbitrary positive constants and S_1, S_2 iterate all possible 2-partitions of S. At this moment, the asymptotic distribution of T is unclear as well as the choice of optimal values for the λ_S. Both problems are subject to future research. Finally, T is expensive to compute which poses algorithmic challenges.

References

[1] R. Gorelick, M.D. Laubichler, *Decomposing Multilocus Linkage Disequilibrium*, Genetics **166** (2004), 1581–1583.

[2] A. Hastings, *Linkage Disequilibrium, Selection and Recombination at Three Loci*, Genetics **106** (1984), 153–164

[3] P. McCullagh, J. Kolassa, *Cumulants*, Scholarpedia **4** (2009), 4699

[4] B. Weir, *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, Sinauer Associates (1996)

Moving beyond genome-wide association studies through the modelling of more complex mechanisms

Heather J. Cordell

Over the past 8 or 9 years, genome-wide association studies (GWAS) have been exatrordinarily successful at identifying genetic variants associated with common, complex disorders. However, a typical GWAS gives little insight into the underlying biological mechanism through which the associated genetic variants are implicated in disease. I outline two strategies that we have been exploring to help elucidate the underlying causal mechanisms leading to an observed association. I outline the methodological approaches we have been taking in relation to both strategies and present the results of computer simulations and real data analyses illustrating the utility of these approaches. One strategy has been through the development of methods for detection of parent-of-origin effects [1]. Parent-of-origin effects, particularly if mediated through mechanisms such as imprinting, represent a more complex, potentially functionally relevant finding than the genetic effects that are typically identified through large-scale case/control GWAS. The requirement for parental data necessarily limits the power of studies designed to detect such effects, however suitable cohorts (particularly of mother/child duos) are often collected, for example, when investigating traits related to pregnancy complications. Genetic variants identified through such investigations still represent the first step along the causal pathway to disease development, and the second

strategy we have been exploring attempts to clarify the underlying causal mechanisms through modelling relationships between genetic factors, factors that are potential mediators (such as DNA methylation and gene expression), and disease outcomes. We focus on methods that assume at least a proportion of subjects will have measurements on all variables (genetic data, "omics" measures such as DNA methylation and gene expression, and variables related to disease phenotype) of interest. Previous studies using such data types [2, 3] have used a filtering strategy to generate triplets of 'interesting' variables corresponding to a genetic variant such as a single nucleotide polymorphism (S), a phenotype of interest (P), and an intermediate trait such as DNA methylation or gene expression (G), based on their pairwise correlations. The resulting triplets are then subjected to a causal inference test such as the 'causal inference test' (CIT) or else are interrogated using techniques such as structural equation modelling or Mendelian randomization to infer the underlying causal structure. We use computer simulations to investigate the performance of such approaches (as well as alternative approaches based on Bayesian Networks or a Bayesian Unified Framework) when the underlying causal structure is known. We find that all methods perform well in simple models where their assumptions are not violated. However, the presence of an unknown/unmeasured common environmental effect can lead to incorrect inference.

References

[1] R. Howey, C. Mamasoula, A. Töpf, R. Nudel, J.A. Goodship, B.D. Keavney, H.J. Cordell, *Increased power for detection of parent-of-origin effects via the use of haplotype estimation*, Am J Hum Genet **97** (2015), 419–434.

[2] Y. Liu, M.J. Aryee, L. Padyukov, M.D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T.J. Ekström, A.P. Feinberg, *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis*, Nat Biotechnol **31** (2013), 142–147.

[3] S.Y. Shin, A.K. Petersen, S. Wahl, G. Zhai, W. Römisch-Margl, K.S. Small, A. Döring, B.S. Kato, A. Peters, E. Grundberg, C. Prehn, R. Wang-Sattler, H.E. Wichmann, M.H. de Angelis, T. Illig, J. Adamski, P. Deloukas, T.D. Spector, K. Suhre, C. Gieger, N. Soranzo, *Interrogating causal pathways linking genetic variants, small molecule metabolites, and circulating lipids* Genome Med **6** (2014), 25.

Integration of biological knowledge, SNP and omics data for gene discovery in multifactorial diseases.

Florence Demenais

Genome-wide association studies (GWASs) have been highly effective in identifying thousands of genetic variants associated with many diseases or traits. However, these variants explain only a part of the genetic component of these diseases (traits). Disease susceptibility is likely to result from the joint and potentially interactive effects of many genetic factors, each making a small contribution to overall disease risk, and the effect of such factors may be missed if they are examined individually as classically done by GWAS. The analysis of the joint effect of

multiple SNPs and their interactions on disease risk together with the integration of biological knowledge can facilitate the discovery of novel genetic factors. We proposed different strategies of data integration that were illustrated in the field of cancer and asthma.

A first strategy is to combine pathway analysis of GWAS outcomes and gene-gene interaction analysis within disease-associated pathways. Pathway analysis based on the Gene Set Enrichment Analysis (GSEA) [1] allows to identify pathways enriched in genes associated with disease. These pathways can be used as statistical and biological filters to investigate cross-gene SNP-SNP interactions within pathways. A major advantage of this approach is to reduce the multiple testing burden as compared to genome-wide gene-gene interaction (GWIS) studies. We proposed a hierarchical bottom-up procedure to correct for multiple testing. We first corrected for multiple interaction tests for each gene pair, then for multiple gene pairs within a pathway, and finally across all disease-associated pathways. Specifically, for each gene pair, the effective number of independent interaction tests was estimated from the eigenvalues of the correlation matrix of pairwise products of SNPs allele dosages (imputed SNPs), by extending the method proposed by Li and Ji [2] for correlated single SNPs to correlated SNP-pairs interactions. The effective number of independent tests in a pathway was estimated by the sum of the effective number of independent tests for a gene pair over all gene pairs tested within that pathway; from this, we computed a Bonferroni-corrected critical threshold for a pathway (Tpathway). To correct for overall statistical significance across all disease-associated pathways, a Bonferroni-correction was applied to the pathway-corrected threshold (Tpathway) to get the overall critical threshold (Toverall). For example, for a set of about 1 million SNPs, this procedure reduced the multiple testing corrected threshold from 5×10^{14} for an agnostic GEWIS to between 3×10^{-7} and 7×10^{-9} depending on the size of the pathway. The combined pathway and gene-gene interaction analysis strategy was applied to GWAS outcomes for cutaneous melanoma obtained from two datasets: the French MELARISK dataset (1179 cases, 2797 controls) that served as a discovery set and the MD Anderson Cancer Center (MDACC) dataset (1801 cases, 1026 controls) that served as a replication set. Five pathways defined by gene ontology (GO) categories were significantly enriched in genes associated with melanoma (FDR$<$ 5% in both studies): response to light stimulus, regulation of mitotic cell cycle, induction of programmed cell death, cytokine activity and oxidative phosphorylation. Epistasis analysis, within each of the five significant GOs, showed significant evidence for interaction for one SNP pair at TERF1 and AFAP1L2 loci ($P = 2.0 \times 10^{-7}$, in the meta-analysis of the two datasets, which met both the pathway and overall multiple-testing corrected thresholds that were equal to 9.8×10^{-7} and 2.0×10^{-7}, respectively) and suggestive evidence for another pair involving correlated SNPs at the same loci ($P = 3.6 \times 10^{-6}$). This interaction has important biological relevance given the key role of TERF1 in telomere biology and the reported physical interaction between TERF1 and AFAP1L2 proteins [3].

This study clearly shows the advantage of using a statistical and biological filtering to identify gene-gene interactions.

A second strategy is to conduct network-based analysis by integrating genome-wide SNP data and protein-protein interaction networks (retrieved from the Human Protein Interaction Network (HPIN) database) to identify a gene sub-network associated with disease. We proposed an algorithm that was applied to two large asthma datasets from the GABRIEL Asthma Consortium that consisted of the outcomes of two meta-analyses of 9 childhood asthma GWASs each (including 3,031 cases/2,893 controls and 2,679 cases/3,364 controls, respectively) [4]. GWAS signals were overlaid to HPIN by assigning SNPs to genes and using gene-wise P-values obtained through circular genomic permutations (CGP) [5]. Modules enriched with childhood asthma-associated genes were generated by a dense module search (DMS) strategy [6]. We selected the gene modules that showed the highest pairwise similarity between the two datasets. These modules were further evaluated for their association with asthma using CGP and for their biological relevance through pathway analysis using DAVID. We identified 10 gene-module pairs that had high similarity between the two datasets. By merging the selected modules within each dataset and intersecting the two gene lists, we identified a sub-network consisting of 91 genes and 106 connections among them. Among these genes, 14 were reported associated with asthma by previous GWASs and 22 with nominally significant gene-wise P-values were novel candidates. The identified sub-network was significantly associated with childhood asthma ($P < 10^{-4}$ using 10,000 CGPs). Moreover, the number of connections among known and novel candidate genes was significantly higher than expected by chance ($P = 3 \times 10^{-4}$). Three KEGG pathways were found significantly enriched in genes from the identified network: cytokine-cytokine receptor interaction (Bonferroni-corrected $P = 3 \times 10^{-8}$), chemokine signaling pathway (Bonferroni-corrected $P = 5 \times 10^{-8}$), natural killer cell mediated cytotoxicity (Bonferroni-corrected $P = 3 \times 10^{-6}$). This study shows the benefit of integrating GWAS data and HPIN to identify novel functionally related genes underlying childhood asthma [7].

