Genetic Epidemiology 36 : 118–171 (2012)

ABSTRACTS FROM THE

ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

1

A Framework To Assess Technology-Specific Error Signatures In Next-Generation Sequencing, With An Application To The 1000 Genomes Project Data

Michael Nothnagel (1) Alexander Herrmann (2) Andreas Wolf (1) Stefan Schreiber (2) Matthias Platzer (3) Michael Krawczak (1) Jochen Hampe (2)
(1) Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany
(2) Department of Internal Medicine I, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany (3) Leibniz Institute for Age Research, Jena, Germany

Next-generation sequencing (NGS) is a key technology in understanding the causes and consequences of human genetic variability. In this context, the validity of NGS-inferred single-nucleotide variants (SNVs) is of paramount importance. We therefore developed a statistical framework to assess the fidelity of three common NGS platforms and to estimate the proportion of false-positives heterozygotes based on read distributions. Application of this framework to aligned DNA sequence data from two completely sequenced HapMap samples as included in the 1000 Genomes Project revealed remarkably different error profiles for the three platforms. Newly identified SNVs showed consistently higher proportions of false positives (3-17%) when compared to confirmed HapMap variants. We show that this increase was not due to differences in flanking sequence features, read coverage or quality, nor was this observation limited to a particular data set or variant calling algorithm. Consensus calling by more than one platform yielded significantly lower error rates (1-4%). This implies that the use of multiple NGS platforms may be more cost-efficient than relying upon a single technology alone, particularly in physically localized sequencing experiments that rely upon small error rates. Our study thus highlights that different NGS platforms suit different practical applications differently well.

2

Efficient Capture Of Allele Frequency Spectra In Resequencing Studies By Selection Of Independent Chromosomes

Todd L Edwards (1) Chun Li (2)
(1) Center for Human Genetics Research, Vanderbilt University (2) Center for Human Genetics Research, Vanderbilt University

Published online 17 January 2012 in Wiley Online Library (www.wileyonlinelibrary.com)
DOI: 10.1002/gepi.21604

Resequencing studies are still more expensive than GWAS on a per-subject basis for accurately calling individual-level genotypes. As a result, subsets of subjects from a larger study often serve as the resequencing sample. To investigate the entire genome the choice of subjects should maximize the number of ancestral lineages to avoid redundant regions that were inherited identical by descent (IBD) from a common ancestor. We present SampleSeq2 (SS2) a greedy algorithm which can select a subset of optimally unrelated subjects, estimate the number of independent chromosomes, G_T, or select the minimum number of subjects with a target G_T. We evaluated SS2 compared to a random draw by simulation and using the Amish study of Successful Aging. Comparing the known value of G_T from simulation to the estimate of G_T, the estimate was close to the true value of G_T, and SS2 increased G_T relative to a random draw across a range of sample sizes. There were 4995 subjects in the full Amish pedigree with 827 in the aging study. We compared SS2 with random selection for K subjects (K=50, 100). For K=50, average G_T was 41.5 using SS2 and 29.7 for random selection. On average, SS2 resulted in 39% more independent genomes. For K=100 the average G_T was 60.6 for SS2 and 39.9 for random selection, 52% more independent genomes. Increasing chromosomes provides a no cost improvement in power, mitigates effects of relatedness on parameter estimates, and increases the yield of alleles from resequencing.

3

On Study Designs For Identification Of Rare Disease Variants In Complex Diseases

Iuliana Ionita-Laza (1) Ruth Ottman (1)
(1) Columbia University

The recent progress in sequencing technologies makes possible large-scale medical sequencing efforts to assess the importance of rare variants in complex diseases. The results of such efforts depend heavily on the use of efficient study-designs and analytical methods.
We introduce here an analytical framework for the analysis of rare variants in family-based or population-based designs. This framework allows us to quantify the enrichment in rare disease variants in families containing multiple affected individuals, and to investigate the optimal design of studies aiming to identify rare disease variants in complex traits. We show that for many complex diseases with small values for the overall sibling recurrence risk ratio, such as Alzheimer's disease, and most cancers, sequencing affected individuals known to have a positive family history of the disease is extremely advantageous for identifying rare disease variants. On the contrary for complex diseases with large values of the sibling recurrence risk ratio sequencing population-based affected individuals is likely to be preferable.

4

A General Framework for Detecting Disease Associations With Rare Variants in Sequencing Studies

D. Y. Lin (1) Z. Z. Tang (1)
(1) University of North Carolina

Biological and empirical evidence suggests that rare variants account for a large proportion of the genetic contributions to complex human diseases. Recent technological advances in high-throughput sequencing platforms have made it possible to generate comprehensive information on rare variants in large samples. We provide a general framework for association testing with rare variants by combining mutation information across multiple variant sites within a gene and relating the enriched genetic information to disease phenotypes through appropriate regression models. Our framework covers all major study designs (i.e., case-control, cross-sectional, cohort and family studies) and all common phenotypes (e.g., binary, quantitative and age-at-onset), and it allows arbitrary covariates (e.g., environmental factors and ancestry variables). We derive theoretically optimal procedures for combining rare mutations and construct suitable test statistics for various biological scenarios. The allele-frequency threshold can be fixed or variable. The effects of the combined rare mutations on the phenotype can be in the same direction or different directions. The new methods are statistically more powerful and computationally more efficient than existing ones. An application to a deep-resequencing study of known or potential drug targets led to a novel discovery of rare variants associated with total cholesterol. The relevant software is freely available.

5

Incorporating Model Uncertainty in Detecting Rare Variants: The Bayesian Risk Index

Melanie Quintana (1) David Conti (1)
(1) Dept. of Preventive Medicine, Division of Biostatistics, Univ. of Southern California

We are interested in investigating the involvement of multiple rare variants within a given region by conducting analyses with two goals: (1) to determine if regional rare variation in aggregate is associated with risk; and (2) conditional upon the region being associated, to identify specific variants that are driving the association. In particular, we seek a formal integrated analysis that achieves both goals. Like previously developed rare variant methods, our framework aims at constructing a risk index based on multiple rare variants within a region. Our analytical strategy is novel in that we use a Bayesian approach to incorporate uncertainty in the selection of variants to include in the index as well as the direction of the effects. Additionally, our approach allows for inference at both the region and variant specific levels. We also extend our approach by introducing a novel informative prior on the marginal inclusion probabilities that incorporates variant specific biological information (such as conservation, genomic region, mutation type, etc...). Using a set of simulations, we show that our methodology has added power over other popular rare variant methods to detect global associations. We also show that there is an added power gain in detecting marginal associations when using the novel informative prior. Finally, we apply the approach to sequence data from the WECARE Study of second primary breast cancers.

6

Graphical Modeling Reveals Primary Genes In The Gene Expression Network Linking Smoking To Atherosclerotic Plaques

Ricardo A Verdugo (1) Maxime Rotival (1) Tanja Zeller (2,3) Philipp Wild (2) Thomas Munzel (2) David-Alexandre Tregouet (1) Francois Cambien (1) Stefan Blankenberg (2,3) Laurence Tiret (1)
(1) Unite Mixte de Recherche (UMRS 937), INSERM (2) Johannes-Gutenberg University Mainz, Germany, (3) University Medical Center Hamburg-Eppendorf, Germany

Smoking is a known risk factor affecting atherosclerosis that has large effects on gene expression in circulating monocytes (BMC Med Genomics 2010, 3:29). We hypothesized that a molecular signature explaining the association between smoking and atherosclerosis may be found in the transcriptome of monocytes. Microarray data and number of atherosclerotic plaques in the carotid was analyzed in nonsmokers (F=407, M=281) and smokers (F=115, M=133). 205 genes differentially expressed by smoking and plaques (FDR<0.1) were considered. A likelihood-based causality test was implemented to identify genes with evidence for a causal effect (smoking->transcript->plaques) among all possible models with three variables. Robustness of the causal inference was assessed by bootstrapping. 18 genes were selected in > 60% bootstraps. The PC-algorithm (J. Mach. Learn. Res. 2007, 8:613) was used to infer the network of conditional independencies between the genes, risk factors and phenotypes. Four genes were independently and directly connected to smoking: PTGDS, SASH1, TJP2, and PPARG in a consensus network from 2000 bootstraps. Together, they explained 81.5% of the covariance between smoking and plaques. PPARG was the only gene directly connected to plaques, representing a hub that merges all network paths from smoking and other risk factors. This gene is a transcription factor with a known role in atherosclerosis. Here we present a candidate topology for its causal network.

7

Robust Methods For Analyzing Secondary Phenotypes In Case-Control Genetic Association Studies

Chuanhua Xing (1) Andrew S Allen (2)
(1) Boston University (2) Duke University

Most case-control genetic association studies will also measure other phenotypes, in addition to case-control status, either because they are readily available or because they are thought to be related to the underlying disease process. As a result, there is considerable interest in assessing the association between genetic variants and these "secondary" phenotypes. However, the biased sampling of a case-control study can result in distortion of association between genetic variants and secondary phenotypes in the population. Thus analysis methods that ignore the case-control design can give biased estimates of the population effect of a genetic variant on a secondary phenotype. We propose an inverse-probablility weighted estimating equation (IPWEE) approach for analyzing secondary phenotypes

in case-control studies. We derive estimators that are appropriate when the disease is rare as well as estimators that utilize existing population level disease prevalence information. We evaluate our methods in an extensive simulation and compare the IPWEE approach with several existing methods. We found that IPWEE had nearly the same power as the (optimal) full-likelihood approach when the model was correctly specified. However, we found that IPWEE was substantially more robust than the full-likelihood approach, both in terms of validity and power, when the model was misspecified.

8

Significance Analysis And Statistical Dissection Of Variably Methylated Regions

Andrew E Jaffe (1) Andrew P Feinberg (1) Rafael A Irizarry (1) Jeffrey T Leek (1)
(1) Johns Hopkins University

It has recently been proposed that variation in DNA methylation at specific genomic locations may play an important role in the development of complex diseases such as cancer. Here we develop one- and two-group multiple testing procedures for identifying and quantifying regions of DNA methylation variability. Our method is the first genome-wide statistical significance calculation for increased or differential variability, as opposed to the traditional approach of testing for mean changes. We apply these procedures to genome-wide methylation data obtained from biological and technical replicates and provide the first statistical proof that variably methylated regions exist and are due to inter-individual variation. We also show that differentially variable regions in colon tumor and normal tissue show enrichment of genes regulating gene expression, cell morphogenesis, and development, supporting a biological role for DNA methylation variability in cancer.

9

PSEA: Phenotype Set Enrichment Analysis - A New Method For Genome Wide Analysis Of Multiple Phenotypes

Janina S. Ried (1) Angela Doring (2) Annette Peters (3) Christa Meisinger (3) Konrad Oexle (4) Juliane Winkelmann (5) Thomas Meitinger (6) H.-Erich Wichmann (7) Norman Klopp (8) Karsten Suhre (9) Christian Gieger (10)
(1) Institute of Genetic Epidemiology, Helmholtz Zentrum Munchen (HMGU), Germany. (2) Institute of Epidemiology I, HMGU, Germany. Institute of Epidemiology II, HMGU, Germany. (3) Institute of Epidemiology II, HMGU, Germany. (4) Institute of Human Genetics, MRI, Technical University Munich (TUM), Germany. (5) Department of Neurology and Institute of Human Genetics, MRI, TUM, Germany. Institute of Human Genetics, HMGU, Germany. (6) Institute of Human Genetics, MRI, TUM, Germany. Institute of Human Genetics, HMGU, Germany. (7) Institute of Epidemiology I, HMGU, Germany. IBE, Chair of Epidemiology, LMU, Germany. Klinikum Grosshadern, Germany. (8) Research Unit of Molecular Epidemiology, HMGU,Germany. (9) IBIS, HMGU, Germany. Dept. of Physiology and Biophysics, Weill Cornell Medical College, Qatar. Faculty of Biology, LMU, Germany. (10) Institute of Genetic Epidemiology, HMU, Germany.

Most genome wide association studies (GWAS) are focused on one phenotype, even if multiple related or unrelated phenotypes are available. However, an integrated analysis of multiple phenotypes can provide insight into their shared genetic basis and may improve the power of association studies. The field of methods for the analysis of multiple phenotypes is still under development.
We present a new method, called "phenotype set enrichment analysis" (PSEA), that uses ideas from gene set enrichment to test sets of phenotypes for association with genes. PSEA does not only allow analyzing predefined phenotype sets, but also identifies new phenotype sets. Apart from application to situations where phenotypes and genotypes are available for each person the method was transferred to GWAS results. For demonstration PSEA was applied to the population based cohort KORA F4 (N=1814) using a panel of five iron related phenotypes and nine whole blood count traits. By confirming associations that were detected in recent large meta-analyses on blood and iron traits, PSEA was shown to be a reliable tool. Many of the identified genes were not found by a GWAS on single phenotypes in KORA F4. Therefore the results suggest that PSEA may be more powerful than a single phenotype GWAS.
PSEA is a valuable method for analysis of multiple phenotypes, which may help to understand biological processes. Its design enables both the use of prior knowledge and the generation of new hypotheses.

10

Interpreting Joint-SNP Analysis Results: When Are Two Distinct Signals Really Two Distinct Signals?

Tae-Hwi Schwantes-An (1) Robert C. Culverhouse (1) Shelina Ramnarine (1) Laura J. Bierut (1) Nancy L. Saccone (1)
(1) Washington University School of Medicine

In genetic association studies, joint analysis of SNPs is often used to distinguish signals in a region of interest. Once a statistically significant SNP is detected, models that include this SNP as a covariate are used to re-analyze the region (or genome); an additional significant SNP is often interpreted as independent from the initial signal. However, this approach does not necessarily rule out the possibility of both SNPs being proxies for a third, untyped causal SNP. We previously reported that rs16969968 and rs588765, two SNPs in the *CHRNA5/A3/B4* region that are modestly correlated (r^2=0.39), show distinct association with smoking behavior, based on joint analysis of European-ancestry data. It would be useful to have a method to determine whether two such SNPs represent genuinely independent signals. We used simulation together with real data to investigate whether these two signals could be explained by a single causal SNP in the region. We identified a space of minor allele frequencies, correlations, and odds ratios for an additively acting causal SNP that could account for the two reported signals. Using 1000 Genomes Project data to impute additional SNPs in this region, we were able to analyze 1867 SNPs in our data. None of them fit the requirements for a single causal SNP model. We conclude that rs16969968 and rs588765 represent two distinct signals. We report this

finding and outline the method, which can be generalized to other joint-analysis results.

11
Rapid Uptake And Use Of A Pharmacogenomic Test For Colon Cancer Treatment
Katrina A.B. Goddard (1)
(1) Kaiser Permanente Northwest

Abstract presented on behalf of the CERGEN Study Team.
Integrated delivery systems with electronic medical records offer unprecedented opportunity within the US to leverage existing infrastructure, health informatics, and specimen resources for broad-based cohorts to support genomic research. These systems can capture the complexity and breadth of real-world clinical decisions. Our study includes seven sites with a combined membership of 5.5 million patients. We investigated the clinical impact of KRAS, a pharmacogenomic test that predicts response to EGFR inhibitors, in 1191 patients with metastatic colorectal cancer diagnosed from 2004 to 2009. In all, 455 subjects (39%) received KRAS testing as part of their clinical care, and 266 (22%) were treated with EGFR inhibitors. Of those who were treated, the proportion who received KRAS testing increased from 47% in 2004 to 100% in 2009. Over 85% of treated patients were KRAS wild type. Most patients with a KRAS mutation (87%) were not treated with EGFR inhibitors. The interval between diagnosis and receipt of KRAS testing decreased from 28 months in 2006 to 4 months in 2009. Most testing occurred after July 2008, following the presentation of clinical trials at a professional oncology meeting. We conducted KRAS testing in codons 12 and 13 for 380 additional subjects, and detected three novel mutations. These findings demonstrate rapid uptake and incorporation of this pharmacogenomic test into clinical decision-making.

12
Integrated Genome-Wide Pathway Association Analysis Using Parallel Computing
Christine Herold (1) Manuel Mattheisen (2) Dmitriy Drichel (1) Tim Becker (1)
(1) German Center for Neurodegenerative Diseases, DZNE Bonn, Germany (2) Harvard School of Public Health, Boston, USA

It has been suggested that the genetic component of complex diseases might be determined by large numbers of common variants with very small effects. For lack of power such tiny effects are not detectable with a one-SNP-at-a-time approach while comprehensive multi-marker modelling is not feasible due to the vast amount of SNP sets that would have to be considered. To overcome these difficulties, prior information on biological pathways can be used to define a treatable amount of meaningful analysis units. In general, pathway association analysis (PAA) tests for an over-representation of nominally significant SNPs in a pathway.
Determining pathway significance is computationally challenging and complicated by varying pathway size and LD between SNPs. We offer a general Monte-Carlo simulation based framework that realizes existing pathway methods as special cases and enables flexible definition of a new pathway score that models independent signals from the same gene in a regression framework and show that it is a good compromise that combines the advantages of existing ideas. The complete work-flow is integrated into a single software tool (INTERSNP), which facilitates the practical conduction of PAA.
Via a power simulation study we show that PAA can outperform GWAS single-marker approach, even if only a fraction of genes in the pathway contain genetic variants relevant to the disease. We present analysis results using KEGG and Gene Ontology for own GWAS on Bipolar disorder.

13
Integration And Visualization Of Genetic And Genomic Data Using A 3-D Video Game Engine
Douglas Hill (1) Jason H. Moore (1)
(1) Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH USA

This is an exciting time in biomedical research due to the availability of technology that allows us to measure tremendous amounts of information about genes, proteins and other biomolecules. However, it is also a challenging time due to the bioinformatics needs associated with integrating, analyzing and interpreting 'omics' data. While we have made great progress in developing the databases and analysis tools for measuring statistical relationships in high-dimensional datasets, the bioinformatics methods for knowledge discovery in the large volumes of statistical results generated from 'omics' analyses are in their infancy. To address this interpretation challenge, we have developed an innovative 3-D visualization approach to the integration and exploration of 'omics' data. The overall goal of this study is to replace the traditional approach of sifting through data in an Excel spreadsheet with an innovative visual approach that presents the results in an interactive 3-D graphical format. To implement this we have harnessed the power of cutting-edge computer graphics technology in the form of 3-D video game engine software (Unity3D). Here, we introduce a 3-D heatmap software package that is able to visualize more than five dimensions of genetic and genomic research results thus permitting the integration of SNPs, gene expression and clinical data, for example. Our 3-D heatmap software is open-source and freely available.

14
Derived SNP Allele Are More Frequently Used As A Risk-Associated Variants In Common Human Diseases
Olga Y Gorlova (1) Jun Ying (1) Christopher I Amos (1) Margaret R Spitz (1) Ivan P Gorlov (1)
(1) UT MD Anderson Cancer Center

The results of more than 200 genome-wide association studies (GWAS) have been published to date. We used the GWAS data to address a question whether ancestral and derived (mutant) alleles serve as risk alleles randomly, which is important for understanding evolution of the genetic control of common human diseases. We found that rarer alleles are more likely to be risk variant and common alleles to be protective. When we analyzed ancestral and derived alleles separately, 0.96 ± 0.01 of rare (0-0.1) derived alleles were risk variants compared to only 0.67 ± 0.04 of rare ancestral alleles being risk alleles. Among minor

alleles, mean proportion of the risk variants was 0.84±0.05 for derived and 0.63±0.02 for ancestral alleles. The mean proportion of the risk variants among derived minor alleles was 0.79±0.06 and for ancestral variants only 0.41±0.05 for early onset diseases (onset < age 30), while the corresponding proportions for the late onset diseases were 0.78±0.07 and 0.62±0.05. Thus early onset diseases are likely to have a rare derived variant as a risk allele. For the late onset diseases risk alleles tend to be more uniform both in terms of population frequency and of the ancestral versus derived status. We also examined whether the population frequency and derived status are independent predictors of the risk variant status. Low frequency but not the derived status was a predictor of being the risk variant overall, but both were predictors in early onset diseases.

15

Inferring The Frequencies And Effect Sizes Of Unobserved Causal Variants By Using The Family History Of Cases

Frank Dudbridge (1) Nick Orr (2) Isabel Dos Santos Silva (1) Olivia Fletcher (2) Julian Peto (1)
(1) London School of Hygiene and Tropical Medicine (2) Institute of Cancer Research

Cases with a family history are enriched for risk variants, and the power of association studies can be improved by selecting cases that have a family history of disease. The gain in power depends both on the type of family history and on the linkage disequilibrium between the tested marker and the underlying causal variant. We show that the allele frequency and effect size of the underlying causal variant can be estimated by combining marker data from studies that ascertain cases based on different family histories. This allows us to learn about the genetic architecture of a complex trait, without having identified any causal variants. We consider several established markers of breast cancer, using estimates from standard case/control studies, from cases with a family history, and from bilateral breast cancer cases. To obtain realistic estimates and to accommodate some prior beliefs, we use a Bayesian estimation of causal effects to show that the causal variants are probably common, with minor allele frequency >5%, and have small effects, with relative risk around 1.2. These results strongly support the common disease common variant hypothesis for these specific disease loci, and agree with recent assertions that synthetic associations of rare variants are unlikely to account for most associations seen in genomewide studies.

16

A Likelihood-Based Framework For De Novo Mutation Detection In Families For Next-Generation Sequencing Data

Bingshan Li (1) Goncalo Abecasis (1)
(1) Department of Biostatistics University of Michigan

Recent studies utilizing high-throughput sequencing indicate that*de novo* mutations play a key role in complex diseases such as sporadic autism and mental retardation. To identify *de novo* mutations, a simple approach is to infer individual genotypes from sequencing assuming all family members are unrelated and compare the genotypes of affected individuals with their parental genotypes. However, due to high sequencing error rates, this naive approach can lead to increased false positive discoveries. In addition, it lacks a handle of assessing the evidence of *de novo* mutation events, making it challenging to sift true *de novo* mutations from a large number of potential candidates. In this study, we developed a likelihood-based framework to detect *de novo* mutations in both nuclear and extended pedigrees by jointly modeling sequencing data in pedigrees and the evidence is assessed via the likelihood ratio of allowing vs. disallowing *de novo* mutations. Through simulations we show that, compared to the naive approach, our framework can achieve markedly improved sensitivity and specificity and a coverage of 30X is required to achieve >98% power without sacrificing specificity. We applied our method to the two trios in the 1000 Genomes Project and observed high concordance with their findings. We hope that our work provides a tool for continuing efforts of hunting genetic factors of complex diseases using family designs through sequencing.