A third strategy is to integrate SNP data and epigenomic data (DNA methylation levels) to uncover the causal mechanism underlying SNP-disease association. This strategy was applied to family data of the co-morbidity of asthma plus allergic rhinitis (AAR). Following a genome-wide linkage scan of AAR in 615 European families that detected linkage to the 4q31 region in presence of parent-of-origin effect (paternal linkage), association analysis with 1,233 single nucleotide polymorphisms (SNPs) covering the significant linkage region was conducted in 162 French families from the Epidemiological study on the Genetics and Environment of Asthma (EGEA) with replication in 154 French-Canadian families from the Saguenay-Lac-Saint-Jean asthma study (SLSJ). Association analysis in this region showed strong evidence for the effect of a paternally inherited allele of the rs10009104 SNP on AAR ($P = 1.1 \times 10^{-5}$, reaching the multiple-testing corrected threshold). Further association analysis of disease and significant SNPs with DNA

methylation levels (DNAm) at CpG sites was performed in 40 SLSJ families. The paternally inherited allele of rs10009104 was significantly associated with DNAm at cg02303933 site ($P = 1.7 \times 10^{-4}$). Causal Inference Test (CIT) showed that differential DNA methylation at this site mediated the identified SNP-AAR association. The CpG site is located within Melatonin receptor 1A (MTNR1A) gene, a receptor for melatonin which was suggested to have immunomodulatory effect in allergic diseases [8] and is thus a relevant candidate for AAR.

In conclusion, joint analysis of sets of SNPs (genes) and integration of biological knowledge and omics data can facilitate the identification of genes associated with complex diseases and can shed light on the underlying molecular mechanisms. However, pathway and network-based analyses only use information on genetic variants that are mapped to genes. There is accumulating evidence that many genetic variants associated with multifactorial diseases map to regulatory elements that reside outside of genes. A major challenge is to define regions enriched in regulatory elements across the genome that could be further grouped as tissue-specific regulatory pathways and potentially used as classical gene-based pathways and networks to help identifying the genomic variation underlying complex diseases. Further work should also aim at combining various types of omics data and other sources of information (e.g. data from the literature through text mining) as well as data on environmental factors to bring more insight into the mechanisms causing multifactorial diseases.

References

[1] Wang K, Li M, Bucan M. *Pathway-based approaches for analysis of genomewide association studies*. Am J Hum Genet 2007, 81:1278-1283.

[2] Li J, Ji L. *Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix*. Heredity 2005, 95: 221-227.

[3] Brossard M, Fang , Vaysse A, Wei Q, Chen WV, Mohamdi H, Maubec E, Lavielle N, Galan P, Lathrop M, Avril MF, Lee JE, Amos CI, Demenais F. *Integrated pathway and epistasis analysis reveals interactive effect of genetic variants at TERF1 and AFAP1L2 loci on melanoma risk*. Int J Cancer, 2015, 137(8):1901-1909.

[4] Moffat M, Gut Y, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO; GABRIEL Consortium. *A large-scale, consortium-based genomewide association study of asthma*. N Engl J Med. 2010, 363(13):1211-1221.

[5] Cabrera CP, Navarro P, Huffman JE, Wright AF, Hayward C, Campbell H, Wilson JF, Rudan I, Hastie ND, Vitart V, Haley CS. *Uncovering networks from genome-wide association studies via circular genomic permutation*. G3 (Bethesda) 2012, 2(9):1067-1075.

[6] Jia P, Zheng S, Long J, Zheng W, Zao Z. *dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks*. Bioinformatics. 2011, 27(1):95-102.

[7] Liu Y, Brossard M, Sarnowski C, Margaritte-Jeannin P, Llinares F, Vaysse A., Dizier MH, Bouzigon E, . Demenais F. *Integration of genome-wide association data and human protein interaction networks identifies a gene sub-network underlying childhood-onset asthma*. Presented at Annual Meeting of the American Society of Human Genetics (ASHG 2015), Oct 6-10 2015, Baltimore, USA; abstract at http://www.ashg.org/2015meeting/pages/onlineplanner.shtml

[8] Sarnowski C, Laprise C, Malerba G, Moffatt M, Dizier M-H, Morin A, Vincent Q, Rohde K, Esparza-Gordillo J, Margaritte-Jeannin P, Liang L, Lee Y-A, Siroux V, Bousquet J, Pignatti

P.F, Cookson W.O.C, Pastinen T, Lathrop M, Demenais F, Bouzigon E. *DNA methylation within MTNR1A mediates paternally transmitted genetic variant effect on asthma-plus-rhinitis*. J Allergy and Clin Immunol, 2015 (in press).

Assessing Cross-Phenotype Effects of Rare Variants

Michael Epstein
(joint work with K. Alaine Broadaway)

Increasing empirical evidence suggests that many genetic variants influence multiple distinct phenotypes. When cross-phenotype effects exist, multivariate association methods that model pleiotropy are often more powerful than univariate methods that model each phenotype separately. Cross-phenotype association tests for common variants have demonstrated considerable success, with novel findings in studies of Crohn's disease and ulcerative colitis, different facial morphology measures, and among bipolar disorder, autism spectrum disorder, major depressive disorder, and schizophrenia. However, while several statistical approaches exist for testing pleiotropy for common variants, there is a lack of cross-phenotype tests for gene-based analysis of rare variants. In this talk, we created such a non-parametric test of independence between a high-dimensional set of phenotypes and a high-dimensional set of rare-variant genotypes in a candidate gene of interest. Our independence test relies on kernel-distance covariance (KDC) techniques that compare pairwise similarity in multivariate phenotypes to pairwise similarity in multivariate genotypes. Our approach allows for both continuous and categorical phenotypes and can further adjust for influential covariates, such as principal components of ancestry to correct for confounding due to population stratification, by residualizing variables prior to analysis.

We show that, under the null hypothesis of independence, our KDC-based test follows a mixture of chi-square variables with the mixture weights a function of the product of the eigenvalues of the phenotype and genotype similarity matrices. We further show we can derive an analytic p-value for the cross-phenotype test quickly using Davies' exact method. This is important, as it enables rapid cross-phenotype testing of rare variants across the genotype. By employing Davies' exact method and using computational shortcuts to calculate the non-zero eigenvalues from phenotype and genotype similarity matrices in quick fashion, we find that evaluation of our test for sample sizes of 5000, 10000, and 20000 require only 13.2 seconds/gene, 68.6 seconds/gene, and 580 seconds/gene, respectively. We therefore can efficiently apply the approach to existing sequencing studies using a small computer cluster.

Using simulated data based on underlying coalescent models based on population-genetics theory, we show our approach for cross-phenotype testing of rare variants has appropriate type-I error even in the extreme tails of the p-value distribution and can be much more powerful than standard univariate testing of rare variants when phenotypes under consideration are correlated as expected.

We further showed using simulated data that our approach was far more powerful for cross-phenotype testing of rare variants compared to an existing approach based on multivariate functional linear models that employed B-spline modeling of rare-variant sites across a gene of interest. We next applied our cross-phenotype method to exome-chip data from 540 subjects collected as part of the Genetic Epidemiology Network of Arteriopathy (GENOA). For phenotypes, we considered body-mass index, high-density lipoprotein, systolic blood pressure, and diastolic blood pressure. For genotypes, we identified and studied 3277 genes possessing 5 or more rare variants with sample frequency $< 3\%$ (excluding singleton sites due to concerns about sequencing artifacts). We applied both our KDC-based approach for cross-phenotype analysis to GENOA, as well as two competing approaches: univariate analysis of rare variants adjusting for multiple testing and multivariate analysis using the multivariate B-spine approach. All analyses were adjusted for gender, age, smoking status, lipid-lowering medication status, and top 10 principal components of ancestry. Overall, no genes were associated with the phenotypes at study-wise significance threshold using any of the methods. However, our approach identified 8 genes with suggestive p-values less than 0.001. Univariate analysis of phenotypes only identified 4 such genes with p-values less than 0.001; all 4 of these genes were identified by our cross-phenotype method. The multivariate B-spline method yielded inflated type-I error across the genotype as noted from the method's QQ plot.

While our cross-phenotype test based on the KDC framework is promising, there are still many open problems related to the method that warrant further investigation. For example, if we identify a cross-phenotype association, a follow-up analysis could be to assess whether the cross-phenotype effect is due to biological pleiotropy (a causal locus directly affecting more than one trait) or mediation pleiotropy (a causal locus affecting only one trait, which in turn affects another trait). Existing mediation analyses are not intended to handle high-dimensional traits; the creation of KDC procedures to identify whether an observed cross-phenotype association is mediated by a different set of phenotypes would have tremendous value. Related to this point, since our KDC approach is an omnibus test, an association with just one of the tested phenotypes could result in a significant finding. While the result is valid, researchers will often wish to identify which underlying phenotypes are directly associated with the gene of interest. A mediation analysis would allow investigators to tease apart these relationships. Additionally, our approach assumes unrelated subjects. Extensions of the technique to handle subjects that are either closely or cryptically related are important to ensure validity of subsequent test statistics. Finally, one might be interested in combining cross-phenotype association results from multiple studies through a meta analysis based on summary statistics.

Integrative analysis of two omics datasets from several heterogeneous studies using probabilistic O2-PLS

Jeanine J. Houwing-Duistermaat
(joint work with Said el Bouhaddani, Hae-Won Uh)

Nowadays many studies comprise several omics datasets (genomics, proteomics, glycomics, metabolomics) aiming identification of potential biomarkers for pathogenesis of several diseases, including cancers and metabolic diseases as well as infectious diseases. These biomarkers would lead to improved understanding of the underlying biological mechanism and might be clinically useful as the molecular targets for better diagnosis, prognosis, and treatment. Since these datasets represent the same underlying biological mechanism, integrated analysis of these datasets should be performed to exploit all information.

To relate two datasets with each other, we consider latent variable regression. Motivated by the fact that structural variation in a dataset diminishes the interpretation of the score-loading correspondence when using PLS methods Trygg et al. [1] proposed O2-PLS. The O2-PLS model decomposes two datasets X and Y in three parts:

$$\begin{array}{ccccccc} X & = & TW^{\mathrm{T}} & + & T_{\perp}P_{Y_{\perp}}^{\mathrm{T}} & + & E \\ \underbrace{Y}_{Data} & = & \underbrace{UC^{\mathrm{T}}}_{Joint} & + & \underbrace{U_{\perp}P_{X_{\perp}}^{\mathrm{T}}}_{Specific} & + & \underbrace{F}_{Noise} \end{array}$$

The relation between the joint parts of Y and X is given by the following linear model $U = TB + H$. Note that only X and Y are observed. T and $T_{\perp}$ are the lower dimensional subspaces of X and U and $U_{\perp}$ are the lower dimensional subspaces of Y. The dimensions of these subspaces need to be specified a priori. The algorithm comprises two steps. The first step is application of PLS to identify the latent space spanned by T and $T_{\perp}$ for X and U and $U_{\perp}$ for Y. Then the X and Y specific parts, $T_{\perp}$ and $U_{\perp}$, are computed and subtracted from the original X and Y, each followed by a new PLS step on these reduced X and Y datasets to obtain estimates for T and U. For more details see [1].

We performed simulation studies, which showed that the algorithm was able to identify the underlying components. We applied the method to metabolomics, gene expression and exome sequencing data available in about 200 subjects. We used one latent component for the joint spaces and for the metabolomic specific space. For both genomic datasets we used 8 specific components. The number of components were found by minimizing the cross-validated mean squared error. The explained variances of the latent components (R^2) are given in Table 1. In addition the relationship between the joint components of the deleterious variants and of metabolomics datasets explained 98.3% of the variance of the joint metabolomics component. For the gene expression dataset, this was 67.1%. With regard to the gene expression data, there is overlap between the top 100 loadings of the joint component and the genome wide expression analysis of 8 meta lipids [2].

TABLE 1. Explained variances (R^2)of X and Y by the latent components; j=joint, spec=specific.

	SNP-j	SNP-spec	Metab-j	Metab-spec
R^2	0.561%	5.15%	56.3%	0.328%
	Gene Expr-j	Gene Expr-spec	Metab-j	Metab-spec
R^2	1.29%	15.5%	52.2%	0.839%
Formula	$\frac{\|\|T\|\|}{\|\|X\|\|}$	$\frac{\|\|T_\perp\|\|}{\|\|X\|\|}$	$\frac{\|\|U\|\|}{\|\|Y\|\|}$	$\frac{\|\|U_\perp\|\|}{\|\|Y\|\|}$

To conclude the O2-PLS method can be used to identify correlated subspaces in high dimensional datasets. Future work will be the development of the probabilistic counterpart of O2-PLS. The advantage of such an approach will be modelling of heterogeneity, including prior information with regard to underlying biology in the latent spaces and dealing with missing data.

REFERENCES

[1] Trygg J, Wold S (2003) O2-PLS, a two-block(X-Y) latent variable regression method wit an integral OSC filter. *J. Chemoterics* 17:53–64

[2] Inouye M , Silander K, Hamalainen E et al (2010) An immune response network associated with blood lipid levels. *PLOS genetics* 6:e1001113

[3] el Bouhaddani S, Houwing-duistermaat J, Jongbloed G et al (to appear). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics* Suppl.

A Spectral Approach Integrating Functional Genomic Annotations for Coding and Noncoding Variants

IULIANA IONITA-LAZA

(joint work with Kenneth McCallum, Bin Xu, Joseph Buxbaum)

Over the past few years, substantial effort has been put into the functional annotation of variation in human genome sequence. Indeed, for any genetic variant, whether protein coding or noncoding, a diverse set of functional annotations is available from projects such as Ensembl, ENCODE and Roadmap Epigenomics. Such annotations can play a critical role in identifying putatively causal variants among the abundant natural variation that occurs at a locus of interest. The main challenges in using these various annotations include their large numbers, and their diversity. In particular, it is not clear a priori which annotation is better at predicting functionally relevant variants. It is therefore desirable to integrate these different annotations into a single measure of functional importance for a variant.

Recent efforts have been made to employ these diverse annotations in a more principled way. In particular, several studies have focused on machine learning tools for the integration of many different functional annotations into one single score of functional importance. For example, Kircher et al. [1] proposed a supervised approach (support vector machine) to train a discriminative model using a

labelled training set. Ideally, the training data would be obtained by sampling variants at random and then applying a gold-standard method to determine deleteriousness (or functionality). Unfortunately, such a gold-standard approach is currently not practical for a large number of variants, and so supervised methods must resort to training data that may be inaccurate or biased. In essence, CADD is based on assessing evolutionary conservation, and may be suboptimal for weakly selected (or possibly not selected) disease mutations for complex traits.

In this talk, I discuss an unsupervised spectral approach (**Eigen** [2]) for scoring variants which does not make use of labelled training data. As such, its performance is not sensitive to a particular labeling of the training dataset. Instead, the approach is based on training using a large set of variants with a diverse set of annotations for each of these variants, but no label as to their functional status. We assume that the variants can be partitioned into two distinct groups, functional and non-functional (although the partition is unknown to us), and that for each annotation the distribution is a two-component mixture, corresponding to the two groups. The key assumption in the **Eigen** approach is that of block-wise conditional independence between annotations given the true state of a variant (either functional or non-functional). This last assumption implies that any correlation between annotations in different blocks is due to differences in the annotation means between functional and non-functional variants. Because of this, the correlation structure among the different functional annotations can be used to determine how well each annotation separates functional and non-functional variants (i.e. the predictive accuracy of each annotation). Specifically, we compute a rank one matrix $\mathbf{R}$ approximation of the annotation variance-covariance matrix, and show that the entries in the eigenvector for the rank one matrix are proportional to the accuracies of the individual predictors. Subsequently we construct a weighted linear combination of annotations, based on these estimated accuracies.

We illustrate the discriminatory ability of the proposed meta-score using numerous examples of disease associated variants and putatively benign variants, both coding and noncoding, from the literature. In addition we consider a related, but conceptually simpler meta-score, **Eigen-PC**, which is based on the direct eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations. Across varied scenarios, the **Eigen** and **Eigen-PC** scores perform generally better than CADD, and any single individual annotation, representing powerful single functional scores that can be incorporated in gene-mapping studies, e.g. in the framework of a hierarchical model. Furthermore, an important advantage of the **Eigen** and **Eigen-PC** scores is that, due to their unsupervised nature, they can be easily adapted to a specific tissue or cell type. Future work includes further methodological developments, such as empirical Bayes nonparametric mixture models, and development of context-specific (tissue/cell type) scores that could be used to infer the relevant tissue for a disease of interest.

References

[1] Kircher M et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* doi: 10.1038/ng.2892.