17

Taking Into Account Imprinting And Maternal Genotype Effects Facilitates Detection Of New Genes

Chloe Sarnowski (1) Giovanni Malerba (2) Catherine Laprise (3) Klaus Rohde (4) Miriam Moffatt (5) Patricia Jeannin (1) Marie-Helene Dizier (1) Pier Franco Pignatti (2) William O.C. Cookson (5) Mark Lathrop (6) Florence Demenais (1) Emmanuelle Bouzigon (1)
(1) INSERM, U946, Paris, France (2) Section of Biology and Genetics, Department of Mother and Child, and Biology-Genetics, University of Verona, Verona, Italy (3) Universite du Quebec a Chicoutimi, Chicoutimi, Canada (4) Max?Delbruck?Center for Molecular Medicine (MDC), Berlin, Germany (5) National Heart Lung Institute, Imperial College, London, UK (6) CEA-CNG, Evry, France

A previous genome-wide linkage scan conducted in 640 families from European ancestry detected linkage of 4q35 to asthma-plus-rhinitis phenotype, with increased evidence when accounting for imprinting (LOD=3.14, $P=2.5\times10^{-5}$). We investigated further this region by genotyping a panel of 3,000 SNPs (spanning 20Mb) in 161 EGEA families (206 offspring). To test for association between these SNPs and asthma-plus-rhinitis phenotype, we used two different methods aiming to detect parent-of-origin and/or maternal genotype effects: 1) the Monte-Carlo Pedigree Parent-Asymmetry-Test and 2) the Parent?of?origin?Likelihood ratio Test. Irrespective of the method used, we identified 50 markers associated with asthma-plus-rhinitis with P-value<0.005. These associations were replicated in 245 French Canadian families for four SNPs (0.005?P-values?0.06) under the same model as in the discovery set. The combination of P-values (P_{comb}) from the EGEA and SLSJ samples using Fisher's method enhanced the evidence for association of asthma-plus-rhinitis with SNPs in two genes. The most significant SNP in one gene had $P_{comb}=2\times10^{-4}$ under a parent-of origin effect model while the best SNP in the other gene had $P_{comb}=5\times10^{-4}$ under a maternal genotype effect model. This study highlights that taking into account complex mechanisms, such as imprinting and maternal genotype effect, facilitates the identification of new genes.

Funded by French Min Education & Research, AFSSET, ANR-CEBS, ANR-CEST, GABRIEL

18

Prioritized-GWAS Based On Linkage Information Identifies Novel Putative Loci Influencing Coagulation

France Gagnon (1) Apostolos Dimitromanolakis (1) Guillemette Antoni (2) Angel Martinez (3) Nicholas Greliche (2) Alfonso de Buil (3) Jose Manuel Soria (3) Pierre E. Morange (4) Philip S. Wells (5) David A. Tregouet (6) Lei Sun (7)
(1) University of Toronto Dalla Lana School of Public Health, Toronto, Canada (2) UMR_S 937 INSERM, Paris, France (3) Institut de Recerca Hospital de la Santa Creui Sant Pau, Barcelona, Spain (4) URM_S 626 INSERM, Universite de la Mediterranee, Marseille, France (5) Ottawa Health Research Institute, Ottawa, Canada (6) UMR_S 937 INSERM, France (7) University of Toronto Dalla Lana School of Public Health & Dept. of Statistics, Canada

Frequently, GWAS do not provide overwhelming evidence for association and results are difficult to validate, while oligogenic genome-wide linkage studies (GWLS) are powerful but imprecise. We hypothesized that coupling the information brought by both strategies can increase efficiency. We applied such combined approach to the study of coagulation FXII levels using 5 multigenerational families (n=253) genotyped for >600K SNPs. We conducted a combined analysis involving: 1) GWLS using a subset of the GWAS SNPs; 2) family-based GWAS; 3) combined analysis of GWLS and GWAS results using the stratified false discovery rate (SFDR) approach (Yoo 2010), allowing for SNPs with some evidence of linkage to have higher weight in the prioritized-GWAS. In addition to the structural gene encoding FXII protein, regions on chr 1 and 11 showed promising evidence for both linkage (LOD>1.5) and association ($p=10^{-5}$). When the SFDR control procedure was applied to these results, several SNPs in chr 1 and 11 loci reached genome-wide significance. The strongest evidence for association was observed at the SLC35F3 locus (q-value=0.0003). This locus was then validated in the GAIT study for ?2-glycoprotein 1 and in a case-control study of venous thrombosis, suggesting pleiotropy of SLC35F3. Our prioritized-GWAS incorporating linkage information allowed the identification of novel loci implicated in coagulation, despite the initial absence of genome-wide significance using GWAS or GWLS alone.

19

Detection and Dissection of Pleiotropy for Complex Multivariate Traits

Ingrid B Borecki (1) Qunyuan Zhang (1) Michael A Province (1)
(1) Washington University in St. Louis

Statistical identification of pleiotropy is of growing interest in dissecting the correlated architechture of traits. Meta-analysis methods provide a good initial screen, since they are computationally simple and only need the marginal GWAS results for each phenotype. The correlated-version of such meta techniques can appropriately account for the non-independence of the scans and give valid p-values. However, such meta-tests still can be misleading since they are based upon the wrong null hypothesis (that the gene effects none of the traits), whereas the proper null for pleiotropy is the compound null that the gene does NOT affect all of the traits. We present a novel, likelihood based, mixed model approach, in which we simultaneously model the effect of a set of test SNPs on a vector of multivariate correlated traits and outcomes, using an appropriate set of random effects (depending on study design). The main measure of the pleiotropic effect is the percent of trait correlation that is explained by the test set of SNPs. We compare this model to other pleiotropy models, such as MANOVA, reversed causation logistic models, and simple compound univariate tests (correcting each trait for the others). The statistical operating characteristics of these approaches are evaluated using simulated data and in relation to pleiotropic genes reported by us and others for traits involved in metabolic syndrome.

20

A Phenome-wide Exploration of Novel Genotype-Phenotype Associations and Pleiotropy using Metabo Chip in the PAGE Study

Sarah A Pendergrass (1) Eric S Torstenson (2) Jose-Luis Ambite (3) Christy L Avery (4) Congxing Cai (3) Megan D Fesinmeyer (5) Chris Haiman (6) Gerardo Heiss (4) Lucia A Hindorff (7) Chu-Nan Hsu (3) Charles Kooperberg (5) Loic Le Marchand (8) Yi Lin (5) Tara C Matise (9) Kristine Monroe (6) Kari E North (10) Lynne R Wilkens (8) Steve Buyske (11) Dana C Crawford (1) Marylyn D Ritchie (1)
(1) Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville TN, USA (2) Center for Human Genetics Research, Vanderbilt University, Nashville TN, USA (3) Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA (4) Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA (5) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA (6) Department of Preventive Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, USA (7) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA (8) Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA (9) Department of Genetics, Rutgers University, Piscataway, NJ, USA (10) Department of Epidemiology, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC, US (11) Department of Statistics, Department of Genetics, Rutgers University, Piscataway, NJ, USA

Phenome-wide association studies (PheWAS) explore the relationship between phenotypic structure and genotypic variation. The Population Architecture using Genomics and Epidemiology (PAGE) network is a collection of diverse, population-based studies with data ideally suited for PheWAS. A PAGE PheWAS has been initiated using 161,098 SNPs from the MetaboChip custom array, including single nucleotide polymorphisms (SNPs) (33.2% of the SNPs) identified from genome-wide association studies of metabolic/cardiovascular traits and fine mapping SNPs (62.2%) covering regions around the index SNPs. Tests of association were performed for 247 phenotypes in 6,359 African-American participants for three PAGE studies: Atherosclerosis Risk in Communities (ARIC), the

Multiethnic Cohort (MEC), and the Women's Health Initiative (WHI). Significant novel associations were identified in ARIC for the phenotype of the fibrin D-dimer test, used to diagnose conditions such as deep vein thrombosis (SNP, p-value): rs79266590,5.10E-13 and rs74022438,4.24E-12. These SNPs are in regions previously associated with HDL levels in European descent populations. Future efforts include expanding this PheWAS by including populations of different ancestry and additional phenotypes. This PheWAS has the potential to elucidate a comprehensive picture of genetic associations by uncovering novel SNP/phenotype relationships; identifying pleiotropy; providing novel mechanistic insights; and fostering hypothesis generation.

21

Principal Components Analysis In Expression Profile Studies: Application To Gene Discovery And Eqtn Detection In An LCL-Based Expression Profile Study On Schizophrenia

Harald H H Goring (1) Alan R Sanders (2) Eugene Drigalenko (1) Winton Moy (2) Jubao Duan (2) Pablo V Gejman (2)
(1) Texas Biomedical Research Institute, San Antonio, TX, USA (2) NorthShore University HealthSystem Research Institute and University of Chicago, Evanston, IL, USA

Many of the genetic risk factors for schizophrenia may affect gene regulation. We therefore performed a gene expression study on lymphoblastoid cell lines (LCLs) from the Molecular Genetics of Schizophrenia collection. Expression profiles (Illumina HT-12v4 microarray) are currently available for 903 Caucasians. Significant expression was detected for 27,118 probes at FDR 0.05. To remove the influence of potential confounder effects in LCLs, we measured putative confounders (such as cell growth rate, energy status, and viral load), and we performed principal components (PC) analysis (PCA) to identify major sources of expression variation. We observed substantial structure in the expression data (the top 1, 5, and 50 PCs account for >9, 28, and 47% of variation, respectively), suggesting that careful accounting for PCs may be warranted. This was confirmed in the expression quantitative trait nucleotide (eQTN) analysis, where the number of detected putatively *cis*-acting eQTNs greatly increased when accounting for expression PCs. Using the same adjustment approach, we successfully detected transcripts correlated with schizophrenia. Intriguingly, our expression profile results converge with signals from recent schizophrenia GWAS. We will present details about the PCA approach, eQTN analysis, and schizophrenia-related findings. Our observations support the utility of LCLs for identifying genetic factors influencing behavioral disorders, but also point out analytical challenges.

22

A Statistical Framework For Environmental Epigenetics

Duncan C Thomas (1) Carrie Breton (1) Muhammad T Salam (1) Talat Islam (1) Frank Gilliland (1)
(1) University of Southern California

Several lines of evidence suggest that epigenetic mechanisms such as DNA methylation can mediate exposure-response relationships, including possible transgenerational effects like the effect of grandmaternal smoking during pregnancy on asthma in the second generation seen in the Children's Health Study. We propose a latent variable modeling framework specified in terms of (1) the risk of disease given exposure, genotype, and methylation, (2) the level of acquired methylation given exposure and inherited methylation, (3) the inherited methylation given parental methylation, and (4) the error structure of methylation measurements. By simulation, we show that all parameters of the model are estimable if the model is correctly specified. In a sample size of 1000 3-generation pedigrees, an RR of 3.2 (comparing 0 vs. 100% methylation) for the mediating effect of methylation on exposure-response relationships is detectable with 90% power. Coefficients of variation for other parameter estimates were 1.8% for the exposure effect and 3.2% for the methylation effect in submodel (2), and 6.7% for the transmission effect in submodel (3). We have also shown associations of PM and ozone with DNA methylation of inducible nitric oxide synthase (iNOS), modified by a promoter haplotype in *NOS2A*, and a 3-way interaction between PM exposure, *NOS2A*, and iNOS methylation on eNO measurements. This presentation will provide a unifying framework for synthesizing such observations.

23

A Novel Permutation Strategy to Correct for Population Stratification in Case-Control Studies of Rare Variation

Michael P. Epstein (1) Yunxuan Jiang (1) Karen N. Conneely (1) Richard Duncan (1) Andrew S. Allen (2) Glen A. Satten (3)
(1) Emory University (2) Duke University (3) Centers for Disease Control and Prevention

Many case-control tests of rare variation [Neale et al. PLoS Genet e1001322; Ionita-Laza et al. PLoS Genet e1001289; Li et al. AJHG 87: 728 among others] are implemented in statistical frameworks that prohibit straightforward correction for confounders such as population stratification. Confounding due to population stratification is possible in resequencing studies since rare variants likely are specific to particular ancestral groups. To correct for this confounding, we propose establishing the significance of a rare-variant test using a novel permutation procedure that preserves the population stratification within the sample. Using Fisher's noncentral hypergeometric distribution, we sample disease outcomes for subjects in a permuted dataset in a manner such that the probability a subject is selected as a case is dependent on his/her odds of disease conditional on ancestry (defined as the stratification score). Our permutation framework allows for adjustment of different measures of ancestry and further can accommodate other confounders of interest to researchers. The permutation approach is applicable to any rare-variant association test used in a case-control study and is implemented in a modified version of the R package 'BiasedUrn' for public use. We demonstrate the value of the approach using simulated data based on coalescent models and also illustrate the method using real sequence data from the FUSION study of type 2 diabetes.

24

A Bayesian Analysis and Optimal Design for Association Studies Using Next-Generation Pooled Sequencing Data

Wei E Liang (1) Duncan C Thomas (1) David V Conti (1)

(1) University of Southern California

With its potential to discover a much greater amount of genetic variation, next-generation sequencing is fast becoming an emergent tool for genetic association studies. However, the cost of sequencing all individuals in a large-scale population study is still high in comparison to most alternative genotyping options. While the ability to identify individual-level data is lost (without bar-coding), sequencing pooled samples can substantially lower costs without compromising power to detect significant associations. We propose a hierarchical Bayesian model that estimates the association of each variant using pools of cases and controls, accounting for the variation in read depth across pools and sequencing error. To investigate the performance of our method across a range of number of pools, number of individuals within each pool, and average coverage, we undertook extensive simulations varying effect sizes, minor allele frequencies and sequencing error rates. In general, the number of pools and pool size have dramatic effects on power while the total depth of coverage per pool has only a moderate impact. This information can guide the selection of a study design that maximizes power subject to cost, sample size, or other laboratory constraints. We provide software to find the optimal design, allowing the user to specify a cost function, cost and sample size limitations, and distributions of effect size, minor allele frequency and sequencing error rate.

25

Two-phase Stratified Sampling Designs for Regional Sequencing

Zhijian Chen (1) Radu V Craiu (2) Shelley B Bull (1)
(1) Samuel Lunenfeld Research Institute (2) University of Toronto

By systematic examination of common tag single nucleotide polymorphisms (SNPs) across the genome, the genome-wide association study (GWAS) has proven to be a successful approach to identify genetic variants that are associated with complex diseases and traits. Although the per basepair cost of genotyping has dropped dramatically with the advent of the next-generation sequencing technologies, it may still only be feasible to obtain deoxyribonucleic acid (DNA) sequence data for a portion of available study subjects due to financial constraints. Two-phase sampling designs have been frequently used in large-scale surveys and epidemiological studies where certain variables are too costly to be measured on everyone. We consider two-phase stratified sampling designs for genetic association, in which tag SNPs for candidate genes or regions are genotyped on all subjects in phase 1, and a proportion of subjects are selected into phase 2 based on genotypes at one or more tag SNPs. Deep sequencing in the region is then applied to genotype phase 2 subjects at sequence SNPs. We investigate alternative sampling designs for selection of phase 2 subjects within strata defined by tag SNP genotypes, and develop methods of inference for sequence SNP variant associations using data from both phases. Results from simulation studies are presented for comparison between the proposed methods that use combined data and methods that use data from phase 2 alone.

26

Investigation and Functional Characterization of Rare Genetic Variants in the Adipose Triglyceride Lipase (ATGL) in a Large Healthy Working Population

Stefan Coassin (1) Martina Schweiger (2) Anita Kloss-Brandstatter (1) Claudia Lamina (1) Margot Haun (1) Gertraud Erhart (1) Bernhard Paulweber (3) Yusof Rahman (4) Simon Olpin (5) Heimo Wolinski (2) Irina Cornaciu (6) Rudolf Zechner (2) Robert Zimmermann (2) Florian Kronenberg (1)
(1) Division of Genetic Epidemiology, Innsbruck Medical University, Innsbruck, Austria (2) Institute of Molecular Biosciences, University of Graz, Graz, Austria (3) First Department of Internal Medicine, Paracelsus Private Medical University Salzburg, Austria (4) Department of Inherited Metabolic Diseases, Evelina Children Hospital, London, UK (5) Department of Clinical Chemistry, Sheffield Children's Hospital, Sheffield, United Kingdom (6) Structural Biology Group, Institute of Molecular Biosciences, University of Graz, Graz, Austria

Background: While extensively investigated in the phenotypic extremes, little data about the effects of rare variants is available from general populations. Since *ATGL* catalyzes the rate-limiting step of the lipolysis, it represents an important candidate gene to evaluate the impact of rare mutations on the lipid metabolism.
Methods: We screened the full *ATGL* gene region for genetic variants in 1473 individuals of the SAPHIR Study and investigated the residual catalytic activity of all identified protein mutations. The effects on free fatty acids (FFA) levels were assessed by linear regression and by comparing the highest and lowest 10% quantiles of the distribution.
Results: We detected 55 mostly very rare variants, including 11 novel rare amino acid exchanges. Indeed, 7.7% of the individuals carried a rare variant. Functional investigations revealed a wide spectrum of residual catalytic activities, ranging from total inactivity to wild type activity. Association studies showed a moderate shift of rare variant carriers towards lower FFA levels and a modest accumulation of rare variants in the lower 10% quantile of the FFA distribution.
Conclusion: Our screening reveals a considerable allelic heterogeneity even in a healthy population. Despite a large variability in the residual activity of protein variants, rare ATGL variants exerted only a minor impact on the FFA levels, suggesting that most naturally occurring rare *ATGL* variants may be only mildly deleterious.

27

Rules For Resolving Mendelian Inconsistencies In Nuclear Pedigrees Typed For Two-Allele Markers

Sajjad Ahmad Khan (1)
(1) Department of Statistics, Islamia College University, Peshawar

Gene-mapping studies regularly rely on examination for Mendelian transmission of marker alleles in a pedigree, as a way of screening for genotyping errors and mutations. For analysis of family data sets, it is usually necessary to resolve or remove the genotyping errors prior to analysis. At the Center of Inherited Disease Research (CIDR), to deal with their large-scale data flow, they formalized their data cleaning approach in a set of rules based on PedCheck

output. We examine via carefully designed simulations that how well CIDR's data cleaning rules work in practice. We found that genotype errors in siblings are detected more often than in parents for less polymorphic SNPs and vice versa for more polymorphic SNPs. Through computer simulation, we conclude that some of the CIDR's rules work poorly in some situations and we suggest a set of modified data cleaning rules that may work better than CIDR's rules.

28

A Novel Test For Differentiating Population Stratification From Genotyping Error Using Family Data

Ronnie Sebro (1) Christoph Lange (2) Nan M Laird (3) Neil J Risch (1)
(1) University of California, San Francisco, San Francisco, CA (2) University of Bonn, Germany (3) Harvard School of Public Health, Boston, MA

Identifying population stratification is important for candidate gene association studies using the Transmission Disequilibrium Test (TDT). Although the TDT retains the pre-specified Type I error in the presence of population stratification, these tests may have decreased power in the presence of population stratification. Identifying genotyping error is also important when using the TDT, because genotyping error could result in increased false positive rate. Differentiating population stratification from genotyping error remains a challenge for geneticists. Both genotyping error and population stratification can result in an increase in the observed homozygosity of a sample relative to that expected assuming Hardy-Weinberg Equilibrium (HWE). We show that when family data are available, evaluating the markers that show statistically significant deviation from HWE with the Mating Type Distortion Test (MTDT) - a test based on the mating type distribution, can reliably differentiate genotyping error from population stratification even if only a few markers are genotyped. We simulate data based on several models of genotyping error in previously published literature, and show how this method could be used in practice to assist in differentiating population stratification from systematic genotyping error.

29

Detection of Genotyping Errors in Dense Markers on Large Pedigrees

Ellen M. Wijsman (1) Charles Y.K. Cheung (1) Elizabeth A. Thompson (1)
(1) University of Washington

Accurate linkage analysis results depend on clean genotypes. Error detection is facilitated by Mendelian inconsistent (MI) and Mendelian consistent (MC) error detection checks in pedigrees. However, computational reasons restrict the detection of MC errors for dense markers to small pedigrees. Here we introduce an efficient computational method to detect errors in large pedigrees that allows linkage disequilibrium (LD) between dense markers. We first sample inheritance vectors (IVs) using moderately sparse markers in linkage equilibrium. Conditional on realized sparse IVs, we sample IVs at dense positions, with possible LD. To detect errors, we calculate either the percentage of IVs inconsistent with jointly observed genotypes (S1) or the posterior probability of error configurations (S2). We tested our method on a simulated 5 generation 52-member pedigree with 34 observed subjects. We simulated clean SNPs at 0.5cM density on a 100cM chromosome to infer IVs, and 25000 denser markers with a 0.1% error rate. Of 825 markers with at least 1 error, only 13% were MI. With specificity of 99.9%, both S1 and S2 flagged ∼85% of remaining MC erroneous markers. S2 attributed error to the appropriate individual with 84% accuracy. These results suggest that our method is effective in detecting MC genotyping errors. Moreover, the much quicker S1 closely matches the performance of S2 while not requiring knowledge of marker allele frequencies or the use of an error model.

30

Reconstructing Pedigrees from Genetic Marker Data

Nuala A Sheehan (1) James Cussens (2)
(1) University of Leicester (2) University of York

Population biobanks of large numbers of unrelated individuals have been enormously successful in detecting common genetic variants affecting diseases of public health concern. Attention is now shifting towards investigating gene-gene and gene-environment interaction effects and finding rarer variants for which related individuals are ideally required. In reality most large population studies will contain sets of (undeclared) relatives. Identification of relatives from existing biobanks would further the use of these studies, both to search for rare variants and to adjusting statistical analyses by taking account of relatedness. Although a crude measure of relatedness might suffice more many applications, having a good estimate of the true relationship, or pedigree, would be much more informative if this could be obtained efficiently.
We propose to exploit fast combinatorial optimisation graph-searching algorithms adapted to search for valid pedigrees by imposing appropriate constraints. Our methods are not restricted to small pedigrees and can often guarantee to return a most probable pedigree, conditional on the available information. By delivering multiple high probability pedigrees, we will also allow for the inherent uncertainty in any particular pedigree reconstruction.
1. J. Cussens. Maximum likelihood pedigree reconstruction using integer programming. In *Proc. Workshop on Constraint Based Methods for Bioinformatics*, Edinburgh, July 2010.

31

A Robust Score Test For Family-Based Association Studies Of Complex Diseases With Ordinal Responses, Interactions And Missing Parental Genotypes

M. Fazil Baksh (1)
(1) University of Reading

Analysis methodology for family-based association studies with ordinal measures of disease that were recently developed specifically for ordinal data have consistently been shown as more efficient than procedures based on the assumption of continuous or dichotomized measures. However, current ordinal data methods require specification of a penetrance function, which in rarely known in practice, and their sensitivity to other sources of potential variability, such as extreme ascertainment schemes and family-specific effects, remains unclear. In this presentation,

findings from extensive evaluations oftest procedures for ordinal data under the above conditions motivate development of a score based modification of a recently proposed, efficient, ascertainment-adjustedprocedure. The score procedure is shown to maintain efficiency, yet is robust to model mis-specification, ascertainment and family-specific effects. Secondly, the robust score approach is shown to be unaffected by population substructure and can easily incorporate covariates and interactions. Finally, the persistent problem of missing parental genotypes is addressed via use of an adjusted profile likelihood and the EM algorithm. Findings from simulation studies and real data are presented throughout. Potential application include studies concerned with finding rare variants for common, complex diseases as here families are likely to bemore biologically informative than unrelated individuals.

32

A Novel Statistical Method For Testing The Association Of Rare Variants In A Case-Parent Trio Design

Kwangmi Ahn (1) James Gao (1) Yohan Lee (1) Judith L Rapoport (1) Yin Yao (1)
(1) National Institutes of Health

Next-Generation Sequencing (NGS) technologies are now available in the discovery of rare variants in complex diseases. Due to low frequency and the expected large number of such variants, however, it is difficult to faithfully analyze these data sets. Towards this end, new methods are being developed to increase the statistical power while keeping the level of nominal type I errors in population-based studies low. However, in rare diseases such as Childhood Onset Schizophrenia (COS), the necessary recruitment of large numbers of patients is impractical and, as an alternative, family-based studies are often applied. Here, we combined population-based methods with Transmission Disequilibrium Test (TDT) to test the association of a collection of rare variantsin a case-parent trio design. We also explored the statistical power and nominal type I error under various conditions. As a proof-of-principle, we will then apply these methods to COS sequence data.

33

Fast Association Testing of Genotyped and Imputed SNPs as well as Gene-Environment Interactions in Case-Parent Trio Studies

Holger Schwender (1) Margaret A Taub (2) Mary L Marazita (3) Terri H Beaty (2) Ingo Ruczinski (2)
(1) TU Dortmund University (2) Johns Hopkins Bloomberg School of Public Health (3) University of Pittsburgh

Case-parent trio designs are frequently used to detect single nucleotide poly-mor-phisms (SNPs) associated with disease. A popular procedure for detecting such genetic variations is the genotypic transmission/disequilibrium test (gTDT), which is equivalent to a Wald test based on a conditional logistic regression model. Usually, the parameters of such a model need to be estimated by an iterative procedure, which can be time-consuming if this model should be fitted to hundreds of thousands of SNPs. However, as we will show in our presentation, there exist closed-form solutions for the parameter estimates when testing a SNP with a gTDT under an additive, a dominant, or a recessive model. As exem-pli-fied by applying the gTDT to genome-wide case-parent trio data, the time required for testing all SNPs in such a study reduces from several hours to a few minutes when employing the analytic estimates instead of the iterative fitting procedure. These closed-form solutions can also be used to test interactions between SNPs and binary environmental variables for asso-ciation with disease. Moreover, the analytic estimates can be adapted to test fuzzy genotype calls usually determined for imputed SNPs. Finally, we present a procedure that makes it feasible to compute genome-wide permu-tation-based p-values for the gTDTs under the three models, as well as for a MAX-test in which the maximum over the statistics of these three gTDTs is used as test statistic.

34

Association Mapping Of Multivariate Phenotypes Using Transmission Disequilibrium

Tanushree Haldar (1) Saurabh Ghosh (1)
(1) Indian Statistical Institute

Most complex traits are characterized by quantitative precursors. However, a single quantitative phenotype may not be a sufficiently good surrogate for a clinical end-point trait and it has been argued that it may be more prudent to analyze a multivariate phenotype vector correlated with the end-point trait. The classical Transmission Disequilibrium Test (TDT) for binary traits circumvents the problem of population stratification as it tests for allelic association in the presence of linkage. We propose a simple logistic regression based test with the transmission indicator as the response variable and the multivariate phenotype vector as a covariate. We can analytically show this test to be statistically equivalent to the TDT for binary traits. We perform simulations under a wide spectrum of genetic models and probability distributions of the multivariate phenotype vector to evaluate the power of the proposed procedure and compare with the FBAT approach with identical data. We find that our method yields more power than FBAT if trios with both parents heterozygous are suitably incorporated in our likelihood as well as the multivariate phenotype vector is modeled in terms of the first principal component. We apply our method to analyze a vector of three endophenotypes associated with alcoholism: the maximum number of drinks in a 24 hour period, externalizing symptoms and the COGA diagnosis phenotype in the Collaborative Study on the Genetics of Alcoholism (COGA) project.

35

The Robustness of Generalized Estimating Equations for Association Tests in Extended Family Data

Bhoom Suktitipat (1) Rasika A. Mathias (2) Dhananjay Vaidya (2) Lisa R. Yanek (2) J. Hunter Young (2) Lewis C. Becker (2) Diane M. Becker (2) Alexander F. Wilson (1) M. Danielle Fallin (3)
(1) Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health (2) Department of Medicine, Johns Hopkins Medical Institutions (3) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

Correlations among family members pose a problem that makes genetic analysis computationally intensive in genome-wide association study. The traditional method for handling correlation within family is variance-component analysis, a likelihood-based test (LBT). However, the computational burden of this problem increases with family size and the number of genetic markers. Alternative approaches that do not require complex computation of familial correlations may be preferable, provided that they do not inflate type I error or decrease power. In this study, we performed a simulation study to evaluate alternatives to LBT that employ regression with generalized estimating equations (GEE) in extended family data. We compared the properties of linear regression with GEE applied to an entire extended family structure (GEE-EXT) and GEE applied to nuclear family structures splits from extended pedigrees (GEE-SPL) to simple linear regression with no accommodation of correlations (LM) and a variance-components likelihood-based method (FastAssoc). We observed similar average type I error rates from GEE-EXT (0.051) and FastAssoc (0.049), compared to GEE-SPL (0.059) and LM (0.093). Type I error rates for the GEE-EXT method were marginally higher than the nominal rate when the MAF was < 0.1, but were close to nominal rate when MAF ? 0.2. All methods gave consistent effect estimates and yielded similar power. In summary, the GEE framework appears to work well in extended families.

36

Association Of Genetic Variants With The Risk Of Dysplastic Nevi In Melanoma-Prone Families With And Without CDKN2A/CDK4 Mutations

Xueying Liang (1) Ruth M Pfeiffer (2) William Wheeler (3) Dennis Maeder (4) Laurie Burdette (4) Yeager Meredith (4) Stephen Chanock (4) Tucker A Margaret (2) Alisa M Goldstein (2) Xiaohong R Yang (2)
(1) Division of Epidemiology, CDRH, Food and Drug Administration (2) Division of Cancer Epidemiology and Genetics, NCI/NIH (3) Information Management Services, Inc. (4) Core Genotyping Facility, SAIC-Frederick, Inc.

Cutaneous malignant melanoma (CMM) is an etiologically heterogeneous disease with genetic, host, and environmental factors contributing to risk.Dysplastic nevi (DN) is a strong risk factor for CMM, particularly in melanoma-prone families. Previous studies have suggested a genetic component for DN. However, no candidate genes have been identified. The goal of this study is to identify genetic variants that are associated with DN risk in melanoma-prone families with and without germline mutations of CDKN2A and CDK4. A total of 505 individuals (311 DN) from 53 families (23 CDKN2A/CDK4+ and 30 mutation negative) were genotyped for 851 tagSNPs in 60 genes that were associated with CMM risk in previous candidate gene and GWAS. Conditional logistic regression, conditioning on families, was used to estimate trend p-values, odds ratios and 95% confidence intervals for the association between DN and each SNP separately, adjusted for age, sex, CMM and CDKN2A status. P-values for SNPs in the same gene were combined to yield gene specific p-values. Two genes, CDK6 and XRCC1, were significantly associated with DN after Bonferroni correction for multiple testing (p=0.0001 and 0.0003, respectively), whereas neither gene was significantly associated with CMM risk. Our findings suggest that additional genetic mechanisms may contribute to DN risk in melanoma-prone families. Further analysis of these genes in independent larger datasets is needed to confirm our findings.

37

Adjusting Relatedness in Family Data for Collapsing Association Test of Rare Variants

Qunyuan Zhang (1) Doyoung Chung (1) Ingrid Borecki (1) Michael A. Province (1)
(1) Washington University School of Medicine

Advances of sequencing and genotyping technologies have been facilitating rare variants (RVs) identification, and family data, as potentially enriched with RVs that transmit and congregate within pedigrees, may provide a great source for detecting association between RVs and human complex traits. Most RV association analysis methods developed in recent years, however, are data-driven and permutation-based collapsing methods, which are inapplicable to family data, because direct permutation test ignores and destroys family structure. In this study, we use simulated and real data to show that direct application of these methods to family data will result in a significant inflation of false positive rate (FPR). To deal with this issue, we propose a mixed model based permutation procedure that incorporates family information with different collapsing methods in a permutation test. We demonstrate that the proposed procedure can appropriately adjust the inflation of FPR. We also investigate the power and receiver operating characteristic (ROC) of the procedure for different collapsing methods. Finally we discuss the computational issue and feasible strategies of applying the procedure to large data sets.

38

SNP Effect Decomposition in Family Data using Mixed Models

Nubia E Duarte (1) Suely R Giolo (2) Mariza de Andrade (3) Julia P Soler (1)
(1) University of Sao Paulo (2) Federal University of Parana (3) Mayo Clinic

Gene mapping is a critical step to the understanding of genetic bases of complex diseases. The genomic advances through the new genotyping platforms using single nucleotide polymorphism (SNP) bring up some new issues related to data analysis. An important problem in genetic mapping is the identification of genes associated with complex diseases using data from families and platforms for SNP markers. These platforms produce high dimension data, containing information on more than one million of common genetic variation in the human population (prevalence greater than 1%). The understanding of the feasibility of SNP platforms for mapping genes in family studies is still questionable and offers some analytical challenges. A possible strategy of data analysis is to measure the effect of ordered sets of SNPs, or blocks of SNPs rather than individual SNPs; this is a proposal in the literature that has been addressed for case-control studies, but not in the context of family data. This work proposes a method for SNP subsets selection under a mixed model framework, based on the theory of decomposition of the added variable plot. This

approach allows revealing the relative importance of each SNP inside a block in terms of genetic and residual variance components estimates, as much as, in terms of the each family contribution. This proposal is being implemented by using the R statistical environment, and applied to a simulated and real data from Brazilian families.

39

Evaluation Of Methods To Detect GXG Interaction In Case-Parent Trio Data

Qing Li (1) Joan E. Bailey-Wilson (1)
(1) Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health

The post genome wide association studies (GWAS) era witnesses an increasing emphasis on searching for missing heritability that is not explained by the moderate genetic effects detected via GWAS for complex traits. This study focuses on methods to detect gene-gene (GXG) interactions using a case-parent trio design. Several methods to detect GXG interaction have been extended to trio data, including MDR, logic regression, logic feature selection and Screen and Clean Tool. In addition, gTDT can be used to detect interaction effects involving two SNPs. These methods have been evaluated separately using simulation studies, but it is useful to compare them using the same data under varied interaction patterns to determine whether they have similar behavior when 1) a large set of genes interact to increase the disease risk, 2) the underlying genetic effects are heterogeneous, i.e., due to different causal variants, 3) if both common and rare sequence variants are analyzed? Because all of these methods have very time-consuming procedures, we conducted a small-scale simulation study of several hundred markers, employing various 2-way, 4-way and 6-way GxG interaction effects to generate case-parent trios using the R package "trio". Two marker panels were used: GWAS panel with moderate inter-marker linkage disequilibrium, and GWAS panel markers plus some rare sequence variants. The 5 methods above are evaluated in terms of how often they identify at least 1, 2, 3 or 4 causal variants.

40

Modeling The Non-Inherited Maternal Antigens Effect In Multi-Case Families

Brunilda Balliu (1) Roula Tsonaka (1) Diane van der Woude (2) Stefan Bohringer (1) Jeanine Houwing-Duistermaat (1)
(1) Dep. of Medical Statistics and Bioinformatics, Leiden University Medical Centre (2) Dep. of Rheumatology, Leiden University Medical Centre

Alleles in the HLA-region form an important risk factor for Rheumatoid Arthritis. One of these alleles, HLA-DERAA is hypothesized to have a protective effect not only when inherited but also as a non-inherited maternal antigen (NIMA). For case-parent designs several methods have been developed to test for the NIMA effect (Feitsma et al. PNAS 104, no. 50, 2007). However, these methods are not appropriate for families containing multiple cases and healthy siblings. They ignore the within-family correlation, the information from healthy siblings and do not account for outcome dependent sampling. To address the limitations of the current methods, we use mixed models with family-specific random effects to model the correlation and an ascertainment correction to account for the various sampling schemes. We then estimate the NIMA parameter by maximizing the Ascertainment-Corrected Prospective Likelihood function. We studied the performance of this method using simulations and found that large data sets and/or large families are required to estimate the NIMA effect. Therefore, we propose an approach that combines information from different studies. To illustrate our proposed methodology we used data from 89 multi-case nuclear families (Worthington et al. British Journal of Rheumatology, no. 33, 1994) and 140 twin pairs.

41

Analysis of Rare Genetic Variants in Family-based Sequence Data

Yun Ju Sung (1) Lihua Wang (1) DC Rao (1)
(1) Washington University in St Louis

Recent advances in next-generation sequencing technology provide a cost-effective approach to characterize human genome sequence variation. For detecting rare genetic variants that influence complex traits, several analytical methods are developed, most designed for population-based data. Due to shared ancestry, family-based data are more homogeneous, offering a better opportunity of observing multiple copies of rare variants. We implemented various collapsing methods for family data and applied them to GAW17 family data that were based on exome sequence data from the 1000 Genomes Project. We observed that collapsing methods provided higher power for detecting true causal genes than single-variant analysis (0.59 vs 0.47). Ignoring family relationship lead to substantial increases in false positives. Due to nature, rare variants were present only in a handful number of families and we observed further increase in power by selecting such families. Furthermore, genes detected by family-based methods were different than those detected by population-based methods, indicating that both population and family designs can be complimentary. We also explored how long range correlation observed in the exome sequence influenced both power and false positive rates.

42

Software Packages to Conduct a GWAS On Data from (Twin) Families: a Review

Maarten M.D. Kampert (1) Maria M. Groen-Blokhuis (2) Jouke J. Hottenga (2) Harmen H.M. Draisma (2) Jacqueline J. Meulman (3) Dorret I. Boomsma (2)
(1) Department of Biological Psychology, VU University. Mathematical Institute, Leiden University (2) Department of Biological Psychology, VU University (3) Mathematical Institute, Leiden University

A comparison of statistical software packages for genome wide association studies (GWAS) of family data, including data from monozygotic (MZ) twins, is (yet) absent in the literature. We address accuracy of results and feasibility of the analyses performed in different packages. The current paper compares the freely available packages Assoc (Uh et. Al, submitted for publication), EMMAX (Kang et al, 2010, Nat Genet. 2010; 42: 348-356.), MERLIN (Abecasis et. al., Nat Genet. 2002; 30; 97-101.), OpenMx (Boker

et. al, Psychometrika. 2011; 76:306-317.), PLINK (Purcell et. al., 2007, Am J Human Genet. 2007; 81: 559-575.), and ProbABEL (Aulchenko et. al., BMC Bioinformatics 2010; 11: 134-143). The feasibility of the analyses is compared on several parameters, including computing time, learning curve of the user interface and compatibility with operating systems. The accuracy of the results is compared on empirical and simulated single nucleotide polymorphism (SNP) data with a binary and a quantitative outcome. In the simulated data we vary effect size, SNP information, minor allele frequencies (MAF) and sample sizes (N=1500, 5000 and 10,000) and if feasible, we analyze both dosage and best-guess genotype data. For the empirical data we perform GWAS on the phenotypes Height (quantitative) and Eczema (binary) on 579 genotyped nuclear twin families from the Netherlands Twin Registry.

43

Some New Analytic Procedures In The Sib-Pair Statistical Genetics Package

David L Duffy (1)
(1) Queensland Institute of Medical Research

The Sib-pair statistical genetics package has been under continuous development since 1994, and implements a wide variety of analyses (segregation, variance components, linkage, association) that commonly generate Monte-Carlo based tests of significance. In modern genome wide analyses, P-values are often small, requiring large numbers of Monte-Carlo simulations. Recently, I have adapted the approach of Davis and Resnick (Ann Statist 1984; 12:1467), who describe a simple nonparametric procedure for the estimation of the tail of a distribution function based on a sample from that distribution, "the tail estimation problem". This relies on extreme value theory. Since only a small number of simulated statistics from the tail of the empirical distribution (usually the highest 10-20) need to be retained, this is computationally inexpensive. I present some applications, and show that the estimated P-values are conservative, but considerably better than the usual estimate 1/(1+B) (where B is the number of Monte-Carlo pseudo-samples) in the situation where the observed test statistic exceeded all simulated statistics. In a related vein, I will also discuss the use of delete-*d* jackknife standard errors for familial correlations and related estimators in familial association.

44

Linkage Analysis Of Hepatitis C Virus Infection In An Egyptian Population Living In A Highly Endemic Area

Vincent Pedergnana (1) Mostafa Kamal Mohamed (2) Naglaa Arafa (2) Anne Boland (3) Mohamed Abdel-Hamid (4) Arnaud Fontanet (5) Laurent Abel (1) Sabine Plancoulaine (1)
(1) Inserm (2) Ain Shams University (3) CNG (4) National Hepatology and Tropical Medicine Research Institue (5) Institut Pasteur

Hepatitis C virus (HCV) infects 170 million people worldwide and is thus a major public health problem. Among individuals exposed to HCV, only a subgroup will develop infection as defined by detection of anti-HCV antibodies. Egypt has the highest HCV infection prevalence in the world. A familial study investigating 4,000 individuals from a Egyptian Nile delta village (HCV seroprevalence of 11.8%) showed strong familial correlations for HCV infection after adjustment for known risk factors. A segregation analysis identified a dominant major gene predisposing to HCV infection in young subjects without known risk factors. We investigated 49 large families (313 subjects aged five to 85 years) most likely to contain genetic cases according to the model obtained previously through segregation analysis (i.e. with at least one HCV infected subject under the age of 20 years). Subjects were genotyped with the Illumina linkage IV panel (6,089 SNPs). We performed a genome-wide linkage analysis based on the model provided by the previous segregation analysis using Merlin software. The analysis showed only two regions presenting a LOD-score close to significance for linkage on chromosomes 4 and 20 (LOD-score at 2.9 and 3.1 respectively). This study suggests that at least 2 loci may be linked to HCV infection in endemic areas. Refined mapping of these regions are ongoing to identify variations and corresponding genes involved in host susceptibility to HCV infection.

45

Pointwise-Haplotype Sharing Decomposition of Lod Scores In Association Analysis

Fredrik Olsson (1) Ola Hossjer (1) Keith Humphreys (2)
(1) Stockholm University, Department of Mathematics (2) Karolinska Institute, MEB

Traditional methods for detecting genetic associations with human diseases compare positions in the DNA of seemingly unrelated cases and controls. The Cochran-Armitage test for trends is a commonly used test for detecting such associations. We present a model where we use the DNA from both the tested position and the DNA in a nearby region. We assume that cases share a genetic variant which increases the risk of the disease and that their DNA around that position are inherited from a common founder. By analysing simulated data we show that the method is more powerful than traditional ones.

46

Why Does Linkage Analysis Often Fail With Complex Diseases?

Antonia Flaquer (1) konstantin Strauch (1)
(1) Ludwig Maximilians University Munchen

The question arises why linkage analysis often fails to identify genes when analyzing complex diseases. We have investigated how the power is affected by the pedigree structure, by the parametric or non-parametric test statistics used and by the complex mode of inheritance. Simulations under the alternative hypothesis of linkage were performed to examine the power of the test statistics to detect linkage for each pedigree structure, always considering different complex modes of inheritance. As expected, a small number of pedigrees with less than three affected individuals has low power to map disease genes with modest effect. This holds especially when the mode of inheritance is recessive. Interestingly, the power decreases when unaffected individuals are included in the analysis, irrespective of the

true mode of inheritance. We conclude that under some constraints linkage is an appropriate and robust technique to map genes for complex disease. We provide recommendations regarding the most favorable test statistics, in terms of power, for a given mode of inheritance and type of pedigrees under study, in order to reduce the probability to miss a true linkage.

47

Covariate-Based Linkage Analysis Of Lung Cancer Risk Reveals Novel Loci On 9p21 And 20q12

Claire L Simpson (1) Tiffany Green (1) Betty Doan (2) Christopher I Amos (3) Susan M Pinney (4) Elena Kupert (4) Mariza de Andrade (5) Ping Yang (5) Ann G Schwartz (6) Pam R Fain (7) Adi Gazdar (8) John Minna (8) Jonathan S Wiest (9) Henry Rothschild (10) Diptasri Mandal (10) Ming You (11) Teresa A Coons (12) Colette Gaba (13) Marshall W Anderson (4) Joan E Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland (2) Johns Hopkins School of Medicine, Baltimore, Maryland (3) Department of Epidemiology, University of Texas, M.D. Anderson Cancer Center, Houston, Texas (4) University of Cincinnati, Cincinnati, Ohio (5) Department of Health Sciences Research, Mayo Clinic Rochester, Minnesota (6) Karmanos Cancer Institute, Wayne State University, Detroit, Michigan (7) University of Colorado, Denver, Colorado (8) University of Texas Southwestern Medical Center, Dallas, Texas (9) National Cancer Institute, NIH, Bethesda, Maryland (10) Louisiana State University Health Sciences Center, New Orleans, Louisiana (11) Medical College of Wisconsin, Milwaukee, Wisconsin (12) Saccomanno Research Institute and John McConnell Math & Science Center of Western Colorado, Grand Junction, Colorado (13) Medical College of Ohio, Toledo, Ohio

Lung cancer (LC) is one of the major killers in developed countries (over 150,000 deaths in the US in 2010). Environmental risk factors such as smoking and asbestos exposures are well known. However, only 15% of smokers develop LC, suggesting genetic effects or gene-environment (GxE) interactions.
We previously mapped a major LC susceptibility locus to 6q23-q25, and discovered a rare risk haplotype in linked families that exhibits a GxE interaction between the 6q susceptibility locus and smoking. Genome-wide association studies (GWAS) have suggested other candidate loci with common alleles of small effect on LC risk. However, these loci do not explain all familial risk of LC, suggesting that additional risk alleles exist.
In this analysis, we used LODPAL from the SAGE package to perform a linkage analysis contrasting identity-by-descent sharing in affected relative pairs and discordant relative pairs, while adjusting for a single environmental covariate (a propensity score for LC risk based on pack-years of cigarette smoking and its square, age and sex). Use of only a single covariate increases power in LODPAL since each covariate adds a degree of freedom to the test. Strong evidence of linkage to LC was observed on 6p (LOD=5.72, 74cM) and 6q (LOD=3.25, 173cM), with novel evidence of linkage on 20q12 (LOD=3.42, 63cM). Linkage to lung and throat cancer was observed on 9p21 (LOD=5.66, 64cM). Permutations are ongoing to determine empirical p-values for these LOD scores.