[2] Ionita-Laza I, McCallum K, Xu B, Buxbaum J (2015) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*, to appear

Pitfalls of Rare Variant Data Association Analysis and Method Development

Suzanne M. Leal

(joint work with Gao Wang, Di Zhang, Hang Dai, Zongxiao He, Biao Li)

With the advent of next generation sequencing exome and genome sequence data can be cost-effectively generated. The vast majority of identified variants are rare with minor allele frequency of less than 1%. Although these rare variants can be analyzed using the same methods applied to the analysis of common variants, these approaches are not powerful due to low allele frequencies and allelic heterogeneity. Therefore rare variant association methods have been developed to increase the power of analyzing rare variants. These methods test for an association by aggregating variants across a region, which is usually a gene. To date over 90 rare variant association tests have been developed to analyze rare variants for population-based data and to a lesser degree family-based data. Although so many tests have been developed only a handful of these methods e.g. Combined Multivariate Collapsed (CMC) method [5], Burden of Rare Variants (BRV) [1], Sequence Kernel Association Test (SKAT) [7] and SKAT-O [4] have been used to analyze more than one dataset. The methods that are used are either fixed effects tests (e.g. CMC, BRV) using a variety of coding and weighting schemes, random effect tests (i.e. SKAT) or omnibus tests (i.e. SKAT-O). Some caveats of using rare variant association tests to analyze sequence data are that exclusion of causal variants and inclusion of non-causal variants can greatly reduce power. Additionally currently application of rare variant association methods are predominately limited to the analysis of coding regions, i.e. genes, since it is unknown how to properly aggregate regions outside of genes. Comparisons of rare variant association methods often show that one test is more powerful than the others with different studies having inconsistent findings. One problem is how variant data is generated. A variety of methods can be used to simulate variant data including forward-time, coalescent and allele frequency from data sets such at 1,000 Genomes. If allele frequencies are used from real world data sets where sample sizes are small, e.g. $< 5,000$ individuals there will be an under representation of rare variants in the generated sample, with the sample having a deficiency of singletons, doubletons, etc. This is true even if the generated samples sizes as small as 500 individuals. This problem can be overcome by using newer population demographic models such as those developed by Gazave et al. [3]. We generated data using forward-time simulation using the European population demographic model described by [3] then compared the generated data to exome sequence data on Europeans from

the NHLBI-Exome Sequencing Project [6, 2]. The two data sets showed excellent concordance in the proportions of rare variants and singletons, doubletons etc. We generated sequence variant data for all genes ($N = 18,397$) across the exome and generated data under the alternative using a variety of disease model. We then compared the power of several rare variant association methods. For almost all genes the random effect test had lower power than fixed effect tests. Although the omnibus test did rank higher than the random effect test SKAT it usually did not perform better than fixed effect tests due to a correction for multiple testing. Although there was always at least one gene for which a particular test was most powerful. This is how previous studies where rare variant association methods were compared could always show that a particular method was more powerful than the others by cherry-picking a simulation scenario. It could be seen that even the genetic architecture of different genes can make one test appear more powerful than another and it is always possible to generate data for a particular gene to make the test of ones choice appear more powerful. Naturally this should not be done and tests should be compared using all genes for a variety of disease models.

References

[1] Auer PL, Wang G, Leal SM. *Testing for rare variant associations in the presence of missing data*. Genet Epidemiol 2013;37:529–38.

[2] Fu W, O'Connor TD, Jun G, et al. *Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants*. Nature 2013;493:216–20.

[3] Gazave E, Ma L, Chang D, et al. *Neutral genomic regions refine models of recent rapid human population growth*. Proc Natl Acad Sci U S A 2014;111:757–62.

[4] Lee S, Emond MJ, Bamshad MJ, et al. *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies*. Am J Hum Genet 2012;91:224–37.

[5] Li B, Leal SM. *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data*. Am J Hum Genet 2008;83:311–21.

[6] Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64–9.

[7] Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011;89:82–93.

Simultaneous Sparse Signal Detection with Applications in Genomics

Hongzhe Li

(joint work with S. Dave Zhao, Tony Cai)

This paper presents a new statistical method for identifying important disease genes. It functions by integrating eQTL study results with GWAS results from an independent set of subjects. Motivated by Figure 1, the method tests each gene for whether there are any SNPs which are associated both with the gene's expression, using the genetical genomics data, and with disease, using the genome-wide association data. Each significant SNP association, whether with expression or with disease, is termed a "signal", and the method detects simultaneous signals.

The rationale behind this procedure is that SNPs can be viewed as perturbations of the underlying biological systems, especially the gene regulatory networks underlying various complex diseases. Therefore for a disease-causing gene, any genetic variation that perturbs its expression is also likely to influence disease risk. Furthermore, unlike differential expression, the proposed approach is able to differentiate causal genes G_R from reactive genes G_C in Figure 1. This is because under that causal model, G_R and SNP_C are independent conditional on disease, while a simultaneous detection procedure will only identify genes that are associated with at least one causal SNP. In other words, G_R will not exhibit any simultaneous signals.

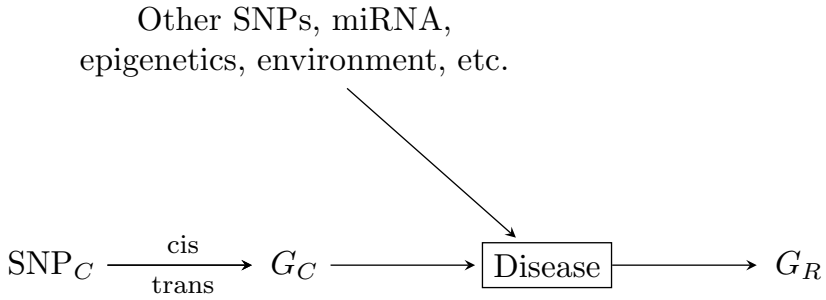

FIGURE 1. A simple causal model illustating a problematic setting for differential expression analysis. SNP_C: causal SNP; G_C: causal gene; G_R: reactive gene. SNP_C can be either *cis* or *trans* to G_C. Only G_C is of interest, but differential expression analysis cannot distinguish between G_C and G_R.

Simultaneous signal detection is conducted one gene at a time. For a given gene, define unobserved signal indicators $X_i, Y_i \in \{0, 1\}$ to indicate whether the ith SNP, $i = 1, \ldots, n$, is truly associated with the disease or the gene's expression, respectively. Significant GWAS and eQTL SNPs are usually rare, or sparse, so very few of the X_i and Y_i equal 1. The observed data consist of test statistics U_i, for the SNP-disease association, and V_i, for the SNP-expression association. These are assumed to follow

$$\begin{aligned} &U_i \mid X_i = 0 \sim F_0^U, \quad U_i \mid X_i = 1 \sim F_i^U, \quad F_i^U \leq F_0^U \\ &V_i \mid Y_i = 0 \sim F_0^V, \quad V_i \mid Y_i = 1 \sim F_i^V, \quad F_i^V \leq F_0^V, \quad U_i \perp\!\!\!\perp V_i, \end{aligned} \tag{1}$$

where the F_0^U and F_0^V are null distributions, which may be known or unknown, and the F_i^U and F_i^V are unknown alternative distributions. The test statistics are assumed to be stochastically larger under the alternatives, which is reasonable for two-sided tests. The U_i are usually obtained from a GWAS study using linear or logistic regression for continuous or binary diseases and the V_i are usually be obtained from an eQTL study using linear regression. Finally, the U_i and V_i are independent for all i because the two studies are assumed to have been conducted in two independent samples.

Under Model 1, let $\epsilon_n = n^{-1}\sum_i I(X_i = 1, Y_i = 1)$ denote the fraction of simultaneous signals. The simultaneous signal detection problem is thus to test

$$H_0 : \epsilon_n = 0 \quad vs. \quad H_A : \epsilon_n > 0 \tag{2}$$

using the observed $(U_i, V_i), i = 1, \ldots, n$. Rejecting H_0 indicates that the expression of the gene being tested is regulated by SNPs which are also associated with disease, suggesting that the gene is likely to be functionally relevant.

To test whether V_i for a given gene and U_i share any simultaneous signals, recall from model (1) that the U_i and V_i are assumed to be stochastically larger when the signal indicators X_i and Y_i equal 1, respectively. Thus if SNP i is truly simultaneously associated with both the disease and the gene's expression, then both U_i and V_i should be large, so it is reasonable to define the statistic $T_i = U_i \wedge V_i$. Intuitively, H_0 of (2) should be rejected if at least one SNP has a observed large value of T_i, so the proposed test statistic is

$$M_n = \max_{i=1,\ldots,n} T_i. \tag{3}$$

A large value of M_n would imply that the gene is functionally relevant for disease. One caveat is that U_i and V_i should be on roughly the same scale, meaning that the null variances of U_i and V_i should be comparable.

A permutation p-value for the proposed M_n statistic can be obtained by a simple hypergeometric probability calculation. Asymptotic optimality in term of detection boundary under the sparse model is derived and demonstrated using both simulated and real data sets. The method was further demonstrated by an application to a study that combined a genome-wide association study of human heart failure and a eQTL study of human heart cardiomyocytes, where genotype data was collected from 1,586 controls and 2,027 heart failure cases using the Illumina OmniExpress Plus. In addition, Left ventricular free-wall tissue was collected from hearts of 177 patients with advanced heart failure who were undergoing transplantation and from 136 donor hearts without heart failure. Genotype data were collected using using the Affymetrix Genome-Wide SNP Array 6.0 and only markers in Hardy-Weinberg equilibrium with minor allele frequencies above 15% were considered, leaving 347,019 SNPs. Our analysis identified nine interesting genes that are potentially causal to heart failure. These genes involve heart muscle contraction, inflammation and angiogenesis. In contrast, differential expression analysis identified over 9,000 differentially expressed genes, which are impossible for any laboratory validations.