48

Familial Relative Risks And Complex Segregation Analysis Of Isolated Cleft Lip With Or Without Cleft Palate In A High-Prevalence Cluster Of South America

Fernando A. Poletta (1) Eduardo E. Castilla (2) Ieda M. Orioli (2) Juan C. Mereb (2) Juan A. Gili (1) Belen Comas (1) Hugo Krupitzki (3) Jorge S. Lopez-Camelo (1)
(1) a) CEGEBI (Centro de Epidemiologia Genetica y Bioestadistica) at CEMIC - CONICET; b) ECLAMC. (2) b) ECLAMC (The Latin-American Collaborative Study of Congenital Malformations). (3) a) CEGEBI at CEMIC (Centro de Educacion Medica e Investigaciones Clinicas), Buenos Aires, Argentina.

The genetic contribution to the etiology of cleft lip and cleft palate (CL/P) is complex and heterogeneous. An area with high prevalence of CL/P was previously identified in Argentine Patagonia, probably associated with Amerindian ancestry and low socioeconomic status.
The aim of this work was to estimate the mode of inheritance and the number of loci involved in CL/P families from Patagonia prior to planned for linkage/association studies.
The sample included 117 extended pedigrees (2,835 total people) ascertained from CL/P probands registered by ECLAMC hospitals in Patagonia. Family Risk Ratios (FRR) were estimated for first-, second-, and third-degree relatives of CL/P probands, and Complex Segregation Analyses (CSA) were conducted using Pointer and SAGE software.
CSA excluded the Sporadic, Environmental and Multifactorial threshold models, and provided evidence that CL/P is most likely determined by a dominant major gene with incomplete penetrance and with residual familial effects on affection status. Furthermore, FRR for relatives equate well with a major gene (or multiple additive or independent loci). One or two loci interacting epistatically with an poligenic background was also shown to be a plausible alternative.
The high-risk allele frequency estimates of 1 to 9% have important implications with regards to the feasibility of identifying causative loci through the typical commercial or customized genotyping platforms used in genome-wide association studies.

49

Significant Confirmation Of Linkage To Melanoma At 9q21 In An Extended Utah Pedigree

Craig C Teerlink (1) James Farnham (1) Lisa A Cannon-Albright (1)
(1) University of Utah

Linkage to ocular and cutaneous malignant melanoma (CMM) at 9q21 has been previously reported in a Swedish pedigree resource (LOD=3.0) and spanned approximately 81.5-89.4 Mb. We have conducted a genomewide scan for CMM using 34 extended pedigrees genotyped on the Illumina 610k SNP platform, and reduced the set to approximately 27K not in linkage disequilibrium. Linkage analysis used both a general dominant and recessive model, all evidence considered here resulted from the dominant model. We used a heterogeneity TLOD score which uses multipoint information but also optimizes over the recombination fraction similar to a conventional two-point

linkage statistic. The overall het-TLOD score in the 9q21 region was only 0.9. However, a single pedigree had a TLOD score of 2.9 in the same region as the previously reported linkage with five cases sharing a segregating haplotype. A 1-LOD drop for this pedigree delineates a region of interest from 87.5-93.4 Mb on chromosome 9, which overlaps the previously reported region by 1.9 Mb on the qter end. Another single pedigree with four cases had a TLOD score of 0.9 encompassing the linkage peak of the first pedigree spanning approximately 81.8-108.2 Mb on chromosome 9. These results indicate a statistically significant confirmation of linkage to 9q21 for CMM and substantially reduce the region of interest if the linkage evidence in the previous report is due to the same underlying polymorphisms.

50

Parameter Estimation and Quantitative Parametric Linkage Analysis with GENEHUNTER-QMOD

Thomas Kuenzel (1) Konstantin Strauch (2)
(1) IMBE, Philipps University Marburg, Germany; Chair of Genetic Epidemiology, Ludwig-Maximilians-Universitat, Munich, Germany (2) Institute and Chair of Genetic Epidemiology, Ludwig-Maximilians-Universitat, Munich, and Helmholtz Zentrum Munchen, Germany

Linkage analysis for quantitative phenotypes has not been explicitly modeled with genotype dependent distributions so far. Here, we present a likelihood-based method that provides a test for linkage as well as an estimate of different phenotype parameters. The phenotype is modeled as a normally distributed variable, with a separate distribution for each genotype. It is possible to adapt the model to imprinting effects as well as to absence of dominance effects or equal standard deviations of the normal distributions. Parameter estimates are obtained by maximizing the LOD score over the normal distribution parameters with a gradient-based optimization called PGRAD method.
While our new quantitative approach has lower power to detect linkage than the variance components analysis (VCA) in case of a normal distribution, it clearly outperforms VCA for non-normally distributed data. Here, the higher power even goes along with conservativeness, while VCA has an inflated type I error. Parameter estimation tends to underestimate residual variances, but works well for expectation values of the phenotype distributions. We have implemented the method in the program GENEHUNTER-QMOD, a new tool to explicitly model the parameters that underlie quantitative phenotypes in the context of linkage analysis.

51

A Comparison Of CNV Calling Algorithms And Analysis Software

Marcel E Nutsua (1) Michael Krawczak (2) Michael Nothnagel (2)
(1) Institute of Clinical Molecular Biology (2) Institute of Medical Informatics and Statistics

Numerous genome-wide association studies based on single-nucleotide polymorphisms (SNPs) have been performed in the past years and have provided new hints to genetic factors involved in the etiology of many human diseases. However, the reported SNPs have so far failed to explain more than a minor fraction of the observed familial aggregation (heritability) for most diseases. Analyses of structural variants, namely copy-number variations (CNVs), may add to the elucidation of the genetic basis of human diseases. A large number of algorithms for CNV calling based on SNP arrays have been proposed. However, these algorithms differ substantially in their sets of predicted CNVs, prediction quality and implementation. We therefore compared a number of CNV calling algorithms, also including the commonly used Affymetrix Power Tools (APT), QuantiSNP and PennCNV, with regard to the underlying mathematical model as well as the usability of their implementation. We present preliminary results from an ongoing benchmarking study using real-world data.

52

Genome-Wide CNV Association With Alzheimer And Parkinson Disease With Shared Controls In France: A Report Of CNV-Specific Limitations

Matthias Mace (1) Gaelle Marenne (2) Jean-Charles Lambert (3) Suzanne Lesage (4) Jean-Francois Dartigues (5) Christophe Tzourio (6) Philippe Amouyel (3) Alexis Brice (4) Maria Martinez (1) Emmanuelle Genin (2)
(1) Inserm UMR 1043, Univ. Paul Sabatier-Toulouse3, Centre de Physiopathologie de Toulouse-Purpan, Toulouse, France (2) Inserm UMR-S946, Univ. Paris Diderot, Institut Universitaire d'Hematologie, Paris, France (3) Inserm UMR 744, Institut Pasteur de Lille, Univ. de Lille-Nord de France, Lille, France (4) Inserm UMR-S975, Univ. Pierre et Marie Curie-Paris6, Centre de Recherche de l'Institut du Cerveau et de la Moelle epiniere, Pari (5) Inserm U897, Univ. Victor Segalen, Bordeaux, France (6) Inserm U708, Univ. Pierre et Marie Curie-Paris6, Paris, France

Copy number variants (CNV) are an important source of variations in the human genome. Various studies supported the role of CNVs in gene expression and phenotypic traits. Advancement in technologies allows their detection at the genome-wide level by using SNP-array data. However, many issues remain in these microarray-based analyses as compared to the CGH-golden standard.
In this work we performed genome-wide association studies of Alzheimer and Parkinson diseases for CNVs with shared French controls. CNV were called using PennCNV on Illumina 610K data. Genetic data were available for 1369 Alzheimer cases [1], 927 Parkinson cases [2], and 1437 shared controls after CNV specific quality control. Fisher exact test was used for assessing association to CNVs, either at the call or gene level.
We observed an important excess of significant association results with the two diseases, highlighting the presence of artifacts in the CNV assessment differentially distributed in cases and in controls. We explored different strategies to correct for this bias specific to CNV study using SNP-array data.
In conclusion, we think that specific care should be given to microarray SNP genetic data provided with an objective of studying CNVs. SNP-array data are widely available. With additional correction, it may be an interesting opportunity to screen for CNV association.
[1] Lambert et al. (2009) *Nature Genetics* **41**:1094 - 1099
[2] Saad et al. (2011) *Hum Mol Genet* **20**:615-627.

53

Evaluating The Role Of Reference Models In Copy Number Variation Analyses

Ivonne Jarick (1) Anke Hinney (2) Johannes Hebebrand (2) Helmut Schafer (1)
(1) Philipps-University of Marburg (2) University of Duisburg-Essen

Besides single nucleotide polymorphisms (SNPs), copy number variations (CNVs) as another important component of genomic variation have gained much attention with regard to human phenotypic diversity. CNVs, being defined as segments of DNA that are larger than 1 kb in size and that are present at variable copy number in comparison with a reference genome, have been discovered to exist on a large scale. Such duplications or deletions are believed to be related to various diseases. Any such CNV association analysis is intimately connected with the detection of CNV loci and with the assessment of individual copy number states.
To date, SNP genotyping arrays are widely used for CNV identification, although the SNP microarray approaches tend to lead to many false positives and negatives. One striking difficulty when quantifying individual copy number signals, is the lack of a canonical reference genome. Several authors, have compared the performce of mean or median values across differently composed sample subgroups, supposed to be of normal 2 copy number state. Other authors suggested to use median or (trimmed) mean values across publicly available samples, such as HapMap. On the basis of family-based as well as publicly available, genome-wide SNP microarry data (Affymetrix 6.0), we evaluated how a sophisticated per-marker within-sample estimation of reference intensity signals, prior to the application of selected calling algorithms, can influence the precision of CNV detection.

54

Copy Number Variant Detection Using SNP-Chips: Impact Of Calling Performances On Association Tests

gaelle marenne (1) Stephen J Chanock (2) Luis Perez-Jurado (3) Nathaniel Rothman (2) Benjamin Rodriguez (3) Manolis Kogevinas (4) Montserrat Garcia-Closas (2) Debra T Silverman (2) Francisco X Real (5) Nuria Malats (6) Emmanuelle Genin (5)
(1) Inserm U946 / CNIO (2) National Cancer Institute (3) Universitat Pompeu Fabra (4) Institut Municipal d'Investigacio Medica (5) CNIO (6) Inserm U946

Previous studies have highlighted the low performances of copy-number variant (CNV) calling algorithms applied to SNP-array data. This can be problematic in case-control association studies if the level of accuracy of CNV callings differs between cases and controls. The aim of the present study is to determine if there exist some data characteristics that can influence CNV calling performances and study their impact on association tests between CNV and disease. We used data on 21 HapMap samples genotyped on Illumina 1M array as part of the Spanish Bladder Cancer Study to evaluate the performance of four CNV detection algorithms: cnvPartition, PennCNV, QuantiSNP and cnvHap. The callings obtained with these algorithms in 3109 autosomal CNV regions were compared to those reported in the Sanger public database. The sensitivities were around 0.1 and the False Positive Rates (FPR) around 0.3. These figures varied depending on the CNV regions (length, number or density of probes).
To assess the impact of differences in the accuracy of CNV detection in cases and controls on association tests, we performed some simulations of CNVs in cases and controls under different models of correlation between calling performance and disease status. We evaluated the type one error and power rates and studied whether accounting for CNV region characteristics related to sensitivity and FPR reduces false positive association signals.

55

A Step Toward An Integrated Map Of Copy Number Variation And Heritable Gene Expression Through Linkage Of Cnvs In Multigenerational Pedigrees

August Blackburn (1) Harald HH G?ring (2) Angela Dean (3) Melanie A Carless (2) Satish Kumar (2) Joanne E Curran (2) Ravindranath Duggirala (2) John Blangero (2) Donna Lehman (3)
(1) University of Texas Health Science Center at San Antonio department of Cellcular and Structural Biology (2) Texas Biomedical Research Institute Department of Genetics (3) University of Texas Health Science Center Department of Medicine

Investigating copy number variation (CNV) in large pedigrees allows us to assess transmission, heritability, and linkage, thus confirming correct CNV calls and genomic location of common and rare variants present in multiple subjects through inheritance. Using PennCNV and QuantiSNP, we identified 2,937 CNV regions (CNVR) from high density Illumina Infinium genotype data from 1,677 subjects of the San Antonio Family Heart Study and San Antonio Family Diabetes Gallbladder Study, both multi-generational pedigree cohorts of Mexican American descent: 59% overlap with RefSeq genes; 417 CNVRs are within 50Kb of disease associated SNPs from the NHGRI GWAS Catalog; 780 CNVRs are putatively novel. In 675 individuals genotyped on the 1Mduo the first principal component, identified by CNVtools, had a statistically significant ($p<0.05$) heritability for 2,752 of 2,761 CNVRs tested. 592 CNVRs were significantly cis-linked ($p<0.05$), 194 of which were significantly linked after Bonferroni-correction ($p<1.81E-05$). Of these 194, 31 are unambiguously duplications. 211 of the cis-linked CNVRs are common CNVRs identified by HapMap 3. The remainder may be unique to Mexican American populations or rare and significantly enriched in our pedigree data. We are currently refining CNVR calls in the remaining subjects and testing the relationship of CNVs with existing heritable gene expression data from the SAFHS. Our results show the promise of investigating CNVs in extended families.

56

Platinumcnv: A Bayesian Gaussian Mixture Model For Genotyping Copy Number Polymorphisms Using SNP Array Signal Intensity Data

Natsuhiko Kumasaka (1) Hironori Fujisawa (2) Naoya Hosono (1) Astsushi Takahashi (1) Michiaki Kubo (1) Naoyuki Kamatani (1)
(1) Center for Genomic Medicine, RIKEN (2) The Institute of Statistical Mathematics

We present a statistical model for allele-specific patterns of copy number polymorphisms (CNPs) in commercial SNP array data. This model is based on the observation that fluorescent signal intensities tend to cluster into clouds of similar allele-specific copy number (ASCN) genotypes at each SNP locus. To capture the tendency of this clustering to be vague due to instrumental errors, our model allows for the cluster memberships to overlap each other, according to a Bayesian Gaussian mixture model (GMM). This approach is flexible, allowing for both absolute scale differences and X/Y scale imbalances of fluorescent signal intensities. The resulting model is also robust toward unobserved ASCN genotypes in a population that can be problematic for ordinary GMMs. We illustrated the utility of the model by applying it to commercial SNP array intensity data obtained from the Illumina HumanHap 610K platform. We retrieved more than 4,000 allele-specific CNPs, though 99% of them showed rather simple allele-specific CNP patterns with just a single aneuploid haplotype among the normal haplotypes. The genotyping accuracy was validated by two approaches: quantitative PCR and replicated subjects. The results of both of these methods demonstrated genotyping error rates of less than 1% of the median values.

57

Characterization Of Germ-Line Copy Number Variations In Melanoma-Prone Families With And Without CDKN2A/CDK4 Mutations

Xiaohong R Yang (1) Sihui Zhao (2) Ruth M Pfeiffer (1) Margaret A Tucker (1) Alisa M Goldstein (1)
(1) Division of Cancer Epidemiology&Genetics, National Cancer Institute, NIH, DHHS (2) Information Management Services, Inc.

Genomic copy number variations (CNVs) have recently been recognized as a significant source of genetic variation and have been related to disease susceptibility. The goals of this study are to compare frequencies of CNVs in melanoma (CMM) cases and controls and to identify CNVs enriched in CMM cases in melanoma-prone families. Methods: We used genome-wide tiling CGH arrays (Nimblegen 720K exon-focused) to identify germline CNVs in 163 CMM cases and 82 unaffected family member/spouse controls from 50 American melnaoma-prone families with and without germline CDKN2A/CDK4 mutations. We used the Nexus Copy Number^{TM} bulit-in Rank Segmentation algorithm to identify significant CNVs (P=1×10^{-6}; number of probes per segment ?10; $\log_2$ ratio>0.25 for gains and <-0.25 for losses). We compared CMM to spouse controls for differences in CNV frequencies and number of genes and length of DNA segment affected by CNVs using t-test and conditional logistic regression adjusting for age and gender. Results: CMM cases did not show significantly different number of CNVs, genes, and length of CNV segments (either overall or gains and losses separately) compared to spouses. Separate analyses in mutation negative and positive families showed similar results. The top three pathways that were enriched in CMM cases were homophilic cell adhesion, cell adhesion, and nervous system development. We are currently evaluating genes covered in the CNV regions and validating the findings using qPCR.

58

A Comparison Of Methods To Detect Complex Trait Rare Variant Associations Implementing The Rarepower Tool

Gao Wang (1) Suzanne M Leal (1)
(1) Baylor College of Medicine

There is currently great interest in detecting rare variant associations using next generation sequence data. A large number of rare variant association methods which aggregate variants across a region e.g. a gene, have been developed. It is not clear which existing method is the most powerful. To compare methods both realistic phenotype models and spectrum of variants across a region must be generated. Power was compared for 12 methods to detect associations for both qualitative (case-control) and quantitative traits (extreme & random sampling). For each method, power was determined for different scenarios which include: 1.detrimental & protective variants within a region; 2.misclassification; 3.different underlying population demographic model for both Africans & Europeans and 4.gene size. It was observed that there is not a single method that is most powerful in all situations and the majority of rare variant methods had only small incremental differences in power. Methods which were developed to detect associations when both protective and detrimental variants are within an associated region are usually less powerful than more general rare variant association methods. The evaluated methods also vary greatly in their computation efficacy. The RarePower tool with its user friendly graphical interface can be used to determine sample sizes for rare variant association studies under a large variety of complex trait and population demographic models.

59

Quality Control And Assurance Strategies To Optimize Variant Calling/Detection Using Next Generation Sequencing (NGS) Data

Hua Ling (1) Kurt Hetrick (1) Kimberly Doheny (1) Elizabeth Pugh (1)
(1) Center for Inherited Disease Research (CIDR), Johns Hopkins University

High sensitivity and few false positive variant calls from NGS data are critical for downstream data analysis using collapsing methods. Usually after variants are called, a set of quality filters are applied to remove uncertain calls. Then they are filtered by position in relation to capturing assay to define NearBait, OnBait and OnTarget calls for either whole exome sequencing (WES) or custom-targeted sequencing (CTS). To investigate the performance of quality filters, we examined a number of QC metrics for the filtered and remaining variant calls at varying level of the filtering thresholds. These metrics include TiTv ratio,%dbSNP, SNP and genotype calling reproducibility for blind and HapMap duplicates. We also compare these QC metrics by comparing NGS variant calls to GWAS array genotypes using concordance and heterozygous sensitivity. We found the optimal cutoff for many filters varies by sequencing depth, they performed well in for our WES samples but for CTS with much higher depths too many true variants were filtered. In both WES and CTS, comparison to GWAS array data using concordance and heterozygous sensitivity are useful indicators for sequencing data quantity and quality. Heterozygous sensitivity is approximately linearly correlated with

mean bait coverage at 8x and deviation from the expected trend suggests an over filtering of variants.

60

Hunting For Rare Susceptibility Variants Using In Genome-Wide-Association Data Of Parkinson's Disease

mohamad saad (1) Suzanne Lesage (2) Alexis Brice (3) Maria Martinez (1)
(1) Inserm, UMR 1043 (2) Inserm UMR_S975, Paris, France (3) UMR_S975, AP-HP, Pitie-Salpetriere Hospital, Paris, France

The common disease-multiple rare variant hypothesis has recently received much attention. Different statistical methods have been developed for testing the hypothesis that collections of rare variants are associated with a disease in a case-control sampling design [Am J Hum Genet 2008, 311-21; PLoS Genet 2009, e1000384; Genet Epidemiol 2010,188-93; Am J Hum Genet 2010, 832-8]. The detection of rare polymorphisms requires high-quality whole-genome sequence data of a large number of cases and controls and remains expensive. Alternatively, recent studies showed that rare variants can be imputed into existing GWAS datasets from publicly available sequenced data, as the 1000 Genomes Project (i.e., pseudo-sequencing data). In addition, it has been reported that collapsing-based test of single-marker dosages in pseudo-sequencing data may have greater power than haplotype-based methods in genotyped data [Am J Hum Genet 2010,718-35]. Here, we evaluate such approaches in the genome-wide association data of Parkinson's disease [Hum Mol Genet 2011, 615-27]. Imputations were conducted with IMPUTE. We tested association at the gene level. The lowest rates of positive signals (exceeding a given significance threshold) are observed when the cumulative sum test is limited to the uncommon variants. We also show that depending on the approach, different sets of genes may be identified as being associated to the disease. Thereby, drastically different conclusions might be reached.

61

Study Design Considerations To Improve Power In Association Tests For Rare Variants

Ingo Ruczinski (1) Rasika Mathias (1)
(1) Johns Hopkins University

The assumption that common complex diseases are attributable in part to allelic variants that are reasonably common in a population is often termed the "common disease, common variant" hypothesis, and is the underlying rationale for genome-wide association studies (GWAs). While GWAs have been successful identifying hundreds of such genetic variants associated with many complex diseases, the individual variants typically only represent a small increment in risk for any particular disease, and together, can usually explain only a small proportion of the familial clustering (heritability) observed. Thus, the paradigm has shifted somewhat towards whole exome and whole genome sequencing approaches to assess the effects of rare variants (with possibly larger effect sizes), which are poorly tagged by standard genotyping arrays. In this presentation, we focus on family and population based study design considerations, and show how family records can be leveraged to improve power even in population based studies.