Multiple Phenotype Association Tests using GWAS Summary Statistics

Xihong Lin
(joint work with Zhonghua Liu)

In this paper, we consider testing for association between a genetic variant and multiple correlated phenotypes without access to the individual level genotype or phenotype data. We first investigate the information contained in summary statistics from genome-wide association studies (GWASs), and show explicitly how the means and correlation structure of the summary statistics are related to the individual level data. Based on this connection, we can aggregate statistical evidence across multiple phenotypes without using individual level data. Since a genetic variant could affect multiple phenotypes in different directions with different magnitudes, and the correlation structure among multiple phenotypes can also be arbitrary, therefore we propose testing procedures that fall into two categories.

The first category contains a series of robust and powerful testing procedures based on linear mixed models. The score testing statistic for the fixed effect aims to detect homogeneous effects and the score testing statistic for the random effect aims to detect heterogeneous effects. We further propose a number of ways to combine these two independent score testing statistics and therefore the resulting combined tests are more robust to effect heterogeneity.

The second category contains a series of tests based on principal components (PCs) performed on the summary statistics. We introduced a novel geometric concept called principal angle which can well explain the powers of single PC test and PC combination based tests. We further used theoretical power analysis to find the most favorable and least favorable alternatives for those PC based tests and conclude that each test could be almost powerless under their least favorable alternatives. To overcome this limitation, we propose two adaptive tests that take the minimum p-value of PC combination based tests and the resulting adaptive tests are more robust to various alternatives and are still powerful.

In addition, we have analytic formulas to compute the p-values for all of the proposed tests. This computational advantage makes our methods practically appealing in large-scale genetic studies. The proposed tests all maintain correct type I error rates and their powers are compared in various settings via simulation studies. We further apply these tests to a GWAS summary statistics data set from the Global Lipids Genetics Consortium and identify hundreds of genetic variants that were missed by the original single-trait analysis. The newly detected genetic variants indicate potentially novel lipids biology by checking their functional annotations. We also develop an R package MPAT freely available for public uses.

Key words: GWAS; Linear mixed models; Multiple phenotypes; Principal components; Principal angles; Score tests; Summary statistics; Variance component test

Learning molecular networks: interventions, joint estimation and causal interpretation

Sach Mukherjee

This talk focuses on the problem of estimating network structure from molecular data. This problem is of current interest in computational and molecular biology but also gives a concrete setting for the exploration of general ideas concerning causal inference that has the virtue of allowing empirical verification by interventional experiments.

The problem can be stated as follows. Given data on molecular variables $V = \{1 \dots p\}$ (e.g. transcripts or proteins), the goal is to estimate a graph G with vertex set V that describes molecular influences between the variables. Here we focus on directed graphs and causal influences and consider the "detection" problem of estimating the presence or absence of edges in G (rather than estimation of quantitative causal effects); we will use the shorthand $(a, b) \in G$ to indicate that the pair (a, b) belongs to the edge set of graph G. Furthermore, motivated by problems that arise in modern "high-throughput" biology we focus especially on approaches that could potentially scale to relatively large p, including variables that can be measured but whose mutual interplay might currently be poorly understood. We focus on time course data in the molecular biological setting, but note that the ideas (and concerns) are general.

Basic model. We first outline a basic model and then discuss in turn the modelling of interventional data and joint estimation of multiple, non-identical networks. For time course data obtained at T discrete time points, the model we employ is a directed graphical model, where each variable at each time point depends on a subset of the variables at the previous time point. Letting X denote the collection of data for all variables at all time points, the likelihood can be written as

$$p(X \mid G, \{\theta_j\}) = \prod_{j=1}^{p} \prod_{t=2}^{T} p(X_j^t \mid X_{\pi_G(j)}^{t-1}, \theta_j), \tag{1}$$

where G is the (latent) graph of interest, X_j^t is the abundance of molecule j at time t, X_A^t denotes abundances at time t for a subset A of the variables, $\pi_G(j) \subseteq V$ is the set of parents of j in graph G and θ_j are parameters describing the dependence of variable j on its parents in the graph. This type of model is sometimes called a Dynamic Bayesian Network (DBN). The specific form (1) is a very basic DBN, but more elaborate models (e.g. allowing G to change over time, as in [2]), are possible. Here, we use standard linear models throughout, but note that more complex models can be used (typically at added computational cost).

Inference is performed within a Bayesian framework that is closely related to Bayesian variable selection. This allows us to perform inference with respect to the latent graph G and report marginal posterior probabilities of the form $P((i, j) \in G \mid X)$. As discussed in [3], models of the form (1) admit computationally advantageous factorization, so that under some restrictions inference can be

carried out exactly (by enumeration). For full details, including a biological case study, see [3].

Modelling interventions. In many settings, data includes interventions. Such data, when available, can be particularly informative with respect to causal relationships, and should therefore be included in analysis. We do so by making modifications to the likelihood for samples in which interventions were carried out (this is an application of well known ideas from the causal directed acyclic graph literature). An important point is that such modifications need to take account of the nature of the intervention, in particular whether it is an intervention on a node, edge or set of edges. For further details see [6].

Joint estimation over multiple, non-identical graphs. Increasingly, experimental designs span multiple contexts - such as disease subtypes or cell types - that may require context-specific models. However, it is not expected that context-specific networks will be entirely different, but rather variations on a common theme. This motivates joint estimation of the networks to share information between the problems, whilst allowing for differences. We have pursued this idea by modifying the approach outlined above to allow simultaneous inferences concerning K networks $G_1 \dots G_K$. This is done via a hierarchical Bayesian formulation, using belief propagation for efficient inference. Full details are reported in [5].

Causal claims and their empirical assessment. The scientific goal of molecular network estimation is usually to obtain causal insight into how molecules influence one another in specific biological or biomedical contexts. However, strong assumptions are needed to guarantee that statistical models - including the graphical models described above - will yield causal insights (even asymptotically), and these assumptions may be difficult or impossible to check in practice. These issues are rather general ones in causal modelling and appear in many places in the literature, see [1] for an enlightening discussion in the context of directed acyclic graph models, and [6] for a brief discussion of causal issues in the specific context of the models above. Furthermore, beyond general concerns about causal inference and statistical models, there are specific issues that arise in biological settings and with biological data that might make estimation of causal structure difficult in practice (for a discussion of some of these points see [4]). The upshot of these concerns is that one cannot be assured in advance that existing estimation methods can truly deliver causal insights. Recently, we have therefore focused in parallel on the question of empirical assessment of causal estimation: in other words, to ask using real data whether proposed methods actually work in practice. We do not discuss this line of work here, but mention these points to remind the reader of these important caveats and limitations.

The methodologies described above are fully described in [3, 5, 6] and were joint work with the co-authors of those papers, especially Steven Hill, Chris Oates and Simon Spencer. The biological applications were performed in collaboration with the Spellman and Gray laboratories at OHSU, Portland, USA and the Mills laboratory at MD Anderson Cancer Center, Houston, USA.

References

[1] Dawid, A. P. (2010). Beware of the DAG! *Journal of Machine Learning Research, W & CP*, 6: 59–86.

[2] Dondelinger, F., Lèbre, S. & Husmeier, D. (2013). Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine Learning*, 90(2), 191-230.

[3] Hill, S. M., Lu, Y., Molina, J., Heiser, L. M., Spellman, P. T., Speed, T. P., Gray, J. W., Mills, G. B. & Mukherjee, S. (2012). Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics*, 28(21):2804–2810.

[4] Oates, C. J. & Mukherjee, S. (2012) Network inference and biological dynamics. *The Annals of Applied Statistics*, 6(3):1209–1235.

[5] Oates, C. J., Korkola, J., Gray, J. W. & Mukherjee, S. (2014). Joint estimation of multiple related biological networks. *The Annals of Applied Statistics*, 8(3):1892–1919.

[6] Spencer, S. E., Hill, S. M. & Mukherjee, S. (2015). Inferring network structure from interventional time-course experiments. *The Annals of Applied Statistics*, 9(1):507–524.

Evolving designs in disease genetics

Dan Nicolae

(joint work with Carole Ober, Oren Livne, Sahar Mozaffari and Matthew Reimherr)

Understanding the role of genetic polymorphisms in human phenotypic variation has been the main motivator for the development of new technologies and the dramatic increase in the variety and scale of data we have seen for the past decade. Classical variant-phenotype association tests performed in genome-wide association studies of thousands of subjects have led to a large number of discoveries, but our understanding of the genetic architecture of human complex diseases is still incomplete. In this talk I will argue that one solution for progress in the field is the use of datasets that are deep in clinical data and/or genetic and genomics measurements.

I will use studies in the South Dakota Hutterites to illustrate methods for integrating whole genome sequencing, array SNP data, RNA-sequencing and pedigree information in gene mapping studies. I will describe advantages of using founder populations for such investigations, including cost-efficient study designs. The efficiency is achieved using a fast phasing and computationally efficient imputation method that combines the advantages of pedigree-based and LD-based methods and obtains accurate genotypes and high call rates in 1317 related Hutterites using whole genome sequencing data on only 98 related individuals. In addition, the algorithm allows accurate parent-of-origin assignments for each allele as well as imputed genotypes of recent ancestors (or other members of the pedigree) with no DNA or available genotype information. The description of the algorithm has been recently published in [1]. I will illustrate how this additional information is used for discovery of parent-of-origin effects and imprinting in cardiovascular traits.

I will also present novel methodology for association testing with longitudinal phenotypes [2]. The methods are based on ideas from functional data. In short, we reconstruct trait trajectories (curves) via smoothing and interpolation, and apply probabilistic tools for function spaces to curves. The inference is based on the following model,

$$Y_n(t) = \alpha(t) + X_{1,n}^T\beta_1(t) + X_{2,n}^T\beta_2(t) + \varepsilon_n(t),\ t \in [0,]1.$$

where Y_n is the phenotype for the n-th subject, X_1 are covariates, and X_2 are the genotypes for the K SNPs that are tested. We assume that $\{\varepsilon_n\}$ are independent and identically distributed in $L^2[0,1]$, $\mathrm{E}[\varepsilon_n(t)] = 0$, $\mathrm{E}\|\varepsilon_n\|^2 < \infty$. The test statistic for K predictors in X_2 is given by,

$$\Lambda = \sum_{n=1}^{N}(\|Y_n - X_{1,n}^T\hat{\beta}_1\|^2 - \|Y_n - X_n\hat{\beta}\|^2),$$

and it can be shown that if $\beta_2 = 0$

$$\Lambda \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty}\lambda_i\chi_i^2(K),$$

where the χ^2 variables are iid and λ_i's are eigenvalues from the spectral decomposition of the covariance function. Applications to a genome-wide association study of longitudinal lung function are shown.

References

[1] Livne OE, Han L, Alkorta-Aranburu G, Wentworth-Sheilds W, Abney M, Ober C and Nicolae DL, *PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population*, PLoS Computational Biology **11(3)** (2015), e1004139.

[2] Reimherr M and Nicolae DL, *A Functional Data Analysis Approach for Genetic Association Studies*, The Annals of Applied Statistics **8(1)** (2014), 406–429.

Confounding in omics data analysis: an example

Michael Nothnagel

(joint work with S. Siegert, A. Wolf, D.N. Cooper and M. Krawczak)

Confounding is a recurrent issue in statistical analyses, in particular of large and retrospective datasets. Integrative approaches to omics data analysis, combining two or more layers of genomic information, can be expected to be subject to multiple sources of confounding at the different layers. Here, I am going to present a recent result for the most basal layer, namely genetic variation, demonstrating the possibility of substantial confounding among causal mutations [1]. More specific, a shared genealogy can induce a negative correlation between variants that act as causal complements, i.e. causing the phenotype of interest independently from each other; see Figure 1 for an illustration. A prerequisite for this phenomenon to occur is some degree of tolerance of an organism to a limited number of deleterious mutations before the trait manifests itself.

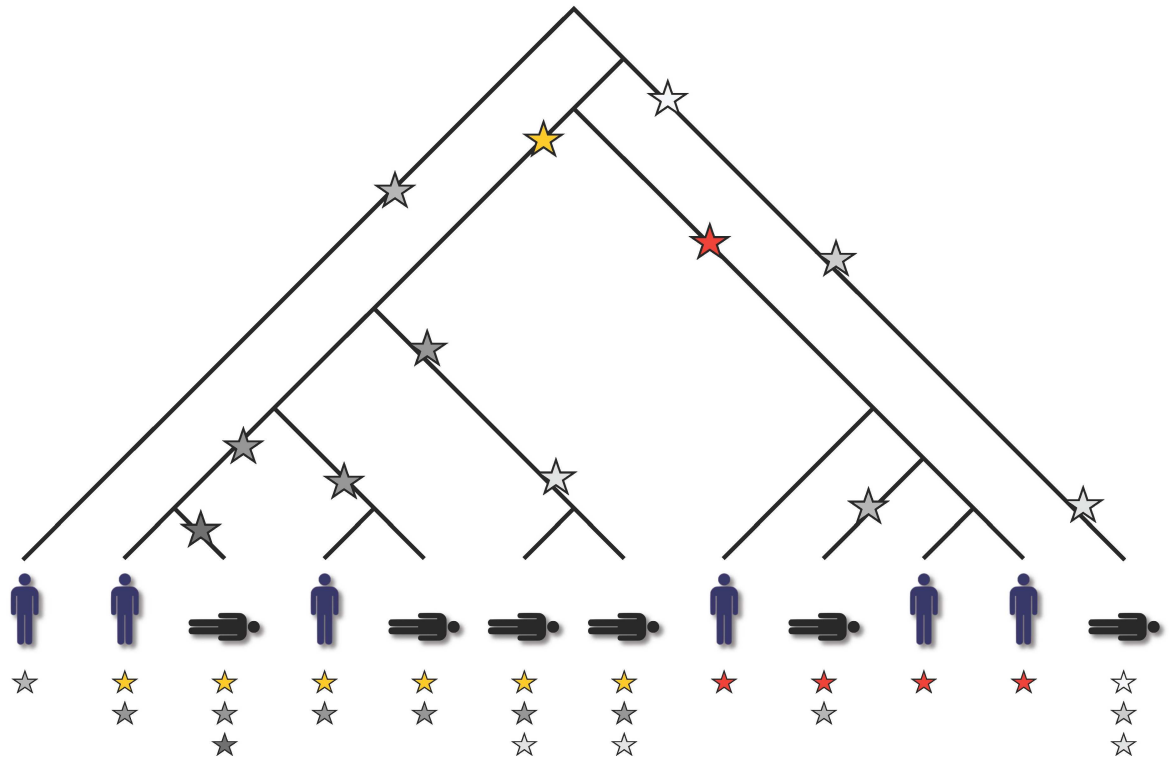

FIGURE 1. **Illustration of confounding phenomenon**. The coalescent tree depicts the genealogy of a population of twelve haploid individuals, with six being affected with some disease (cases; lying symbol) and six healthy (controls; upright symbol), as well as the carrier-ship of deleterious mutations of equal effect. Zero and one mutation are assumed to be fully tolerated (0% affection probability), whereas two mutations are partially tolerated (50%) and carrier-ship of three or more mutations inevitably leads to affection (100%). Occurrence of the red and the yellow mutations is negatively correlated ($\Phi_{cases} = -0.63$, $\Phi_{controls} = -0.71$). Following expectation, the yellow mutation is enriched in cases (odds ratio of 2.00). However, the red mutation is depleted in cases (odds ratio of 0.33) and would thereby, although deleterious, appear protective in an epidemiological study.

A major consequence of this phenomenon is a possible depletion of a part of causal variants in patients compared to unaffected individuals. Such disease-causing mutations then appear 'protective' in genetic epidemiological studies. Possible consequences for such apparently protective mutations in omics analyses include a power loss due to the exclusion of causal variants from the integrative model and in some burden association tests of rare genetic variation, but also consideration of the wrong, not harmful allele in such models and in functional

annotation. The described phenomenon adds to the list of 'strange' genetic epidemiological phenomena, such as 'flip-flop associations' [2], 'synthetic association' (see [3] and others) and 'indirect associations' [4], but is different from them.

We set out to evaluate the relevance of this confounding phenomenon through coalescent simulations. To this end, we repeatedly simulated haploid populations of 10,000 individuals under a Wright-Fisher model of a single non-recombining locus without selection, randomly placing mutations on branches proportional to branch length. Later, simulations were extended to up to ten loci in order to approximate the effect of recombination between multiple causal loci. Disease affection status was randomly assigned to each individual based on the number of carried deleterious mutations at any of the loci, $K = \sum_{l=1}^{L} k_l$ where L denotes the number of loci and k_l the number of deleterious mutations at locus l. We considered two affection probability functions, namely a multiplicative one,

$$\mathrm{P}(K) = 1 - (1 - \gamma)^K \quad ,$$

in order to model little-tolerance scenarios and a logistic one,

$$\mathrm{logit}(\mathrm{P}(K)) = \alpha - \beta \cdot K \quad ,$$

for describing scenarios of tolerance to a limited number of deleterious mutations before trait manifestation. Two sets of scaling parameter values α, β and γ were used in the simulations. For given number of loci, affection probability function and parameter values, simulations were repeated until 1000 populations were available for each of three prevalence classes, namely rare (0.1-1%), common (1-5%) and pandemic (10-20%). We found that oligo- and even multi-locus models for common diseases can yield substantial proportions of disease-causing mutations that appear 'protective' in genetic epidemiological studies by being depleted in patients compared to unaffected individuals.

Our reported phenomenon implies a negative trend of the mutation effect sizes with increasing disease prevalence. In order to evaluate this prediction, we analysed publicly available data from the GWAS catalogue [5]. More specific, we considered only traits with at least ten SNP markers having been reported to be associated with this trait at a significance level of 5×10^{-4}. We indeed observed the predicted negative trend for different sets of traits (34-51) and SNP markers (1495-2437), resulting from different filtering thresholds for disease prevalence and significance level. The reported confounding phenomenon is, thus, consistent with data on disease-associated variants from genome-wide association studies, although the observed negative trend is also explicable by other causes.