62

Estimating Genetic Effects and Quantifying Missing Heritability for Rare Variant Complex Trait Association Studies via Sequence Data

Suzanne M Leal (1) Dajiang J Liu (1)
(1) Baylor College of Medicine & Rice University

Complex trait rare variant association (RVA) studies using sequence data are being widely performed. Analyzing rare variants individually is extremely underpowered; therefore many powerful RVA methods have been developed, which are all based upon jointly analyzing multiple variants within a gene. After an association is identified, it is also important to estimate genetic parameters of interest and quantify the proportion of heritability explained by the gene. A drawback of RVA methods is that it is not possible to tease apart causal from non-causal variants. Consequently, the causative-variant-effect is not estimable. We describe how to efficiently estimate the locus-average-effect. Due to the presence of non-causative variants, genetic variance explained by the locus-average-effect will be underestimated but provides a lower bound for the true underlying locus genetic variance. It is also shown how an estimate of the upper bound for the true locus genetic variance can also be obtained. An additional problem is due to the winner's cure the naive estimator can be seriously inflated. The bias is quantified and it is shown that even for poorly powered studies a boot-sample-split procedure can be used for any RVA method to greatly reduce the bias of the genetic estimates. Not only are these methods vital for estimating the amount of missing heritability due to rare variants, but they are also important for designing replication studies and risk prediction.

63

A Powerful And Flexible Framework For Rare-Variant Analysis

Dalin Li (1) Xiuqing Guo (1) Jerome I. Rotter (1)
(1) Cedars-Sinai Medical Center

Rare variants may explain some of the missing hereditability in current GWAS. The traditional approaches for rare-variant analysis such as Combined Multivariate and Collapsing method (CMC) can be powerful when all causal variants are deleterious, but suffer from low power when a small proportion of the rare variants are protective. The C-alpha approach, which tests the over-dispersion of rare variants count in cases and controls, offers a powerful and robust alternative. However this approach is not flexible and its application might be hampered by practical issues like population stratification. Here we propose a powerful and flexible method for rare-variant analysis. We first apply a Poisson-distribution based analysis for each single rare variant, in which covariates can be easily incorporated. Under the null hypothesis, the test statistics for rare variants can be shown to be symmetrically distributed with mean of 0 and variance of 1 (but not necessarily normally distributed). We then proposed an omnibus test for multiple rare-variants in a given gene region via testing mean, variance, and skewness simultaneously. Our

simulation demonstrated that the proposed approach can be 5-10% more powerful than the C-alpha approach when most of the causal rare variants are deleterious and much more powerful (with more than 20% increase in power) comparing to CMC across all scenarios. We then applied the proposed analysis framework to the GAW17 dataset successfully.

64

A Unifying Framework For Analyzing Rare Variant Quantitative Trait Associations In Selected Samples: Application To Sequence Data

Dajiang J. Liu (1) Suzanne M. Leal (2)
(1) Rice University (2) Baylor College of Medicine

Next-generation sequencing is being used to map rare variant (RV)/quantitative trait (QT) associations. Sequencing individuals with extreme QT or combining publically available phenotyped cohorts e.g NHLBI ESP-the exome sequencing project can be applied to reduce cost and improve power. Many methods for mapping RVs do not have a likelihood model and cannot estimate genetic parameters. Some leading methods e.g. the weighted sum statistic(WSS) do not allow controlling for covariates, such as sequence read depth and population substructures. Failure to control for confounders can lead to spurious associations which cannot be eliminated by permutations. Additionally, some methods are developed for binary traits, and will be underpowered in QTL mapping. To overcome these limitations, a unifying method was developed to detect and interpret RV/QT associations(UNI-QTL) in any QT study with known designs, extending all methods in a rigorous likelihood framework. The performance of all extended tests was extensively evaluated by analyzing real and simulated data under realistic population genetic and complex trait models. We show that the power of almost all extended RV tests can be consistently improved for QTL mapping. There does not exist a uniformly most powerful method, but the extended methods of WSS, variable threshold(VT) and kernel based adaptive cluster(KBAC) perform well under most scenarios. In conclusion, UNI-QTL will be greatly important for sequence based QT studies.

65

Association Testing For Rare Variants Via Pooled Design

Ioanna Tachmazidou (1) Maria De Iorio (2) Mario Falchi (2)
(1) Wellcome Trust Sanger Institute (2) Imperial College

It is thought that much of the genetic susceptibility to complex diseases is due to rare variants. DNA pooling could prove a cost effective approach for identifying associations between rare variants and disease.
We propose to use pooled DNA samples to estimate allele and haplotype frequencies for rare SNPs, and we investigate the performance of collapsing methods in the context of a binary trait. We use the EM algorithm to estimate haplotypes from pooled DNA, and we use a logistic regression framework to model case/control outcome as a function of i) the presence or absence of a minor allele at any causal variant, and ii) the proportion of causal variants at which a minor allele is present. This approach is compared to the collapsing method with the estimated allele frequencies in a simulation study of N individually genotyped subjects versus N*K pooled subjects, where K is the pool size and N is the number of pools.
We find the allele frequency test to be more powerful than the haplotype-based tests even in the presence of high LD. The differences in power become smaller as N and/or the genetic signal increases. The methods also seem robust to error due to unequal amounts of probe material. In conclusion, next generation sequencing of DNA pools provides a cost-effective approach for studying association of disease with rare SNPs, offering an advantage in terms of ability of detecting rare variants and power for association testing over individual DNA sequencing.

66

Detecting Rare Variants In Admixed Populations

Xiaofeng Zhu (1) Huaizhen Qin (1)
(1) Department of Epidemiology and Biostatistics, Case Western Reserve University

Admixed populations such as African-Americans and Mexican-Americans have many advantages in mapping genes underlying complex traits. One of the methods in admixed populations is the admixture mapping which exploits the long-range disequilibrium generated by the admixture between genetically distinct ancestral populations. It is possible that the evidence detected by admixture mapping may be caused by multiple rare variants. It is still unclear how we can efficiently detect these variants. Here we proposed a novel statistical approach that uses the admixture mapping evidence to improve statistical power of testing rare variants. Power and type I error of the proposed approach will be presented using simulations.

67

A New Approach To Prove Involvement Of A Rare Variant In Disease Susceptibility

Bertram Muller-Myhsok (1) Herve Perdry (2) Broet Philippe (2) Francoise Clerget-Darpoux (2)
(1) MPI Psychiatry (2) INSERM U669

Large-scale sequencing projects are increasingly commonly used to demonstrate the involvement of rare variants in disease susceptibility. The effects of such rare variants as measured in genotypic relative risks are believed to be considerably higher than those of common variants. Genotypic relative risks as high as 2 or even 4 may be realistic.
This leads to two differing but coinciding consequences for the demonstration of the effect of a rare variant in a complex disease. Firstly, we can demonstrate that in index patients recruited from affected sib-pairs the frequency of the rare alleles is greatly increased compared to samples composed of unrelated cases. To give an example, for an allele with a population allele frequency of 1% and a genotypic relative risk of 4 the expected frequency of unrelated cases carrying the risk variant is around 8%; whilst in patients with an affected sibling this frequency is more than doubled to roughly 17% thus lowering the necessary sample size from more than 1100 to less than 400. Secondly in affected sib pairs an additional orthogonal aspect of genetic information, IBD conditional on the genotype of the index patients, can be used. This information is absent from case-control studies and offers a handle to model the genetic effect of a given variant.

We would like to point out that these concepts are readily transferred also to approaches considering ensembles of variants in a given genetic region rather than single variants.

68

Semiparametric Maximum Likelihood Method For Rare Variant Analysis Under Quantitative Trait-Dependent Sampling Designs

Yildiz E Yilmaz (1) Jerald F Lawless (2) Shelley B Bull (1)
(1) Samuel Lunenfeld Research Institute of Mount Sinai Hospital (2) University of Waterloo

For rare variant analysis, selection of individuals for sequencing according to their quantitative trait (QT)-value can improve cost-efficiency. In such sampling designs, standard linear regression methods that treat the QT as the dependent variable, ignoring the selection, are not valid. We consider likelihood methods developed for settings with expensive covariates missing by design (Zhao et al., 2009, Biometrical Journal 51, 1-14) and apply semiparametric maximum likelihood (SML) to fit linear regression models and test for association of the QT with a rare variant score (e.g. total count of minor alleles) as the covariate. Analysis of sequencing data for an entire cohort is the ideal against which we compare the likelihood method for QT-dependent sampling designs where only a fraction of the cohort is sequenced. We evaluate efficiency for various designs, including extreme phenotype selection, and designs in which all individuals have a non-zero probability of being selected, but those with extreme phenotypes have a proportionately higher probability. In finite sample simulation studies, the SML method yields nearly unbiased and relatively efficient estimates of regression coefficients. Moreover, we find the SML test statistic for association to be valid under the null, and under certain designs, for example moderate signals with a 50% sampling rate, power can approach that of tests based on complete data.

69

Are Studied Phenotypes up to the Next Generation Sequencing Challenge?

Aldi T. Kraja (1) Ingrid B. Borecki (1) Michael Y. Tsai (2) Jose M. Ordovas (3) Paul N. Hopkins (4) Robert J. Straka (5) James E. Hixson (6) Michael A. Province (1) Donna K. Arnett (7)
(1) Div. of Statistical Genomics, Washington U. Sch. of Medicine, St. Louis, MO (2) Laboratory Medicine & Pathology, U. of Minnesota, Minneapolis, MN (3) Nutr. and Genomics Laboratory, Tufts U., Boston, MA (4) Div. of Cardiovascular Genetics Research, U. of Utah, UT (5) Dep. of Experimental and Clinical Pharmacology, U. of Minnesota, Minneapolis, MN (6) Human Genetics Cntr., Sch. of Public Health, U. of Texas Health Science Cntr. at Houston, Houston, TX (7) Dep. of Epidemiology, U. of Alabama, Birmingham, AL.

Lipid profiles of nuclear magnetic resonance (NMR) compared to traditional lipid measures have been controversial. Are NMR phenotypes which provide discerning particles that go to the crux of complex traits, up to the NGS challenge? The GOLDN study, with its post-prandial lipid challenge and 3 week fenofibrate treatment intervention, quantified NMR and classical lipid profile changes at 0, 3.5 and 6 hours before and after drug exposure in over 1000 subjects. We used latent modeling based on factor analysis to capture the intricacies of twenty NMR correlated phenotypes by accepting 4 factor scores at each 6 time points. These latent factors and the original traits were the phenotypes for exploring additive effects on a genome wide scan of 'hybrid' genotypes of 2,543,887 SNPs. Two markers, rs727477 an intron in*SLC8A1*gene, and rs9349940 in between*CD83*and*JARID2*genes were at 10^{-9}, 45 unique SNPs at 10^{-8}, and 302 unique SNPs were in the range of 10^{-7}p-value threshold. Because 10^{-7}p-value range is a weak genome wide threshold, we built a lipid's gene network, including interactions, pathways and GO terms, where was observed enrichment with genes related to lipid metabolism. We conclude that in pharmacogenomics/nutrigenomics research NMR lipid profiles can play an important role to better understand lipid metabolism.

70

Genome-Wide Association Analysis Of Rare Variants With Crohn's Disease

Reedik Magi (1) Andrew P Morris (1)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford

Genome-wide association studies of common variants have identified Crohn's disease (CD) susceptibility loci, which account for 23% of the heritability of the disease. The aim of this study was to assess the evidence for association of CD with rare genetic variation, defined here to have minor allele frequency (MAF) less than 1%, through imputation up from a scaffold of existing GWAS genotyping data.
We performed imputation in 1,748 CD cases and 2,940 controls of European descent. Imputation was undertaken using reference panel from the 1000 Genomes Project. We tested for association of CD with accumulations of minor alleles at rare variants within genes, modelling disease status as a function of the proportion of rare variants at which an individual carries at least one minor allele in a logistic regression framework.
The strongest signal of association of rare variants with CD was observed for *PTGER4* ($p=1.3\times10^{-6}$, genome-wide significant correcting for 30,000 genes). This gene contains common variants that have been previously associated with CD and plays an important role in immune response. Strong evidence of association of rare variation with CD was also observed for *CD247* ($p=2.5\times10^{-6}$). Defects in the gene are a primary cause of primary T-cell immunodeficiency.
Our results highlight the potential for the identification of rare variant associations using existing GWAS genotyping data, supplemented with imputation, without the need for costly re-sequencing experiments.

71

Alternative Test Statistics for Sparse Data in Genome-wide Association and Whole-genome Sequencing Analysis

Shelley B Bull (1) Michael A Rotondi (2)
(1) Samuel Lunenfeld Research Institute and Dalla Lana School of Public Health, University of Toronto (2) Samuel

Lunenfeld Research Institute of Mount Sinai Hospital, Toronto, Canada

Issues of inference from sparse SNP data are of heightened concern as the field expandsinto whole-genome sequencing studies. With low minor allele frequency the probability of observing individuals heterozygous for rare alleles is low while the effect size may be large; maximum likelihood estimates may not have finite values (complete separation in logistic regression,or monotone likelihood in Cox regression); and test statisticsnot attain asymptotic distributions especially at extreme testing levels. Aggregation approaches to rare variant testing are designed to improve power to detect association, but p-values based on standard asymptotic theory, particularly Wald-type test statistics, nevertheless may be misleading. Several authors consider alternative inference based on a penalized likelihood (Firth, 1993, *Biometrika* **80**, 27-38; Heinzeand Schemper, 2002, *Statistics in Medicine* **21**, 2409-19), with extensions to multinomial outcomes (Bull et al., 2002, *Computational Statisticsand Data Analysis* **39**, 57-74) and non-linear models (Kosmidisand Firth, 2009, *Biometrika*, **96, 793-804**). In binomial logistic regression, this leads to a generalization of Haldane's statistic for sparse 2×2 table analysis with covariates; however, it has seen little application to date in genetics. Here, with examples of sparseSNP data analysis, we illustrate some features of penalized likelihood ratio and score test statistics using available R code.

72

Power to Detect Gene-Environment Interactions Involving Rare Variants

Remi Kazma (1) Niall J. Cardin (1) John S. Witte (1)
(1) Department of Epidemiology and Biostatistics and Institute for Human Genetics, University of California San Francisco

Gene-environment (GxE) interactions and rare variants are often cited as possible sources of the unexplained heritability in studies of common variants and complex diseases. Not accounting for GxE interactions may decrease the power to detect disease association with a genetic variant, in particular if the interaction is strong and the genetic marginal effect is weak. In the last decade, many methods to account for GxE interactions in genetic association studies have been developed, albeit with very few successes in practice. Recently, the focus in genetic epidemiology has shifted toward the study of rare variants. Many aggregating strategies have been suggested to improve the power of association tests with rare variants. However, very little is known about the benefits of accounting for GxE interactions when studying rare variants. To determine whether accounting for GxE interactions improves the power to detect associations with rare variants, we carried out a simulation study to compare the power of various methods under different disease models. We extended three methods, which aggregate rare variants, to account for GxE interactions using multivariate and multinomial models (the cohort allelic sum test, a weight-based aggregation test and the C-alpha test). We used as reference the single marker test accounting for GxE interactions. Our results show how one can account for GxE interactions when studying rare variants and that power can be improved in specific cases.

73

Simrare: A Program To Generate And Analyze Sequence-Based Data For Association Studies Of Quantitative And Qualitative Traits

Biao Li (1) Gao Wang (1) Suzanne M Leal (1)
(1) Baylor College of Medicine, Rice University

Many methods have been developed to detect complex trait rare variant associations. In order to fairly compare type I error and power it is necessary to generate data using realistic models. Currently it is difficult to compare methods because there is no standard to generate data and often comparisons are biased. SimRare generates variant data for "gene" regions using forward-time simulation which incorporates population demographic models. It is possible to generate both case-control and quantitative trait data. The phenotypic effects of variants can be detrimental, protective or non-causal. For causal variants the effect size can be determined by frequency or purifying selection coefficients. It is possible to model interactions (GxG & GxE) and confounders (environmental or genetic). SimRare has a user friendly interface. To evaluate novel association methods R libraries can be imported or conversely the simulated data can be written to external files. SimRare has built in functions to evaluate the performance for 15 currently available rare variant association methods. SimRare can also be used to evaluate computational efficiency, control of confounders and ability to detect interactions. Using SimRare it is demonstrated that there is not a single most powerful method and most methods are not robust to confounders. Additionally due to the computational speed some methods are more advantageous to use especially for the analysis of exome/genome sequence data.

74

Association Testing In Sequencing Studies: Accommodating Risk And Protective Variants

Abra Brisbin (1) Brooke L. Fridley (1)
(1) Mayo Clinic

Many existing methods address the question of identifying associations between a phenotype and a set of rare variants. However, the majority of these methods implicitly assume that the direction of effect is the same for all rare variants, and are subject to loss of power in the presence of both risk and protective rare alleles. We developed a new method for analysis of rare variants, the Difference in Minor Allele Frequency (D-MAF), which allows combined analysis of common and rare variants, and makes no assumptions about the direction of effects. We tested our method and 9 others on simulated genomic regions with varying mutation and recombination rates, and a variety of phenotypic models. We found that several methods, including D-MAF, performed well when all rare variants were either risk alleles or neutral; however, D-MAF and two other methods, C-alpha and CMC, outperformed the others when protective variants were present. D-MAF can also be extended to the analysis of pooled sequencing data, for which many collapsing methods are not applicable.

75

Boosting Ensemble As A Tool To Combine Genetic Signals

Wei Yang (1) C. Charles Gu (1)
(1) Washington University School of Medicine

Common variants identified by genome-wide association (GWAS) studies typically account for a small portion of trait variability, implicating other factors (copy number variation, rare variants, etc.). Effectively combining such weak signals becomes a pressing issue in post-GWAS era. Boosting is a learning method capable of building ensemble classifiers by combining many simple ones. But the performance in high-dimensional data is unclear. We evaluated boosting algorithms using a simulated GWAS dataset (24,487 SNPs in 3,205 genes in 697 samples) with common and rare causal variants. Collapsing was applied to combine effects of rare variants. We first studied boosting as a means of gene-set analysis and compared with gene-set enrichment test (GSEA) and our variable set enrichment test (VSEA). Power of the boosting was often better than GSEA and sometimes better than VSEA. We then tested using boosting to prioritize risk genes from all available variants. Current implementations of boosting were able to evaluate 1,000s variants in minutes. However, it can only reliably pick up the strong risk factors. Moreover, as number of variables increases demand on memory became prohibitively large. In summary, application of boosting to gene-set analysis can potentially improve performance for testing gene-sets with relatively small number of variants. However, direct application of boosting to GWAS data is impractical using current implementations and further investigation is warranted.

76

Identifying The Genetic Variation Of Gene Expression Using Gene Sets: Application Towards Pharmgkb Gene Sets

Ryan P Abo (1) Gregory D Jenkins (1) Leiwei Wang (1) Brooke L Fridley (1)
(1) Mayo Clinic

Genetic variation underlying gene expression levels in humans may provide key insights to the molecular mechanisms of human traits. Current methods to map genetic loci or SNPs associated with gene expression have applied previously developed linkage and/or association methods without further consideration of the high dimensionality and multiple testing issues involved with such approaches when applied towards genomewide SNP and mRNA expression data. Here we present a novel approach to model and test the association between genetic variation and mRNA expression levels using gene sets, referred to as eGeneSet (eGS). Using gene sets reduces the dimensionality and multiple testing by grouping the SNP and mRNA expression data based on *a priori* biological knowledge. We apply eGS to analyze SNP-expression associations with cell line genomic data using the pathways defined in PharmGKB. We found a large number of significant eGS associations in which the most significant associations arose between genetic variation and mRNA expression from the same pathway. Our proposed method effectively addresses a key limitation in eQTL studies by reducing the multiple testing and testing SNP-expression associations between biologically relevant gene sets. By applying our method to PharmGKB gene sets we have identified notable associations involving drug pathways which might lead to insight into genomic variation and its potential influence on therapeutic response.

77

Multi-Ethnic Fine-Mapping Of Cis Expression-Qtls With Fixed-Effect Meta-Analysis

Christopher P Grace (1) John C Whittaker (2) Julie Huxley Jones (2) Andrew P Morris (1)
(1) The Wellcome Trust Centre For Human Genetics, Oxford University (2) GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, UK

We have considered genotyping data made available through the Phase 2 HapMap Project for 270 samples across three ethnic groups. We have tested for association of up to 2.2 million SNPs with *cis*-expression of up to 18,000 transcripts in lymphoblastoid cell lines. We combined the data across the ethnic groups by using fixed-effect meta-analysis. Differences in patterns of linkage disequilibrium between ethnic groups would be expected to improve the mapping resolution of causal variants within *cis*-eQTL signals.
Our trans-ethnic meta-analysis identified genome-wide significant ($p<1\times10^{-12}$) evidence of association for 4,368 *cis*-eQTLs. Of these, there was evidence of heterogeneity (Cochran's Q $p<1\times10^{-3}$) at 513 *cis*-eQTLs, which is significantly higher than expected by chance (binomial test: expected: 4, 95% interval: 1 to 10). However, within many cis-eQTL signals with homogeneous effects across ethnic groups, there is a noticeable improvement in mapping resolution. For example through trans-ethnic meta-analysis, causal variants of the probe GI_21553312-S (CHURC1) (peak signal, $p=3.85\times10^{-304}$ Cochran's Q p-value=0.008) were localised to a 47kb interval, compared to 78kb in CEU alone.
Our results demonstrate the potential for trans-ethnic meta-analysis of cis-eQTL studies to improve mapping resolution, but also highlight the need for the development of novel methodology to take account of heterogeneity in allelic effects between distantly related populations.