References

[1] Siegert S, Wolf A, Cooper DN, Krawczak M, Nothnagel M (2015) Mutations causing complex disease may under certain circumstances be protective in an epidemiological sense. *PLoS One* **10**:e0132150.

[2] Lin P-I, Vance JM, Pericak-Vance MA, Martin ER (2007) No gene is an island: the flip-flop phenomenon. *Am J Hum Genet* **80**:531–8.

[3] Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**:e1000294.
[4] Platt A, Vilhjalmsson BJ, Nordborg M (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**:1045–52.
[5] National Human Genome Research Institute (NHGRI) & European Bioinformatics Institute (EMBL-EBI). GWAS catalog. URL: http://www.ebi.ac.uk/gwas/

Disentangling transcriptional heterogeneity among single-cells: a Bayesian approach

Catalina A. Vallejos
(joint work with John C. Marioni, Sylvia Richardson)

Multiple levels of biological heterogeneity define distinct populations — going from different species (e.g. mouse, human) to organs within a subject (e.g. heart, brain) and the individual cells that constitute an organ (e.g. neurons inside the brain). Among other molecular phenotypes, such populations can be characterised by differences in their gene expression profiles. Up until recently, transcriptomic studies have focused on examining bulk expression, measured as an average across thousands of cells. Some biological processes, however, require the study of variation in gene expression at the single-cell level.

A few years ago, single-cell mRNA sequencing (scRNA-seq) emerged as a tool for quantifying gene expression profiles of individuals cells. This novel technology can uncover cell-to-cell heterogeneity in seemingly homogeneous populations of cells. In addition to experimental challenges (such as the isolation of individual cells), statistical analysis of scRNA-seq data is itself a challenge. In particular, compared to bulk RNA-seq, scRNA-seq datasets lead to increased variance estimates of gene expression. This is partially related to biological differences (e.g. changes in mRNA content and the existence of cell sub-populations or transient states), which disappears when measuring bulk gene expression. Nonetheless, this variance inflation is also due to unexplained *technical noise*, which is confounded with genuine cell-to-cell heterogeneity [1].

To deal with these issues, we developed BASiCS (Bayesian Analysis of Single-Cell Sequencing data) [2], a hierarchical Bayesian model for the analysis of scRNA-seq datasets. It borrows information between genes that are intrinsic to the population of cells under study and technical *spike-in* genes which are artificially added to each cell in known amounts. These known quantities provide a control or *gold standard* to which empirical measurements of spike-in genes' expression can be compared, enabling a quantitative calibration of the technical noise. For each gene i ($i = 1, \ldots, q$) and cell j ($j = 1, \ldots, n$), our model is defined as

$$(1) \quad X_{ij}|\mu_i, \phi_j, \nu_j, \rho_{ij} \overset{\text{ind}}{\sim} \begin{cases} \text{Poisson}(\phi_j \nu_j \mu_i \rho_{ij}), & \text{if gene } i \text{ is intrinsic;} \\ \text{Poisson}(\nu_j \mu_i), & \text{if gene } i \text{ is a spike-in, with} \end{cases}$$

$$(2) \quad \nu_j|s_j, \theta \overset{\text{ind}}{\sim} \text{Gamma}(1/\theta, 1/(s_j\theta)) \quad \text{and} \quad \rho_{ij}|\delta_i \overset{\text{ind}}{\sim} \text{Gamma}(1/\delta_i, 1/\delta_i),$$

In this Poisson formulation, ϕ_j's and s_j's act as cell-specific normalising constants, bringing expression counts into a comparable scale. In addition, cell-specific random effects ν_j capture unexplained technical noise, whose strength is controlled by global noise parameter θ. Additional random effects ρ_{ij} account for heterogeneous expression of a gene i across cells, whose strength is quantified by gene-specific over-dispersion parameters δ_i. Finally, μ_i's represent gene-specific expression rates, as an average across all cells. These parameters are known in case of spike-in genes, hence the identifiability restriction $n^{-1}\sum_{j=1}^{n}\phi_j = 1$ ensures identifiability of all model parameters. Our software is available at `https://github.com/catavallejos/BASiCS`. Importantly, our method avoids stepwise approaches where datasets are firstly normalised and secondly technical noise is removed prior to other downstream analyses, ignoring the uncertainty related to the initial steps [1].

BASiCS can highlight genes showing particularly large or low heterogeneity across the analysed cells. Highly variable genes (HVG) constitute key drivers of cell-to-cell heterogeneity and are potential markers of novel cell sub-populations. In contrast, lowly variable genes (LVG) related to core processes of the cell. To detect HVG and LVG, we use a probabilistic approach based on tail posterior probabilities associated to high and biological heterogeneity components. These are calibrated by controlling a trade-off between false discovery and false negative rates. We demonstrate our method using gene expression measurements from mouse Embryonic Stem Cells. Cross-validation and meaningful enrichment of gene ontology categories within genes classified as highly (or lowly) variable supports the efficacy of our approach.

More recently, we extended BASiCS to include other downstream analyses that help functional characterization of multiple pre-specified populations of cells (defined by experimental conditions or cell-types) [3]. In particular, we focus on *differential expression* analyses where the aim is to identify genes that exhibit changes in expression between the analysed populations. Our method goes beyond traditional differential expression tools, where changes in expression are restricted to differences in overall expression. Instead, we are also able to identify changes in cellular heterogeneity. To validate our method, we compared expression between mouse embryonic stem cells and *pool-and-split* samples consisting of pooled RNA from thousands of cells split into single-cell equivalents. As expected, BASiCS rules our a global shift in gene expression levels between cells and pool-and-split samples. Additionally, we infer a substantial decrease of biological over-dispersion on the pool-and-split samples, which is intuitive as they reflect pooled expression levels across thousands of cells.

References

[1] P. Brennecke et al., *Accounting for technical noise in single-cell RNA-seq experiments*, Nature Methods **1** (2013), 1093–1095.

[2] C. Vallejos, J. Marioni and S. Richardson, *BASiCS: Bayesian Analysis of Single-Cell Sequencing data*, PLoS Computational Biology **11** (2015), e1004333.

[3] C. Vallejos, S. Richardson and J. Marioni, *Disentangling transcriptional heterogeneity among single-cells: a Bayesian approach*, In preparation (2016).

Playing musical chairs in multi-phenotype studies improves power and identifies novel associations

Noah Zaitlen

(joint work with Hugues Aschard, Peter Kraft)

Testing for associations in big data faces the problem of multiple comparisons, with true signals buried inside the noise of all associations queried. This is particularly true in genetic association studies where a substantial proportion of the variation of human phenotypes is driven by numerous genetic variants of small effect. The current strategy to improve power to identify these weak associations consists of applying standard marginal statistical approaches and increasing study sample sizes. While successful, this approach does not leverage the environmental and genetic factors shared between the multiple phenotypes collected in contemporary cohorts. Here we develop a method that improves the power of detecting associations when a large number of correlated variables have been measured on the same samples. Our analyses over real and simulated data provide direct support that large sets of correlated variables can be leveraged to achieve dramatic increases in statistical power equivalent to a two or even three or four fold increase in sample size.

The objective of this work is to develop a method that keeps the resolution of univariate analysis when testing for association between an outcome Y and candidate predictor X, but takes advantage of other available covariates $C = (C_1, C_2, \ldots, C_m)$ to increase power. A first step toward that aim is to consider the inclusion of covariates correlated with the outcome in a standard regression framework. This may increase the signal-to-noise ratio between the outcome and the candidate predictor when testing: $Y = X + C$. The selection of which covariates C_i are relevant to a specific association test is usually based on causal assumptions. Putting aside the estimation of indirect and direct effect of X on Y, epidemiologists and statisticians recommend the inclusion of two types of covariates: those that are potential causal factors of the outcome and independent of X, and those that may confound the association signal between X and Y, i.e. variables such as PC covariates that capture undesired structure in the data that can lead to false association. All other variables that vary with the outcome because of shared risk factors are usually ignored. However, those variables carry potential interesting information about the outcome, and more precisely about the risk factors of the outcome. Because of their shared dependences they can be used as proxies for risk factors of the outcome. As such they can be incorporated in C to improve the detection of associations between X and Y. However, as we discuss further, when these variables depend on the predictor X, using them as covariates can lead to both false positive and false negative results depending on their underlying causal structure. The presence of interdependent explanatory