78

Beyond Comparing Means: The Usefulness Of Analyzing Interindividual Variation In Gene Expression For Identifying Genes Associated With Cancer Risk And Development

Ivan Gorlov (1) Jinyoung Byun (1) Christopher Logothetis (1) Olga Gorlova (1)
(1) The University of Texas MD Anderson Cancer Center

An identification of genes associated with cancer development is typically based on a comparison of mean expression values in normal and tumorous tissues to identify differentially expressed (DE) genes. Interindividual variation (IV) in gene expression is indirectly included in identification of DE genes. We explored the possibility of using the IV in gene expression to identify candidate genes associated with cancer development. We focused on two common cancers: prostate (PCa) and lung (LC).
We found that compared with all the other genes in the human genome, cancer-related genes have greater IV in normal tissues and a greater increase in IV during the transition from normal to tumorous tissue. We have identified

a number of genes that have a bimodal (e.g. low vs high) expression in normal tissue and unimodal expression in tumorous tissues. Typically only one mode (low or high) is presented in tumor tissue. This suggests that for some genes the observed changes in mean expression might be due to selection rather than modification. Differences between distinct expression patterns in PCa risk may explain why only one mode is observed in tumorous tissue. We noted that GWAS-identified PCa genes have higher variance compared to the average gene in the human genome and also have lower variance in tumor tissue compared to the normal one. In conclusion, our results suggests that the analysis of IV may be used to identify novel candidate genes associated with cancer development.

79

Robust Statistical Methods For Genome-Wide Eqtl Analysis

Mattias Rantalainen (1) Chris Holmes (1)
(1) Department of Statistics, University of Oxford

Expression Quantitative Trait Loci (eQTL) analysis enable characterization of how common genetic variants influence the expression of individual genes. EQTLs are commonly analysed using a linear regression model including relevant covariates, an additive genetic effect and assuming a Gaussian error term. However, Gaussianity may not hold in noisy biological data (e.g. gene expression), which may be more heavy-tailed or have outliers present. Such departures from model assumptions may result in an increased rate of both type I and type II errors. Careful model checking can reduce the risk of spurious findings, however, this may be prohibitively expensive in eQTL analysis with a high number of models. In this case robust statistical methods, which are less sensitive to departures from the common model assumptions, provide an attractive alternative. We compare eQTL results between the conventional linear model and a robust alternative (M-estimator) in two publicly available eQTL data sets and in a simulation study in respect to bias in estimates, concordance of hypothesis test results and statistical power. Our results indicate that robust statistical models provide more reliable eQTL results over conventional linear models with reduced number of both type I and type II errors. This suggests that unless careful model checking can be carried out on each evaluated model, robust alternatives provide a valuable alternative to conventional linear models in genome-wide eQTL analysis.

80

Genome-Wide Epistasis Screening For Crohn's Disease

Elena S Gusareva (1) Kristel Van Steen (1)
(1) Systems and Model, Montefiore Institute, University of Liege; GIGA-R, University of Liege

Genome-wide association (GWA) studies of Crohn's disease have identified numerous genes. However, a substantial portion of the heritability of this disease remains unexplained. Some gene variants, not detectable via main effects GWA study, may manifest themselves only in interaction with other variants. To search for interacting genes involved in the regulation of Crohn's disease, we performed GWA epistasis screening in a large human cohort (1851 cases/2938 controls) belonging to the Wellcome Trust Case Control Consortium (WTCCC). All subjects were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip). SNPs that passed our quality control (359,479 SNPs) were processed in Biofilter (a software package that looks for candidate epistatic genes contributing to disease risk) giving rise to 14,185 SNPs. Subsequent MB-MDR epistasis screening discovered four pairs of interacting SNPs on chromosome 4q35.1 and eight pairs on chromosome 11q23.2. The identified pairs of SNPs were confirmed with synergy-based measures. Notably, despite their mapping to the same genomic regions, the interacting SNPs were not in LD ($r^2 < 0.5$). Our findings support the idea of close chromosomal localization of two pairs of interacting genes that are involved in development of Crohn's disease.

81

A Robustness Study To Investigate The Performance Of Parametric And Non-Parametric Tests Used In Model-Based Multifactor Dimensionality Reduction Epistasis Detection

Jestinah M Mahachie John (1) Elena Gusareva (1) Francois Van Lishout (1) Kristel Van Steen (1)
(1) Montefiore Institute and GIGA-Research (University of Liege)

Model-Based Multifactor Dimensionality Reduction (MB-MDR) is data mining technique to identify gene-gene interactions among 1000nds of SNPs in a fast way, without making assumptions about the mode of genetic interactions. By construction, one of the implementations of MB-MDR involves testing one multi-locus genotype cell versus the remaining cells, hereby creating two imbalanced groups for trait distribution comparison. To date, for continuous traits, we have adopted a standard F-test to compare these groups. When normality assumption or homoscedasticity no longer hold, highly inflated results are to be expected. The power and type I error control of MB-MDR under these assumptions has been thoroughly investigated in Mahachie John et al [1].
The aim of this study is to assess, through simulations, the effects of ANOVA model violations on the performance of Model-Based Multifactor Dimensionality Reduction (MB-MDR). We quantify their effect on MB-MDR using default options, but at the same time introduce alternative options with increased performance. The better handling of imbalanced data using robust approaches [2] within a MB-MDR context is exemplified on real data for asthma-related phenotypes.
1. EJHG (2011), Early view
2. David Freedman, Statistical Models: Theory and Practice, Cambridge University Press (2000), ISBN 978-0521671057

82

An Ensemble Pipeline to Enable Detection of Epistasis in Genomic Data

Benjamin J. Grady (1) Tom Cattaert, (2) Kristel Van Steen (2) Marylyn D. Ritchie (1)
(1) Center for Human Genetics Research (2) Systems and Modeling Unit, Montefiore Institute, University of Liege, Grande Traverse 10, 4000 Liège, Belgium1

Although our capability to gather information on genetic variation has increased exponentially, our ability to understand the complexity of the relationship between this variation and disease has been unable to keep pace. While GWAS has had great success in uncovering novel variation associated with disease, these variants account for vanishingly small modulations in disease risk and leave unexplained large amounts of the heritability estimated for many traits. One of the primary issues with current treatment of genetic data lies in an implicit assumption of simplicity, with most analyses focusing only on effects of singular loci in isolation. In this study we describe an ensemble approach to exploring epistasis in genetic studies. The approach includes filtering the data prior to analysis to reduce both computational and multiple-testing issues. For filtering, the Evaporative Cooling software and a genotypic Chi-square test are utilized. Within the genetic data passing filtering criteria, we perform separate, exhaustive pair-wise analysis utilizing the Model-Based Multifactor Dimensionality Reduction (MB-MDR) software and a Likelihood Ratio Test (LRT) containing terms for interaction between the pair of genetic loci being examined. A permutation procedure is used to determine significance of results from each method. Through simulations, we show that use of this data analysis pipeline increases the power to detect a breadth of potential genetic interactions.

83

Entropy Based Genetic Association And Gene-Gene Interaction Tests

Xin Wang (1) Mariza de Andrade (1)
(1) Mayo Clinic

In the past few years, several entropy-based tests have been proposed for testing either single SNP association or gene-gene interaction. These tests are mainly based on Shannon entropy and have higher statistical power when compared to standard ?2tests. In this paper, we extend some of these tests using a more generalized entropy definition, Renyi entropy, where Shannon entropy is a special case of order 1. The order ? (>0) of Renyi entropy weights the events (genotype/haplotype) according to their probabilities (frequencies). Higher ? places more emphasis on higher probability events while smaller ? (close to 0) tends to assign weights more equally. Thus, by properly choosing the ?, one can potentially increase the power of the tests or the p-value level of significance. We conducted simulation as well as real data analyses to assess the impact of the order ? and the performance of these generalized tests. The results showed that for dominant model the order 2 test was more powerful and for multiplicative model the order 1 or 2 had similar power, which indicated that Shannon entropy depends on the underlying genetic model and it is not necessarily the most powerful entropy measure for constructing genetic association or interaction tests.

84

Application Of A Novel Method For Testing Gene-Gene Interactions To Genome-Wide Association Studies Of Seven Complex Human Diseases

Kanishka Bhattacharya (1) Andrew P Morris (1)
(1) Wellcome Trust Centre for Human Genetics

Genome wide association studies (GWAS) have proved to be immensely successful in discovering novel loci which contribute effects to common complex traits. However, most of these studies have employed single SNP approaches and have explained only a small fraction of the trait heritability. One plausible explanation for this gap in our understanding is interaction between SNPs, a phenomenon commonly observed in animal models. When considering pairs of SNPs, an exhaustive scan of the genome is the most powerful approach to detect interactions. However, the biggest hurdle to researchers in performing such interaction studies is the computational burden of the analyses.
IntRapid (http://www.well.ox.ac.uk/INTRAPID) employs a novel two-stage strategy to conduct computationally efficient, exhaustive pair-wise SNP interaction scans. In the first stage, rapid interaction tests are performed to screen pairs of SNPs for a more computationally intensive, second stage of interrogation within a generalised linear modelling framework. We have applied IntRapid to GWAS of seven diseases from the Wellcome Trust Case Control Consortium. Our results highlight multiple pairs of SNPs with evidence of interaction, at a nominal significance level of $p<10^{-10}$, which warrant follow-up in additional cohorts. None of these interacting SNPs show strong marginal effects ($p<10\text{-}^{5}$), and hence would not have been identified through traditional single-SNP approaches.

85

Natural And Orthogonal Interaction Framework For Modeling Gxg And Gxe Interactions

Jianzhong Ma (1) Feifei Xiao (1) Christopher I Amos (1)
(1) University of Texas M.D. Anderson Cancer Center

In the Natural and Orthogonal Interaction (NOIA) framework, originally developed for analysis of quantitative trait loci, orthogonal estimates of parameters for the statistical model do not change in reduced models, and are thus convenient for model selection for genetic architecture of traits. In this study, we first extended the NOIA framework to all the three reduced genetic models: additive, recessive and dominant, and for any combinations among the full and reduced models. We then extended the NOIA framework to model the interaction between a binary environmental exposure and a gene with either a full model or any one of the reduced models. The statistical model of NOIA also directly leads to a proper, orthogonal decomposition of the genetic variance, which makes it easy to compute important measures, such as the heritability of a trait. We also proposed to apply the NOIA coding to the case-control data analysis by treating the genetic effects as log-odds of the disease. Simulation results showed that, although power for detecting the interaction effects are the same for the statistical model and the usual functional model, at least for some of the scenarios in our simulations, the statistical model seemed to be more powerful in detecting the main effects. We applied the NOIA coding to the case-control melanoma data and found that the false positive rate seemed to be lower than the method using the usual coding.

86

A Maximum Likelihood Approach to Prioritize SNPs for Interactions Using Variance per Genotype

Wei Q. Deng (1) Angelo J. Canty (2) Guillaume Pare (1)
(1) Departments of Clinical Epidemiology & Biostatistics, McMaster University (2) Department of Mathematics and Statistics, McMaster University

Gene-gene and gene-environment interactions have been suggested as a source of "missing heritability" in complex trait genetics yet are difficult to identify due to low statistical power. Although an exhaustive search for interactions on a whole genome basis is feasible, most interactions fail to be statistically significant after Bonferroni correction given the large number of tests involved. In the context of quantitative traits, we have previously developed a prioritization scheme - variance prioritization - based on differences in variance between the three possible genotypes of biallelic genetic variants. We showed that this method has increased power over exhaustive search under a variety of scenarios. Nevertheless, our method uses Levene's test to compare the variance of a quantitative trait for each genotype, without taking into account that variance will either increase or decrease with the number of minor alleles when interactions are present. To address this issue, we herein propose a maximum likelihood approach to test for differences in variance between genotypes when variances are expected to be ordered. We further derive a closed-form representation of the likelihood ratio test (LRT) statistics, which greatly improves computational speed in genome-wide settings. We use simulations to investigate the accuracy and power of LRT in detecting interactions. We show that LRT is more powerful than both exhaustive search and variance prioritization using Levene's test.

87

Bloat Control Methods Substantially Reduce Computation Time For Detecting Gene X Gene Interactions In ATHENA

Emily R Holzinger (1) Scott M. Dudek (1) Eric S. Torstenson (1) Marylyn D. Ritchie (1)
(1) Vanderbilt University

Recent advancements in human genetics research have spurred the increase in popularity of the genome-wide association study (GWAS) to identify the etiology of common disease. GWAS have been successful at identifying thousands of single nucleotide polymorphisms (SNPs) that associate with human traits; however, these variations only account for a tiny portion of the overall estimated heritability. One hypothesis is that the heritability lies in non-linear interactions that would be missed by single-locus analyses. Testing for interactions is not a trivial task due to the computational burden of exhaustive analysis. To address this problem, our lab has incorporated two machine learning methods, Grammatical Evolution Neural Networks (GENN) and Grammatical Evolution Symbolic Regression (GESR) into ATHENA (Analysis Tool for Heritable and Environmental Network Associations). Both methods use genetic programming (GP)-based techniques to evolve a random population of solutions to find the correct model. One issue with GP-based methods is that the solutions tend to bloat. This results in a substantial increase in computation time and over-fitting. For this experiment, we compared two bloat-control techniques, prune and plant (P&P) and double tournament (DT), using simulated data. Our results show that P&P and DT significantly reduce computation time without reducing power.

88

Interaction Detection with Random Forests in High-Dimensional Data

Stacey Winham (1) Xin Wang (1) Mariza de Andrade (1) Robert Freimuth (1) Colin Colby (1) Marianne Huebner (1) Joanna Biernacka (1)
(1) Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic

Identifying variants associated with complex human traits in high-dimensional data is a central goal of genome-wide association studies. However, complicated etiologies such as gene-gene interactions are ignored by typically applied univariate analyses. Random Forests (RF) are a data-mining technique that can be used to predict disease status based on a large number of loci, allowing for potentially complex genetic models such as gene-gene interactions. Although designed for prediction, RFs give measures of variable importance (VI) that can be used to rank SNPs and are gaining popularity as a filter approach that considers interactions. We explore the ability of VI measures to detect interactions and their potential effectiveness as filters, particularly as the data becomes increasingly high-dimensional. We compare performance of various RF VI measures to univariate logistic regression under different data-generating models for complex disease and different RF tuning parameter settings. We show that as the total number of predictors increase, probability of detection declines more rapidly for interacting SNPs than for non-interacting SNPs and the advantage over univariate logistic regression is lost, indicating that the VI measures are capturing marginal effects rather than allowing for potential interactions as often claimed. To improve interaction detection, weproposealternative VI measures based on permutation decrease in accuracy.

89

Comparison Of Different Methods For Detecting Gene-Gene Interactions In Case-Control Data

Tom Cattaert (1) Jose A Rial Garcia (2) Elena Gusareva (1) Kristel Van Steen (1)
(1) Montefiore Institute and GIGA-R, University of Liege, Belgium (2) Universidade da Coruna, Spain

It is generally believed that epistasis makes an important contribution to the genetic architecture of complex disease, and numerous statistical and bioinformatics methods have been developed to detect it.
We compare several state-of-the-art epistasis detection methods in terms of empirical power, type-I error control, and CPU time. The methods compared include Model-Based Multifactor Dimensionality Reduction (MB-MDR) [1, 2], BOolean Operation-based Screening and Testing (BOOST) [3], EPIBLASTER [4], Random Jungle (RJ) [5], Logistic Regression and PLINK.
Our comparative study is based on an extensive simulation study using different two-locus models, exhibiting both main effects and epistasis [3]. In these simulations, 100 SNPs are generated, no LD between them. All genotypes are assumed to be in Hardy-Weinberg equilibrium.

Furthermore, 2 disease-associated SNPs are selected, with MAFs set to 0.1, 0.2 and 0.4. The MAFs of the non-disease associated SNPs are uniformly distributed on [0.05, 0.5]. In order to achieve high accuracy in empirical power estimation, all simulation settings involve 1000 replicates. All methods are applied to WTCCC Crohn's Disease data.

[1] Calle, M.L. et al. (2008), Tech. Rep. No. 24, Dep. of Systems Biology, Univ. de Vic
[2] Cattaert, T. et al. (2011), Ann. Hum. Gen. 75, 78-89
[3] Wan, X. et al. (2010), Am. J. Hum. Gen. 87, 325-340
[4] Kam-Thong, T. et al. (2011), Eur. J. Hum. Gen. 19, 465-471
[5] Schwartz, D.F. et al. (2010), Bioinf. 26, 1752-1758

90

Epistatic Interactions among Genes with Known Evidence of Protein-Protein Interaction

Yan V. Sun (1) Wei Zhao (2) Sharon LR Kardia (2) Kerby A Shedden (3)
(1) Department of Epidemiology, Emory University (2) Department of Epidemiology, University of Michigan (3) Department of Statistics, University of Michigan

Epistatic associations could be responsible for a portion of the heritability in complex traits that is not explained by additive genetic effects. However, undirected epistatic studies on a genome-wide scale are challenging, due to the need to control false positives when considering an enormous number of candidate variants that may participate in epistasis. We implemented two complementary ideas to make epistatic studies more feasible. First we integrate prior biological knowledge, using physical interactions in protein-protein interaction (PPI) networks to limit the search space; second we focus on interactions among statistical summaries of genetic variation, rather than interactions between individual SNPs. This allows us to identify a reduced number of composite genetic variates for interaction testing. We performed three analytical approaches using data from a GWAS of serum lipid levels. We assessed for interactions between Principal Component Analysis (PCA), between variates defined using Tukey's test for non-additivity, and using variates defined using a hybrid of Tukey's method with PCA. Using an initial set of 99 genes, we identified 11 pairs of genes with PPI of their protein products. Corrected for multiple testing, we identified epistatic associations with total cholesterol level. Combining with appropriate statistical models for testing epistasis, PPI may assist in identifying epistatic interactions and generating hypothesis of molecular functions.

91

Interaction Of Oxidative Stress Pathway Genes With Particulate Matter And Tobacco Smoke On The Course Of Airflow Obstruction During 11 Years

Ivan Curjuric (1) Medea Imboden (1) Rachel Nadif (2) Ashish Kumar (3) Sally LJ Liu (1) Harish Phuleria (1) Thierry Rochat (4) Christian Schindler (1) Florence Demenais (5) Nicole M Probst-Hensch (1)
(1) Swiss Tropical and Public Health Institute SwissTPH, Basel, and University of Basel, Switzerland (2) Inserm, CESP Centre for research in Epidemiology and Population Health, U1018, and Universite Paris Sud 11, Villejuif, France (3) Swiss Tropical and Public Health Institute, Switzerland, and Wellcome Trust Centre for Human Genetics, University of Oxford, UK (4) Division of Pulmonary Medicine, University Hospitals of Geneva, Switzerland (5) Inserm U946, Fondation Jean Dausset- Centre d'Etude du Polymorphisme Humain (CEPH), and Universite Paris Diderot, Paris, France

Background: Ambient air pollution and smoking effects on airflow obstruction might be better explained by testing interactions with entire genes and pathways than just single nucleotide polymorphisms (SNPs).
Methods: Interactions of 12'679 SNPs from 152 oxidative stress genes with exposure to particulate matter <10um in diameter (PM10) or packyears of smoking on decline in the ratio of forced expiratory volume in one second over forced vital capacity (FEV1/FVC) were examined in 650 adults from the SAPALDIA cohort. SNP interaction p-values were projected onto gene and pathway levels using the adaptive rank truncation product (ARTP) method. Interaction estimates for exposure contrasts of one interquartile range (IQR) and percent variability in decline explained were compared between PM10 and packyears for strongest interacting SNPs in nominally significant genes ($p<0.05$).
Results: Seven genes interacted nominally, *CRISP2* significantly ($p_{interaction}=3.0\times10^{-4}$), and *SNCA* marginally with PM10. Five different genes interacted with packyears. The strongest SNP-interaction in *CRISP2* accelerated decline by 1.1%. Top SNPs of nominally significant genes showed similar interaction effects across exposures. Models including all SNPs explained 18.9% of decline variability for PM10, 15.6% for packyears, and 22.5% for both combined. Only *SNCA* SNPs were significant at SNP level.
Conclusions: PM10 and tobacco smoke impact similarly on FEV1/FVC decline, but via different susceptibility genes.

92

Smoking Modifies The Effect Of Lipoprotein Lipase Gene Polymorphism On Serum HDL-C Concentration In Japanese General Population

Hidetoshi Kitajima (1) Kayo Kurotani (2) Keizo Ohnaka (3) Ryoichi Takayanagi (4) Ken Yamamoto (1)
(1) Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University (2) Department of Preventive Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan (3) Department of Geriatric Medicine, Graduate School of Medical Sciences, Kyushu University, Fukuoka (4) Department of Medicine and Bio-Regulatory Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

Genetic and lifestyle factors regulate serum high-density lipoprotein cholesterol (HDL-C) concentration. Recently, genome-wide association (GWA) studies have identified loci influencing HDL-C. Epidemiological studies have also reported lifestyle factors regulating HDL-C. The aim of this study is to identify the gene-lifestyle interactions that modify the level of HDL-C in the members of the Japanese general population, who have different genetic backgrounds and lifestyles from Caucasians. We genotyped 15 SNPs for 9,281 subjects of the Fukuoka Cohort study in Japan. The seven SNPs were significantly associated with HDL-C adjusted for sex, age, BMI, drinker, smoker, physical activity, carbohydrate, protein, saturated fatty acid, and n-3 highly unsaturated fatty acid. A significant interaction

(P=0.00082) between rs10503669 (*LPL*) and current smoking was observed after adjustment for covariates. The adjusted means of HDL-C were 56.1 mg/dl, 56.7 mg/dl, and 56.6 mg/dl among current smokers and 60.4 mg/dl, 63.9 mg/dl, and 66.2 mg/dl among non-current smokers in risk allele homozygosity, heterozygosity, and non-risk allele homozygosity of rs10503669, respectively. The results indicate that the positive effect of the non-risk allele of rs10503669 on HDL-C level is negated by smoking.
In summary, the seven loci were confirmed to be significantly associated with HDL-C in the Japanese general population. The effect of rs10503669 (*LPL*) on HDL-C might be modified by current smoking.

93

Imputation-Based Genome-Wide Gene-Environment Interaction Screening In Colon Cancer Using A Case-Only Design

Sabine Siegert (1) Ute Nothlings (1) Michael Nothnagel (2)
(1) Institute of Experimental Medicine (2) Institute of Medical Informatics and Statistics

Genome-wide association studies (GWAS) have identified a large number of potential genetic susceptibility factors for numerous diseases by testing for marginal effects. However, these variants cannot explain more than a small proportion of the heritability of diseases. Gene-environment (GxE) interactions have been suggested to play an important role in the etiology of common diseases and may account for some fraction of this "missing heritability". Indeed little work has been done to investigate these interactions in GWAS data. In this study, we investigated GxE interactions between genetic markers and smoking status, alcohol consumption and body mass index (BMI) in the etiology of colon cancer on a genome-wide scale, using a large North-German cohort. Marker genotypes were imputed to enable the combined analysis of cohorts that were genotyped on different arrays and as a means of additional quality control. In a first stage, we screened for gene-environment interactions using a case-only design, which has been shown to provide superior power under a range of disease etiology scenarios. In a second stage, interactions of top-ranking SNPs with environmental covariates were then validated in an independent case-control study. We present preliminary results of this analysis.