variables, also known as multicollinearity, can induce bias in the estimation of the predictors effect on the outcome. We recently discussed this issue in the context of genome-wide association studies that adjusted for heritable covariates. Take the simple case of two independent covariates U_1 and U_2 that are true risk factors of Y. When testing for association between X and Y, adjusting for U_1 and U_2 can increase power, because the residual variance of Y after the adjustment is smaller while the effect of X is unchanged. Consider the situation where U_1 and U_2 are unknown but a covariate C that also depends on U_1 and U_2 has been measured. Y and C display positive correlation, and when X is not associated with C, adjusting Y for C increases power to detect (Y, X) association, although the gain in power will be smaller than directly adjusting for U_1 and U_2. Problems arise when C is associated with X. In this case adjusting Y for C biases the estimation of the effect of X on Y, decreasing power when the effect X is concordant between C and Y, and inducing false signal when the effect is discordant. The same principles apply for any number of variables correlated with the outcome provided the sample size is large enough such that the effect of all covariates can be estimated in a multiple regression. When none of the covariates depend on the predictor, their inclusion in a regression can reduce the variance of the outcome without confounding, leading to increased statistical power while maintaining the correct null distribution. This gain in power can be easily translated in terms of sample size increase. The non-centrality parameter (ncp) of the standard univariate test equals $ncp = N \times v_X(\sigma^2_Y)$ where N, v_X and σ^2_Y are the sample size, the variance of the outcome explained by the predictor, and the total variance of the outcome respectively. When reducing σ^2_Y by a factor τ, $ncp = N \times v_X(\sigma^2_Y/\tau) = (N/\tau) \times (v_X(\sigma^2_Y))$. For example, when the covariates explain 30% of the variance of the outcome, the power with the covariates is equivalent to analyzing a 1.4 fold larger sample size without the covariates. When covariates explain 80% of the phenotypic variance - a realistic proportion in some genetic datasets - the power gain is equivalent to a 5 fold increases in sample size. The central problem that must be solved is how to intelligently select a subset of the available covariates to optimize power while preventing induction of false positive or false negative associations. To do this all covariates associated with the outcome should be included except those also associated with the predictor. A naive solution would consist in filtering based on a p-value threshold from the association test between the predictor and each covariate. However, unless the sample size is infinitely large, some associations will be missed and unwanted covariates will be included. Furthermore, because a number of the covariates will be associated with the predictor by chance, the overall distribution of p-values from the covariate-adjusted test can be inflated, again potentially inducing false association signal. The underlying problem with p-value based filtering is that p-values are used to reject the null hypothesis in favor of the alternate. In this case the objective is to reject those covariates under the alternative hypothesis. Therefore, instead of only using p-values to filter covariates, we additionally develop a heuristic based on equivalence testing to improve the filtering of covariates while controlling the type I and type II error rate. Consider

$\widehat{\beta}$, the estimated marginal effect of the predictor X on the outcome Y. Using $\widehat{\beta}$ along with the estimated correlation between between C and Y, we can derive the expected distribution of $\widehat{\delta}$, the estimated regression coefficient between X and C under a complete null model ($\beta = 0$ and $\delta = 0$). When the observed regression coefficient between X and C is far from expectation we filter out the covariate C. Thus we have two types of filtering we use to select covariates, based filtering and equivalence test based filtering. For each of these filters, rejection thresholds are set according the variance explained by the potential covariates. We refer further to our approach as the Musical Chair (MC) algorithm because the list of covariates differs for each pair of outcome/predictors (Y, X) tested. More formally, for an outcome Y, and a predictor X_s, $s = 1 \dots n$, the MC algorithm uses three features to select covariates and perform statistical tests: i) p_MUL, the p-value for the overall association between all C_I, and X_s; ii) r^2_C the amount of total outcome variance explained by the potential covariates; and iii) $\widehat{\beta}_s$, the estimated effect of the predictor on Y. The first two features are used to define the stringency of the filtering, being very high for low values of p_MUL, which reflects the likelihood of the presence of undesired covariates, and high values of r^2_C, because of potential bias. The third feature, $\widehat{\beta}_s$, is used to make inference on the expected null distribution of $\widehat{\delta_I}$, the regression coefficient between X_s and the C_l. It leverages the correlation between $\widehat{\beta}_s$ and $\widehat{\delta_I}$ under a complete null model ($\beta_s = 0$ and $\delta_l = 0$). These features are combined to derive a confidence interval Δ_l for each $\widehat{\delta_I}$, which determines whether a covariate can be safely included in the model. We explored, through extensive simulation studies, a set of parameters to weight each of these components in order to optimize power and robustness.

An Orientational walk in the random forest: About first steps, solid grounds and interactions in a random forest

Andreas Ziegler

(joint work with Marvin N. Wright, Inke R. König)

Untangling the genetic background of complex diseases requires the identification of interaction effects and genetic variants involved in these interactions. Random forests (RF) have repeatedly been heralded to be suitable for this endeavor. Specifically, it is mostly argued that RF variable importance measures (VIM) take interaction effects naturally into account. However, it has been shown that especially in the high-dimensional situation, standard VIM fail to detect interaction effects if the interaction partners do not have strong marginal effects (e.g.,Winham et al., 2012). Some authors have argued that VIM from non-totally randomized trees suffer from combinations of defects, making them not useful at all (Louppe et al. 2013). One approach is to use totally randomized trees or paired VIM. The latter have not been investigated in simulation experiments regarding their ability to detect interaction effects (Ishwaran, 2007). In addition, previous simulation studies only investigated specific simple interaction settings. In contrast, many

interaction scenarios are conceivable in reality including, for example, synergistic, modifying, or redundant interaction effects (Lanktree & Hegele, 2009). In the first part of this presentation, we provide a simple introduction to RF, starting with the generation of single classification trees or probability estimation trees. A link to nearest neighbor approaches is made. Next, different extensions are considered which may be used with RF. This is followed by a detailed discussion of the tuning parameters of RF, namely the number of trees in an RF, the terminal node size of a tree and the number of independent variables made available at a split point. VIM and variable selection procedures are introduced. In the second part of the presentation, several statistical properties of RF are summarized. In the third part, we report the results of a comprehensive simulation study for investigating the ability of RF to detect interactions. We specifically simulated several realistic interaction scenarios described before (Lanktree & Hegele, 2009; Musani et al. 2007). With these simulations we demonstrate that the VIM of RF are unable to adequately capture interactions. In conclusion, RF are a simple-to-use machine learning approach suitable for the analysis of genetic data. The statistical properties of this machine are convincing. However, with the standard RF procedure, statistical interactions cannot be adequately detected.

Reporter: Iuliana Ionita-Laza

Participants

Prof. Dr. Heike Bickeböller
Abteilung für Genetische
Epidemiologie, Bereich Humanmedizin
Universität Göttingen
Humboldtallee 32
37073 Göttingen
GERMANY

Prof. Dr. Stefan Böhringer
Leiden University Medical Center
Postbus 9600
2300 RC Leiden
NETHERLANDS

Prof. Dr. Heather J. Cordell
Institute of Genetic Medicine
Newcastle University
International Centre for Life
Central Parkway
Newcastle upon Tyne NE1 3BZ
UNITED KINGDOM

Dr. Florence Demenais
Institut National de la Santé et de la
Recherche Médicale (INSERM), UMR-S
946
Université Paris Diderot
101, rue de Tolbiac
75654 Paris Cedex 13
FRANCE

Prof. Dr. Michael P. Epstein
Department of Human Genetics
School of Medicine
Emory University
615 Michael Street
Atlanta GA 30322
UNITED STATES

Prof. Dr. Jeanine Houwing-Duistermaat
Leiden University Medical Center
Building 2, Room T-05-058
Einthovenweg 20
2333 ZC Leiden
NETHERLANDS

Prof. Dr. Iuliana Ionita-Laza
Data Science Institute
Columbia University
722 W. 168th St., 6th Floor
New York, NY 10032
UNITED STATES

Prof. Dr. Suzanne M. Leal
Department of Molecular and Human
Genetics
Center for Statistical Genetics
Baylor College of Medicine
One Baylor Plaza, 700 D
Houston TX 77030
UNITED STATES

Prof. Dr. Hongzhe Li
Department of Biostatistics and
Epidemiology
University of Pennsylvania
Perelman School of Medicine
423 Guardian Drive
Philadelphia PA 19014
UNITED STATES

Prof. Dr. Xihong Lin
Department of Biostatistics
Harvard School of Public Health
655 Huntington Ave.
Boston, MA 02115
UNITED STATES

Dr. Sach Mukherjee
Deutsches Zentrum für
Neurodegenerative Erkrankungen e.V.
(DZNE)
Ernst-Robert-Curtius-Straße 12
53117 Bonn
GERMANY

Prof. Dr. Dan Nicolae
Department of Statistics
University of Chicago
5734 S. University Avenue
Chicago IL 60637
UNITED STATES

Prof. Dr. Michael Nothnagel
Cologne Center for Genomics (CCG)
Universität Köln
Weyertal 115b
50931 Köln
GERMANY

Dr. Catalina Vallejos
Institute of Public Health
MRC Biostatistics Unit
Robinson Way
Cambridge CB2 2SR
UNITED KINGDOM

Prof. Dr. Noah Zaitlen
Department of Medicine
University of San Francisco
Byers Hall
1700 4th Street
San Francisco CA 94158
UNITED STATES

Prof. Dr. Andreas Ziegler
Institut f. Medizinische Biometrie &
Statistik
Universität zu Lübeck
Ratzeburger Allee 160
23538 Lübeck
GERMANY