94

Two-Phase Case-Control Study Design For Biomarker Measurements In Geneenvironment Interactions Studies

Duncan C Thomas (1) Sandrah Eckel (1) Kiros Berhane (1)
(1) University of Southern California

Two-phase case-control designs are a cost-efficient way of obtaining biomarker data by subsampling on genes *G*, exposure *E*, and disease *D*, combining main and substudy data in the analysis. We consider the optimization of studies of *GxE* interactions mediated through a latent biological process *X* for which a biomarker *Z* is available. We derive the optimal sampling fractions for measuring *Z*, stratifying jointly on (*D*,*E*,*G*), using the full likelihood. More complex problems may require semi-parametric maximum likelihood or Horvitz-Thompson estimating equations to avoid having to estimate many nuisance parameters. For binary data, the optimal design for estimating *GxE* interactions usually samples all *E+*/*G+* cases, some of the other cases, and only a few controls. When substudy costs are 16x main study costs, the optimal choice yields a 4-fold asymptotic relative cost-efficiency compared to measuring *Z* on all subjects. Similar results will be shown for continuous variables.
In the Children's Health Study, longitudinal measurements of exhaled nitric oxide (eNO)–a marker of inflammation– are available on 2144 subjects. Air pollution and maternal smoking were found to interact with genes in the nitrosative stress pathway (e.g., nitric oxide synthase, *NOS2A*) for eNO and asthma. By repeated subsampling 3/8 the real data in various ways, we found that the optimal design would have increased SEs by only 33% relative to no subsampling, while providing a considerable savings in cost.

95

Impact of Population Stratification on Gene-Environment Interaction Analysis

Elena Viktorova (1) Melanie Sohns (1) Heike Bickeboller (1)
(1) University Medical Center, Georg-August-University Gottingen, Germany

Methods for Gene-Environment (GxE) interactions in genome-wide association scans (GWAS) should help to explain part of the heritability in complex diseases. Compared to main effects requirements regarding power, sample sizes, number of estimated parameters are much higher. Confounding factors, particularly, population stratification (PS) and gene-environment association across the genome may complicate the situation even further. Recently bias of GxE interactions due to PS has been studied for case-control and case-only designs (e.g. Wang et al. 2008).
We will investigate the possible bias due to population stratification for the interaction term in six popular methods for GxE interaction in GWAS (approaches: case-control, case-only (Piegorsch et al. 1994), general two step (Albert et al. 2001) and Murcray's two-step (Murcray et al. 2009), Empirical Bayes (Mukhrerjee et al. 2008), hierarchical Bayes (Lewinger et al. 2007).
We simulate data under a range of realistic scenarios, similar to the procedures described by Wang et al. (2006, 2008). We consider two or more ethnicities and a binary outcome and exposure. We will implement specified disease risks, genotype frequencies and exposure prevalence. Analysis will be carried out with and without adjustment for population stratification by principal component analysis in order to consider the bias with and without this adjustment for the respective methods.

96

Integrating Multiple Genetic and Environmental Factors Using Structural Equation Modeling: An Application to Obesity, Adipokine and Cytokine Signaling Pathways and Prostate Cancer Risk

Nora L Nock (1) Albert Levine (2) Christine Neslund-Dudas (2) Jennifer Beebe-Dimmer (3) Cathryn Bock (3) Andrew Rundle (4) Deliang Tang (4) Robert Elston (1) Benjamin A Rybicki (2)
(1) Case Western Reserve University (2) Henry Health Ford System (3) Karmanos Cancer Institute (4) Columbia University

Statistical approaches that simultaneously model multiple genetic and environmental factors in a hierarchical manner that reflects the underlying pathophysiology are needed to better understand the role that these factors play in a complex disease. We present an approach that integrates multiple data sources and models genes as latent constructs, defined by multiple SNPs within each gene, using the multivariate statistical framework of structural equation modeling (SEM). Prediagnostic circulating levels of adipokines and cytokines have been associated with aggressive prostate cancer risk, yet few studies have examined the effects of SNPs in these genes on prostate cancer and, no studies have simultaneously modeled multiple genes in these pathways. Thus, we applied our SEM approach to modeling multiple genes in adipokine and cytokine signaling on prostate cancer using candidate gene and genomewide SNP arrays. We found that LEPR, ADIPOQ, ADIPOR1 and IL-6 genes were associated with prostate cancer risk in Caucasians (rs1887285, rs822391, rs1342387, rs2069835) and African-Americans (rs6700896, rs1342387, rs1548216). Obesity-associated increased risk of aggressive prostate cancer was markedly reduced among African-Americans who had the IL-6 rs154821 G/C or C/C genotype. By comparing our multivariate approach to "one-SNP-at-a-time" regression methods, we show that the SEM approach improves control of confounding and more efficiently captures overall gene effects.

97

Explaining The Tails Of Quantitative Distributions

Cornelia Marja van Duijn (1) Maksim Struchalin (1) Najaf Amin (1) Kelly Benke (2) Lennart Karssen (1)
(1) ErasmusMC (2) Ontario Institute for Cancer Research

The past decade has seen major progress in the discovery of genes involved in common complex disease. Hundreds of genes have also been identified for quantitative risk factors. The variants identified so far by themselves explain only a small proportion of the heritability. Yet, we have shown that adding up the relatively small effects of even a limited number of loci (N=18) for total cholesterol levels explain a similar proportion of the variation as the joint effect of major epidemiological risk factors age, sex and body mass index. In this study we compared to what extend genetic variants and epidemiological risk factors explain the presence of an individual in the extreme of the distribution of various quantitative traits. The traits studied include outcomes for which a large number of loci are known (height, lipid levels, macular degeneration) and outcomes for which our genetic knowledge is limited (e.g. blood pressure, body mass index). We used two population-based studies well characterized for genetic and environmental risk factors: the Rotterdam study (N=12,000) and the Erasmus Rucphen Family Study (ERF;N=3000). In ERF the relationship between subjects for whom the presence in the extremes could **NOT** be explained by known genetic and environmental risk factors was estimated. The findings combining genetic and epidemiologic approaches have major implications for effective screening for rare mutations with large effects in those in the extremes of distributions.

98

Genome-Wide Association Study of Melanoma Progression and Blood Biomarkers

Shenying Fang (1) Li-E Wang (1) Jeffrey E. Gershenwald (1) Wei Chen (1) Christopher W Schacherer (1) Julie M Gardner (1) Yuling Wang (1) Tim D Bishop (2) Jennifer H Barrett (2) Elizabeth A Grimm (1) Caitlin McHugh (3) Cathy Laurie (3) Kim F Doheny (4) Elizabeth W Pugh (4) Qingyi Wei (1) Christopher I Amos (1) Jeffrey E Lee (1)
(1) The University of Texas MD Anderson Cancer Center (2) University of Leeds (3) University of Washington (4) Johns Hopkins University

Melanoma is a rare, albeit potentially highly aggressive malignant tumor of melanocytes, accounting for the majority of deaths from skin cancer. Standard clinicopathological features (eg, primary tumor thickness, ulceration, sentinel lymph node status) cannot completely predict which patients will recur. For those who do recur, current therapies are effective for only a minority of patients. Therefore, further identification of the role of tumor-associated biomarkers and their genetic variation in melanoma progression is needed. In order to accomplish this, we conducted a genome-wide association analysis of melanoma prognosis using samples and data from 1804 melanoma cases from M.D. Anderson Cancer Center and 1026 age and sex-matched healthy controls.Tumor thickness was found to be associated with both disease-free survival and overall survival. Elevated IL-12p40 was discovered to predict poorer overall survival. We identified association of MUC2(rs12365253, $P=5.62\times10^{-8}$) and CALR/RAD23A(rs1049481, $P=5.83\times10^{-7}$) genes with tumor thickness, and strong association of EBF1(rs6895454, $P=7.65\times10^{-11}$) and IL12B (SNP5-158735664, $P=1.49\times10^{-19}$) genes with IL-12p40 level. Our results further demonstrated evidence associating these genes withdisease progression. These findings will help to elucidate the mechanism of melanoma progression, define high-risk groups for adjuvant therapy, and provide new clinical indicators for response to therapy over time.

99

Genome-Wide Association Studies (Gwass) In Homogeneous Subgroups Of Caucasian Samples Identify Six Novel Loci For Renal Function: The Ckdgen Consortium

Cristian Pattaro (1) on behalf of the CKDGen Consortium (2)
(1) Institute of Genetic Medicine, European Academy of Bolzano/Bozen (EURAC) (2) The CKDGen Consortium

Previous GWASs uncovered several loci related to renal function but there has been no systematic examination of SNPs that may have subgroup-specific effects.
We conducted 9 new meta-analyses on a larger set of GWASs (26 studies, 74,354 individuals of European descent) to identify SNPs related to renal function, as assessed by estimated glomerular filtration rate (eGFR), in the overall population and in subgroups by age, sex, diabetes and hypertension, the major risk factors for chronic kidney disease (CKD). We assessed replication of the most significant SNPs in 19 additional studies (N=56,246). In the combined sample (N=130,600), we tested replicated loci for between-group difference (two-sample Z-test) and association with CKD ($eGFR<60ml/min/1.73m^2$) and severe CKD ($eGFR<45$).
We identified 6 novel loci: 3 in the overall group (p from 4.3E-08 to 8.4E-18), 1 in the non-diabetic group (*SLC47A1*,

p=2.1E-09) and 2 in the young group (*CASP9*, p=1.5E-17; *CDK12*, p=9.0E-13). *CDK12* was associated with eGFR in younger but not in older subjects (Z-test p=0.0008). No sex or hypertension subgroup-specific SNPs were identified. The majority of novel and known loci was more strongly associated with severe CKD than with CKD.
Subgroup analyses identified novel renal function loci, specifically *SLC47A1* and *CDK12*, which would have otherwise been missed in the overall GWAS. Despite reduced sample size, subgroup analysis can increase statistical power by reducing between-study heterogeneity.

100
Genome-Wide Association Studies Of Functional Traits: An Application To Lipid Density Profiles In Type 1 Diabetes
Marie-Pierre Sylvestre (1) Angelo J Canty (2) Darryl Waggot (3) Andrew D Paterson (4) Andrew P Boright (5) John D Brunzell (6) Shelley B Bull (3)
(1) CHUM Research Centre (2) McMaster University (3) Samuel Lunenfeld Research Institute (4) Hospital for Sick Children (5) University Health Network (6) University of Washington

Introduction: Genetic studies of lipid traits generally use fasting serum total cholesterol, HDL and LDL. Plasma lipid measures obtained by density gradient ultracentrifugation (DGUC) of lipid particles produces a more informative characterization of an individual's profile of values on 38 ordered lipid fractions permitting specific genetic analysis of lipid fractions. This work is motivated by the analysis of lipid profiles from 1259 individuals with type 1 diabetes in the Diabetes Control and Complications Trial.
Methods: We describe two methods to conduct GWAS analysis of the lipid density profiles: 1. pointwise analysis of each of the DGUC fractions with a combined p-value for each SNP using the bootstrap to account for correlation; and 2. semi-parametric mixed model (SPMM) analysis to assess association between the SNP and the entire profile. The SPMM uses penalized splines to model the entire profile and allows plots of fitted profile by genotype. We compare these approaches with analysis of conventional lipid traits.
Results: SNPs achieving GWAS significance varied across the methods with some SNPs only identified by full analysis of the profiles. Combining fraction specific test statistics or p-values for each SNP caused inflated type-1 error unless the correlation between the fractions was accounted for. The plots of the lipid profiles generated by the SPMM models informed the specific fractions affected by a SNP, and the genetic model for the association.

101
Genome-Wide Association Study (GWAS) Of Lactose Consumption Measured Longitudinally Identifies A Novel Variant 500kb Downstream Of The LCT Gene Region
Karen M. Eny (1) Shelley B. Bull (2) Angelo J. Canty (3) Lei Sun (4) Andrew P. Boright (5) Mohsen Hosseini (1) Patricia A. Cleary (6) John M. Lachin (6) DCCT/EDIC Research Group (7) Andrew D. Paterson (1)
(1) Program in Genetics and Genome Biology, The Hospital for Sick Children (2) Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute of Mount Sinai Hospital (3) Department of Mathematics and Statistics, McMaster University (4) Dalla Lana School of Public Health, University of Toronto (5) Department of Medicine, University of Toronto (6) The Biostatistics Center, The George Washington University (7) DCCT/EDIC Research Group

Lactase non-persistence occurs due to a decrease in lactase activity, resulting in the inability to digest lactose in adulthood. rs4988235, 14kb upstream of the lactase (*LCT Chr 2q21*) gene has been associated with lactase non-persistence and lactose consumption. *In vitro* studies demonstrated differential enhancer activity of the *LCT* promoter by rs4988235, suggesting this SNP to be the causal variant in Europeans. However, given the long-range LD known to occur in this region, it is conceivable a variant other than rs4988235 could be the causal polymorphism. To date, no GWAS has been conducted on lactose consumption. We therefore conducted a GWAS of mean lactose intake assessed using a diet history questionnaire an average of 3.6 times per subject over a mean of 6.5 years in 1304 white participants with type 1 diabetes from the Diabetes Control and Complications Trial. Linear regression analysis of 841,342 SNPs with mean lactose intake transformed into a normal score identified several SNPs in the LCT region. The top SNP, rs1561277, (?±SE=–0.23?0.04, p=1.2e-8) is located in the *ZRANB3* gene, 517kb centromeric from rs4988235 (?±SE=–0.21±0.04, p=1.8e-7) with pairwise r^2=0.77. In analysis of rs1561277 with rs4988235, the association of rs4988235 was diminished (p=0.68), while the effect at rs1561277 was attenuated, but not completely diminished (?±SE=–0.20±0.09, p=0.02). Our GWAS, therefore, suggests that rs1561277 or another variant in LD with it are alternative causal variants.

102
Meta-analysis of Genome-Wide Associations Studies of Lung Cancer
Maria N Timofeeva (1) Paul Brennan (1) Maria T Landi (2) Thorunn Rafnar (3) Richard Houlston (4) Rayjean Hung (5) Christopher I Amos (6) on behalf of Area 1 Transdisciplinary Research in Cancer of the Lung Research Team (7)
(1) International Agency for Research on Cancer, France (2) Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA (3) deCODE genetics, Iceland (4) Institute of Cancer Research, UK (5) Samuel Lunenfeld Research Institute, Canada (6) University of Texas M.D. Anderson Cancer Center, USA (7) n/a

Several loci including region 15q25, 5p15 and 6p21 have been identified influencing susceptibility to lung cancer in genome-wide association studies (GWAS). Lung cancer being highly heterogeneous disease might have different etiology depending on histology, smoking history and other factors and therefore stratification by demographic, epidemiological and clinical parameters might be of particular interest.
As a part of Transdisciplinary Research in Cancer of the Lung (TRICL) Research Team nine existing GWAS have combined their data comprising over 10000 cases and 25000 controls with the aim to conduct a meta-analysis within specific subgroups; in particular smoking status, histology, sex, age of onset, stage and family history.

An analysis protocol and research environment to facilitate data exchange have been developed within the initiative. The propose strategy includes joint analysis stratified by subgroups of interest using both random and fixed effect models based on the heterogeneity test. The initial analysis is currently in the progress. In order to allow multiple users to analyze the data, we have established a memorandum of understanding process that includes shared access to data and a supportive computational environment.
This analysis will hopefully lead to new susceptibility loci that are specific for particular subgroups and will help generate additional hypotheses that can be followed up within the International Lung Cancer Consortium and TRICL.

103

To What Extent Genotype Imputations Are Able To Identify Causal Variants In Genome-Wide Association Studies?

Myriam Brossard (1) Eve Corda (1) Mark M Iles (2) Jenny H Barrett (2) Alisa M Goldstein (3) Peter Kanetsky (4) Elizabeth M Gillanders (5) Bert Bakker (6) Nelleke Gruis (7) Julia A Newton-Bishop (2) D. Timothy Bishop (2) Geno MEL (8) Florence Demenais (1)
(1) INSERM U946, Fondation Jean-Dausset-CEPH, Paris, France (2) Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, Leeds, UK (3) Genetic Epidemiology Br., DCEG, National Cancer Institute, NIH, Bethesda, MD (4) Dept of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA (5) Inherited Disease Research Br., National Human Genome Research Institute, NIH, Baltimore, MD (6) Dept. of Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands (7) Dept. of Dermatology, Leiden University Medical Centre, Leiden, The Netherlands (8) Melanoma Genetics Consortium

GWAS have been successful in uncovering many loci associated with complex diseases, but the causal variants are often unknown. Our goal was to investigate to what extent genotype imputations were able to identify causal variants. We studied *MC1R*, a gene known to be involved in melanoma, located in 16q24 region associated with this cancer in a GWAS but with no variant on the genotyping chips. We conducted imputations on 16q24 in 2,985 GenoMEL cases and 7,787 controls using IMPUTE2 and Hapmap3+1000 Genomes as templates. A total of 1,543 SNPs passed QC. Single SNP regression was followed by joint analysis of all SNPs using a penalised likelihood approach (HyperLasso). Four functional *MC1R* variants (R151C,R160W,V92M,R163Q) were imputed. Univariate analysis showed the strongest signal with R151C ($p=4\times10^{-30}$) while R160W had $p=9\times10^{-10}$ and 70 SNPs in other genes had p?10^{-8}. Joint analysis of all SNPs showed significant effects of R151C and/or R160W in all models that converged plus effects of other SNPs. We sequenced *MC1R* in 937 cases and 907 controls. Analysis of sequenced variants alone or with imputed SNPs showed significant effects for R151C ($p=4\times10^{-15}$), R160W ($p=1\times10^{-9}$) and D294H ($p=2\times10^{-4}$), a rare variant absent from the imputation panels. Thus, imputations can be of great use in pinpointing a gene that has a functional role in disease but may not be identified by genotyping. A full picture of the effect of causal variants on disease is revealed by resequencing.

104

Methods For Meta-Analyses Of Genome-Wide Association Studies: Critical Assessment Of Empirical Evidence

Cosetta Minelli (1) Martin Gogele (1) Ammarin Thakkinstian (2) Alex Yurkiewich (3) Cristian Pattaro (1) Peter P Pramstaller (1) Julian Little (3) John Attia (4) John R Thompson (5)
(1) EURAC research (2) Mahidol University (3) University of Ottawa (4) University of Newcastle (5) University of Leicester

A large number of genome-wide association (GWA) studies have been recently performed in search of genetic variants associated with common diseases, and this has been followed by a steep increase in meta-analyses of such studies aimed at detecting increasingly smaller genetic effects. The pressure to discover and publish new genetic associations has limited the time available for careful consideration of all methodological aspects of GWA meta-analysis. Here we survey the literature and provide empirical evidence on the methods used in 86 published GWA meta-analyses, including their organization, requirements about uniformity of methods used in primary studies, methods for data pooling, investigation of heterogeneity in study results and quality of reporting. We also review the underlying assumptions and implications of the different methods used. Our findings highlight how important aspects have received insufficient attention, potentially leading to missed opportunities for improving gene discovery and characterization. These include inadequate evaluation of power to replicate, with no association between number of variants sent to replication and replication sample size, and low proportion of GWA meta-analyses investigating presence and magnitude of between-study heterogeneity. Our findings also show the great variety of methods that researchers use and in doing so draw attention to areas that require further methodological research and to the need for consensus guidelines.

105

To Stratify Or Not To Stratify: What Can Be Learned From Power Considerations And A Practical Genome-Wide Search On Sex-Difference In The GIANT Consortium

I M Heid (1) T W Winkler (1) J C Randall (2) G Behrens (1) Z Kutalik (3) S I Berndt (4) A U Jackson (5) T O Kilpelainen (6) K L Monda (7) L Qi (8) T Workalemahu (8) J Czajkowski (9) F Day (6) T Esko (10) M F Feitosa (9) R Magi (2) I Mathieson (2) V Steinthorsdottir (11) G Thorleifsson (11) I B Borecki (9) R JF Loos (6) K E North (7) C M Lindgren (2)
(1) Regensburg University Medical Center, Department of Epidemiology and Preventive Medicine, Regensburg, Germany (2) Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom (3) Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland (4) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA (5) Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan

48109, USA (6) Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK (7) Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill,North Carolina,USA (8) Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA (9) Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri 63108, USA (10) Estonian Genome Centre, University of Tartu, Tartu, Estonia (11) decode Genetics, Reykjavik, Iceland

Genome-wide association (GWA) studies have previously reported sexually dimorphic associations on autosomes, but systematic searches for sex difference were lacking.
Our theoretical power considerations showed that stratified analyses have less power than overall analyses as long as the effect is pronounced in both strata at least to some extent and into consistent effect direction (CED), but are better powered to pinpoint signals with zero or small effects in one stratum and clearly for opposite effect direction (OED).
We conducted sex-stratified GWA analyses in the GIANT consortium (height, weight, waist and hip circumference, body-mass-index). We selected 619 independent SNPs from 58 GWA studies (60586 men, 73137 women) at a false discovery rate (FDR) of 5% based on sex-specific P. We did not find any loci based on the P for sex difference (Pdiff) tailored to detect OED signals. Follow-up (63 studies, 62399 men, 74660 women, including Metabochip Illumina) of the 619 SNPs identified 205 loci (joint sex-specific P $<5\times10$-8). Of these, 4 showed significant (Pdiff$<0.05/204$) and 14 suggestive (Pdiff$<=0.05$) evidence for sex-difference in the follow-up including known (e.g. GRB14, :LYPLAL1) and novel (MAP3K1, sex-difference Pdiff=1.1×10-4, HSD17B4, Pdiff=$3\times3\times10$-3, PPARG, Pdiff=8.4×10-3) sexually dimorphic loci for waist phenotypes.
Our theoretical and practical investigations underscore the gain from sex-stratified GWA analyses.

106
Selection Of Top Snps For Genome-Wide Association Study Using P Values And Magnitude Of Odds Ratios
Jian Wang (1) Sanjay Shete (1)
(1) UT MD Anderson Cancer Center

Genome-wide association (GWA) study is a powerful approach for detecting genetic variants for common diseases. Generally, in GWA studies, the most significant SNPs associated with top-ranked p values are selected in stage one, with follow-up in stage two. However, when minor allele frequencies are relatively low, less-significant p values can still correspond to higher odds ratios (ORs), which might be more useful for prediction of disease status. Therefore, if SNPs are selected using an approach based only on significant p values, some important genetic variants might be missed. We proposed an approach for selecting SNPs from stage one of GWA study, based on both p values and ORs, and conducted a simulation study to demonstrate the performance of our approach. The simulation results showed that our selecting approach was more powerful than the existing ranked p value approach for identifying relatively less-common SNPs. Therefore, in GWA studies, SNPs should be considered for inclusion based not only on ranked p values but also on ranked ORs.

107
A Non-Parametric Regression Approach To The Analysis Of Genomewide Association Studies
Pianpool Kirdwichai (1) M. Fazil Baksh (1)
(1) University of Reading

We present a novel application of non-parametric regression to identify candidate regions with moderate signals of disease-gene association using high-dimensional genomewide data. Despite substantial research the existing analysis methodologies are still largely inadequate. The recent interest in regression inspired methods have been motivated by the need to avoid the stringent thresholds imposed by multiple testing methods which typically results in only genomic regions with very strong signals of disease-gene association being taken forward as potential locations of disease genes and moderate associations being missed. However, in presence of a true association there is high probability the association signal will in fact be moderate and extremely strong signals will be far less common. Here we present an alternative model-free approach to establishing a significance threshold that inherently accounts for the correlation structure in the data. We evaluate the proposed approach using both theory and simulations. Issues such as efficiency, weight, cut-off and choice of bandwidth are elucidated. To avoid the optimal bandwidth selection problem we develop and evaluate an approach based on LD and chromosomal location. We also develop appropriate cut-offs for distinguishing true signals from noise based on the asymptotic distribution of the fitted values for different kernels. Finally, the method developed is illustrated using data from the WTCCC study (2007) of Crohn's disease.

108
Sample Selection Study Designs To Follow-Up GWAS Signals With Targeted Sequencing
Brooke L Fridley (1) Ryan Abo (1) Abra Brisbin (1) Gregory D Jenkins (1)
(1) Mayo Clinic

The association signals identified from the first wave of genome-wide association studies (GWAS) for complex phenotypes has resulted in follow-up studies using targeted sequencing. However, there is limited research on optimal study design for targeted sequencing studies. Therefore, we completed an extensive simulation study to assess various study designs that utilize existing phenotypic and/or genotypic data. The objective of our study was to determine the best approach to select the most informative samples for targeted sequencing with the goal of detecting all variants in the region. The study design scenarios varied with regard to: phenotypic information; genotypic information at the GWAS "signal"; haplotypic information around the GWAS "signal"; genetic diversity based on principal component analysis. A total of 1080 simulation scenarios were assessed. Each simulation scenario was repeated 1,000 times with the% of variants detected summarized for each scenario. In general, as expected, the biggest factor that impacted the% of detected variants was the sample size. For many scenarios, sampling from cases only performed well, while sampling based on haplotypes performed poorly. However, for a majority of scenarios we observed little difference in the% of variants detected between the various approaches. Further research is on-going to

investigate impact of the various study designs on association testing.

109

Adjusting Rare Variant Association Tests for Population Stratification Using the Stratification Score

Glen A Satten (1) Bruce Ling (2) Michael P Epstein (2) Andrew S Allen (3)
(1) Centers for Disease Control and Prevention (2) Emory University (3) Duke University

Association studies that collect sequence-level data can be used to compare the proportion of case and control participants that have rare variants in a region of the genome (or exome). As with SNP asssociation studies, it is important that findings reflect true association, not confounding due to population stratification. However, many such tests like the *C(?)* test [Neale et al. PLoS Genet e1001322] are not regression-based, so principal-components-based adjustment for population stratification is problematic. We show how the stratification score [Epstein, Allen & Satten ASHG 2007 80:921-930] can be used to form strata such that stratified versions of any test statistic account for confounding. Alternatively, we reformulate tests like the *C(?)* test as a comparison between the proportion of case and control participants having a rare variant at each locus, then standardize these proportions over the strata using the stratification score after Allen and Satten [Amer J Epid 2001 173:752-60]. This gives the test statistic that we would have obtained if cases and controls each had the same population structure. Significance can be assessed either asymptotically or by within-stratum permutation, which preserves population substructure but removes disease-variant association. We study test performance using coalescent-based simulated data that mimics the African-American population.

110

Testing and Genetic Model Selection in Genome-Wide Association Studies

Christina Loley (1) Inke R Konig (1) Ludwig Hothorn (2) Andreas Ziegler (1)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, Universitatsklinikum Schleswig-Holstein, Campus Lubeck (2) Institut fur Biostatistik, Leibniz Universitat Hannover, Hannover, Germany

In genome-wide association (GWA) studies, trend test statistics based on the additive genetic model are the standard approach to test for association. But this decision can substantially reduce power if the true genetic model is a recessive or dominant one. MAX tests have been proposed to simultaneously test for additive, recessive, and dominant genetic models and to select the most likely one out of these three. P-values are estimated using permutation, although these are computationally demanding and therefore not feasible in GWA studies. To circumvent this drawback, the analytical asymptotic distribution of the MAX test statistic needs to be derived. In this contribution, we show that the asymptotic distribution of the MAX test can be approximated by an asymptotic multivariate normal distribution. Our approach is based on a generalized linear regression model with two dummy variables, and the parameters are transformed by the delta method to model specific inheritance patterns. The approach naturally allows adjustments for environmental factors or population stratification. In a simulation study, we demonstrate the validity of the method and compare its performance to existing tests. We illustrate its application by re-analyzing GWA data on Crohn's disease.

111

FaST Linear Mixed Models for Genome-Wide Association Studies

Christoph Lippert (1) Jennifer Listgarten (2) Ying Liu (2) Carl M. Kadie (2) Robert I. Davidson (2) David Heckerman (2)
(1) Microsoft Research and Max Planck Institutes Tuebingen (2) Microsoft Research

Linear mixed models (LMMs) are among the richest class of models used today for genome-wide association studies as they are capable of correcting for population structure, family structure, and cryptic relatedness. Although their popularity is rapidly increasing for this reason, their use on contemporary data sets is limited because the required computations are prohibitive when many individuals are analyzed, with runtime increasing as the cube of the number of individuals and memory footprint increasing as the square of the number of individuals. We introduce a mathematical reformulation of LMMs called *FaST-LMM* to overcome this barrier, wherein results remain exact, but both runtime and memory footprint become linear in the number of individuals. On data from the Wellcome Trust with 15,000 individuals[1], our approach is an order of magnitude faster than the state-of-the art LMM implementation[2]. On a data set containing 120,000 individuals, the state-of-the-art implementation is unable to run, whereas *FaST-LMM* completes in just a few hours. Source and executable code are available online.

[1]Burton, P. R. et al. Genome-wide association study of 14,000 cases of seven common diseasesand 3,000 shared controls. *Nature* 447, 661-678 (2007).
[2] Kang, H. M. et al. Variance component model to account for sample-structure in genome-wide association studies. *Nature Genetics* 42, 348-354 (2010).

112

A Bayesian Network Approach For Pathway Analysis Using The Gene Ontology Database

Ronja Foraita (1) Janine Karl (1) Annika Leseberg (1) Frauke Gunther (1)
(1) Bremen Institute for Prevention Research and Social Medicine

Pathway analyses incorporate prior biological knowledge and focus on the association of disease with combined information of genetic variants in the same pathway. We propose Bayesian networks (BNs) to discover pathways enriched with disease-associated SNPs. BNs rely on the concept of conditional independence and are able to detect direct and indirect influences, e.g. pathways that are mediated through other pathways, and interactions.
In our framework, we map SNPs to genes within a pathway given the external knowledge from the gene ontology database. In order to reduce the large dimensionality of a

pathway, we apply principal component analysis to build up a pathway score by using the first principal component for each pathway. This score is used as surrogate for the genetic variability within the pathway and included in the BN analysis. BNs are learned by the tabu search algorithm with BIC as selection criteria.
The proposed method is investigated in a simulation study using data with realistic LD structure, different effect sizes for "causal" SNPs and different pathway-disease dependency structures. First results for scenario with just direct disease-pathway associations show that only 0.6% false disease-pathway associations are found. True dependencies are detected in about 50%, where small pathways have much higher detection rates than large pathways.
However, the approach is computational efficient and helps to derive parsimonious models that can be visualized in a graph.

113
Bayesian Hierarchical Modelling of SNPs and Pathways for Identifying Associated Pathways
Marina Evangelou (1) Frank Dudbridge (2) Lorenz Wernisch (1)
(1) MRC-Biostatistics Unit, University of Cambridge,UK (2) London School of Hygiene and Tropical Medicine, UK

Pathway analysis incorporates the biological knowledge of SNPs and genes for revealing the underlying genetic structure of a phenotype. Several methods have been proposed for identifying pathways. Some of these methods combine the pathway SNP p-values into a single pathway p-value, e.g. Fisher's Method (FM) and Tail Strength Measure (TSM). Other pathway analysis methods look for enrichment of pathways within the list of top ranking SNPs, e.g. Fisher's Exact Test (FET).
We propose a Bayesian Hierarchical framework that includes the pathway membership of SNPs for modelling both pathway level effects and SNP effects within pathways. The phenotype of each individual is assumed to depend both on its genotype data and on the sum of its alleles within each pathway. The goal of our framework is to identify pathways associated with the phenotype; SNP parameters are considered as nuisance parameters and integrated out of the analysis. A sparse and a standard normal distribution are considered and tested for the pathway parameters.
We carried out a simulation study for testing the performance of the proposed Bayesian Hierarchical framework and comparing it with other methods like FM, TSM and FET. The results of the study show that our proposed framework outperforms the other tested methods. Our framework does not rely on the results of single-SNP analysis, unlike the other methods, and uses the advantage of hierarchical modelling to decrease the signal-to-noise ratio.

114
Pathway Analysis Of The Genomel Consortium Genome-Wide Association Study Of Melanoma: Analysis Of Genes Related To Tumour Immunosuppression
Nils Schoof (1) Mark M Iles (2) Tim Bishop (2) Julia A Newton Bishop (2) Jenny Barrett (2)
(1) Department of Medical Epidemiology and Biostatistics, Karolinska Institute (2) Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, University of Leeds

Systemic and sunburn-induced immunosuppression are risk factors for melanoma, making genes in immunosuppression pathways candidates for melanoma susceptibility. If variants within these genes individually have weak effects on risk, the association may be undetected in genome-wide association (GWA) studies. Pathway-based approaches have been suggested as a way of including *a priori* knowledge into the analysis. Association of 1113 single nucleotide polymorphisms (SNPs) in 43 genes related to immunosuppression was analysed using a gene-set approach in 1539 melanoma cases and 3917 controls from the GenoMEL consortium GWA study. Association was tested between melanoma and the whole set of genes, as well as predefined functional subgroups, using a measure formed by summing the evidence from the most significant SNP in each gene. Significance was evaluated by case-control permutation. Based on 5000 simulations, an association was found between melanoma and the whole set of genes ($p=0.002$), as well as subgroups related to secretion of suppressive factors ($p=0.0004$) and the generation of tolerogenic dendritic cells ($p=0.006$), thus providing preliminary evidence of involvement of immunosuppression gene polymorphisms in melanoma susceptibility. The analysis was repeated on a second phase of the GenoMEL study but showed no evidence of association. As one of the first attempts to replicate a pathway-level association this suggests that low power and heterogeneity present challenges.

115
Comparison of Methods for Enrichment Tests in Pathway Analysis
Marina Evangelou (1) Lorenz Wernisch (1) Frank Dudbridge (2)
(1) MRC-Biostatistics Unit, University of Cambridge, UK
(2) London School of Hygiene and Tropical Medicine, UK

The available biological knowledge of SNPs and genes is incorporated through pathway analysis for revealing the underlying genetic structure of a studied phenotype. Several methods have been published for pathway analysis, which can be classified into enrichment and association methods according to the null hypothesis being tested. Although association tests are statistically more powerful than enrichment tests, they can be difficult to calibrate because biases in analysis accumulate across multiple SNPs or genes. Furthermore, enrichment tests can be more biologically relevant than association tests, as they detect pathways with relatively more evidence for association than the remaining genes. Here we show how some well known association tests can be simply adapted to test for enrichment, and compare their performance to some established enrichment tests. We propose versions of the Adaptive Rank Truncated Product (ARTP), Tail Strength and Fisher's combination of p-values for testing the enrichment null hypothesis. We compare the behaviour of these proposed methods with the established Gene-Set Enrichment Analysis and the Fisher's Exact Test. The results of our simulation study show that the adapted version of the ARTP method has the greatest power for all tested cases. This is in agreement with

previous work on association tests and suggests that the ARTP should be preferred for both enrichment and association testing.

116

Joint Modeling Of Multiple Biomarkers Predicting Longevity In Families Using Weighted Penalized Logistic Ridge (WPLR) Regression

Hae-Won Uh (1) Bart J A Mertens (1) Fabrice P R Colas (1) Marian Beekman (2) Jeanine J Houwing-Duistermaat (1)
(1) Dept. Medical Statistics and Bioinformatics, LUMC (2) Section of Molecular Epidemiology, LUMC

Several methods are available for detecting biomarkers form high-dimensional data observed in independent samples, such as ridge regression and multifactor dimensionality reduction approach. These methods, however, are not directly applicable to family data.
To search for biomarkers associated with longevity, we propose a weighted penalized logistic ridge (WPLR) regression. To deal with the correlation between the biomarkers, we impose a ridge-type penalty. The weights used in WPLR are to adjust for the family size. In particular, the family-specific contribution to the likelihood function is down-weighed by a factor inversely proportional to the family size. WPLR implements a double cross-validation procedure each on leave-one-family-out basis to keep family structure intact. The inner level cross-validation is to select the optimal penalty parameter and the second level is to optimize the calibration of ridge regression estimator.
We present application to data from the Leiden Longevity Study. Molecular and phenotypic parameters (N=60) were measured in offspring of nonagenarian siblings from 420 families (cases, N=1671) and in the partners of the offspring (controls, N=744). The size of family members ranges from 1 to 10. The results using WPLR show that multiple parameters might be potential biomarkers. Subsequently, we investigate which families/individuals contribute most to drive the classification using calibrated predicted probabilities.

117

Of Preliminary Test And Shrinkage Estimators For Evaluating Multiple Exposures

Jaya M Satagopan (1) Qin Zhou (1) Susan A Oliveria (1) Stephen W Dusza (1) Martin A Weinstock (2) Marianne Berwick (3) Allan C Halpern (1)
(1) Memorial Sloan-Kettering Cancer Center (2) Brown University (3) University of New Mexico

Epidemiology studies examine multiple exposures (genetic or environmental) in relation to disease. In addition to the empirically ascertained exposure data, external information about certain characteristics of the exposures is also becoming increasingly available either by design or through prior related work. For example, a questionnaire for evaluating sun-related data may be designed to elicit information about themes such as sun exposure and sunburn. Gene functions or gene expressions are also becoming available through public data bases or from related studies. It is now anticipated that incorporating such external data may provide better insights into the disease relevance of the individual exposures. However, this approach may lead to inflated bias and type I errors when the manner in which the external data are incorporated is mis-specified. We investigate preliminary test estimators and shrinkage estimators as alternative approaches to address this issue. We show that these estimators are intimately related under certain assumptions and that the shrinkage estimator derived under the assumption of an exchangeable prior distribution provides precise estimates of the exposure effects and is robust to mis-specification issues. The benefits and limitations of these estimators are illustrated using questionnaire data on sun-related factors from the Study of Nevi in Children, where the exposures are the individual questions and the outcome is (log) total back nevus count.

118

Comparisons Of Shrinkage Estimation Methods For Improved Prediction Of Quantitative Phenotypic Traits Related To Individual CVD Risk

Helen R Warren (1) Juan-Pablo Casas (2) Clive Hoggart (3) Frank Dudbridge (1) John Whittaker (1)
(1) London School of Hyigene & Tropical Medicine (2) London School of Hyigene & Tropical Medicine and University College London (3) Imperial College, London

There has been great interest in the potential of genetics to improve individual risk prediction, but most work has concentrated on the relatively small number of confirmed genetic determinants, which explain only a small proportion of the overall genetic variance. We explore ways to use much larger sets of SNPs, allowing for both common variants of small effect and rare variants of large effect, so that a larger amount of genetic variation is exploited, increasing the potential utility of genetic based prediction of phenotypic traits related to disease risk.
Current shrinkage methods from genome-wide association studies can be adapted for risk prediction models. From a Bayesian perspective, alternative shrinkage approaches relate to different priors on the effect size. We consider the Lasso and Hyper-Lasso methods, using Double Exponential and Normal-Exponential-Gamma priors respectively, and compare these to established approaches, such as Ridge Regression, which corresponds to using a Gaussian prior. Shrinkage methods are expected to yield greater accuracy than single SNP analysis, since effect sizes in large-scale genetic studies tend to follow exponential family distributions.
These models are applied to the genetic prediction of lipid levels, which are established risk factors for cardiovascular disease. Our methods are tested on real data-sets, including the Whitehall II cohort study, which contains data for almost 50,000 SNPs.

119

Regularized Heritability Estimation In Multivariate Traits

Stefan Boehringer (1)
(1) Leiden University Medical Center

In multivariate phenotypes, principal component of heritability (PCH) is defined as the linear combination maximizing heritability. A naive estimator standardizes the intra-familial covariance matrix to be spherical and extracts the PCH from the inter-familial covariance matrix (Ott &

Rabinowitz 1999, Hum Hered, 106-11). The standardization introduces high variance into the estimator and regularization techniques can be used to improve on cross-validated heritability. We use two data-adaptive bootstrap procedures to estimate biases in the spectral decompositions of the involved matrices that can be used to correct the estimator and compare with other proposed estimators. We apply the methods to a facial data set (110 sib pairs, 96 variables) and integrate standard epidemiological measurements (2 variables). We show that the best cross-validated heritability for the first PCH using our methods (13.6%) outperforms the naiv estimator (6.2%). Visual inspection of the estimators reveal differing PCHs for estimators that perform similar in terms of heritability. In conclusion, for complex phenotypes such as the face several directions with similar heritability seem to exist. These methods can be used to integrate many data sets by analyzing their joint heritability. Ensuing association or linkage studies should have increased power when using the PCHs as phenotypes.

120

The Linkage Disequilibrium LASSO for SNP Selection in Genetic Association Studies

Samuel G Younkin (1) J. Sunil Rao (2)
(1) Case Western Reserve University (2) University of Miami

In recent years the field of disease-gene mapping has been dominated by the genome-wide association study in which the entire genome is interrogated for single nucleotide polymorphisms (SNPs) responsible for disease. This method has resulted in the identification of only a small portion of the expected disease susceptibility sites, and it is our belief that by integrating data on haplotype block structure and association signal shape to disease-gene mapping we may overcome some of the inherent difficulties that arise from the vast amount of multiple testing associated with the agnostic approach. Here we develop a statistical method designed to detect a genetic association signal that is more representative of the remaining disease susceptibility SNPs. The genetic association signal that we seek is no longer a peak, for the increase in SNP density, coupled with low effect sizes, gives rise to a plateau-like signal with gaps. We address this by formulating our method as a penalized least squares regression estimator based on the linkage disequilibrium present between SNPs in the same haplotype block. The method known as the LD LASSO is an adaptation of the fused LASSO used for subset selection in a regression framework in which the signal is sparse and block-like. We implement this method in the R package *ldlasso*. We demonstrate the use of the method by examining six regions on chromosome 8 suspected to contain variants associated with Late Onset Alzheimer's Disease.

121

Integrating Molecular (Mrna And Mirna) And Immunohistochemical (IHC) Data To Identify Subgroups Of Estrogen Receptor (ER) Positive Breast Cancer Patients

Dushanthi Pinnaduwage (1) Dylan Ehman (1) Irene L Andrulis (1)
(1) Samuel Lunenfeld Research Institute

Profiling patterns of messenger RNA (mRNA) expression reveals several distinct breast cancer subtypes. Of the two subtypes of ER positive tumors (luminal), Luminal B tumors have higher proliferation and poorer prognosis than luminal A. mRNA expression profiling can discriminate between the subtypes due to increased expression of proliferation genes in luminal B, however the classification system is not yet robust. Many cell processes are controlled by microRNA (miRNA), which has an important role in differentiation and proliferation as it regulates both tumour suppressors and oncogenes. We hypothesize that miRNA expression profiling will discriminate between A and B subtypes with superior accuracy to mRNA expression profiling. We have mRNA array and immunohistochemistry data for subgroups of 137 and 888 tumors, respectively, from our large Axillary Node Negative breast cancer cohort. After collecting miRNA data for a subset of these tumors, we integrate IHC, mRNA, and miRNA data to classify luminal tumors. We analyze the data using unsupervised and supervised methods to explore subgroups of luminals and to identify differentially expressed genes (DEG) and miRNAs (DEmiRNA). Luminal groups defined by IHC data will be used in supervised methods. We perform an enrichment analysis to identify pathways associated with DEG and DEmiRNA. Further, we explore similarities between the pathways associated with DEG and miRNA to find possible interactions between miRNA and mRNA.

122

Searching Through The Folate Metabolism Pathway For Genetic And Nutritional Risk Factors For Lung Cancer

Michael D. Swartz (1) Christine B. Peterson (2) Philip J Lupo, Jr. (1) Ladia H. Hernandez (3) Marina Vannucci (2) Sanjay Shete (3)
(1) University of Texas School of Public Health (2) Rice University (3) University of Texas M. D. Anderson Cancer Center

Lung cancer continues to be one of the deadliest cancers, owing its etiology to genetic and nutritional factors. This project investigates the epidemiology of lung cancer risk focusing on genes from the folate metabolism pathway as well as nutrients important to folate metabolism. We analyzed 1239 lung cancer cases and 1692 controls collected as part of an ongoing lung cancer study from the University of Texas M. D. Anderson Cancer Center. Genotypes were obtained from blood samples, and nutrition information was obtained from food frequency questionnaires. We are one of the first groups to use stochastic search variable selection to search through Single Nucleotide Polymorphisms marking genes in the folate metabolism pathway while jointly searching through nutritional factors related to folate metabolism (stratified by smoking status). We used prior probabilities of inclusion that control the false positive rate, and incorporate linkage disequilibrium to improve the search through potential genes. For current smokers, preliminary results indicate *MTRR*, *SHMT1*, *CBS*, alcohol intake and vitamin B6 intake as factors affecting lung cancer risk. For former smokers, we identify *MTHFR* and alcohol intake as potential factors influencing lung cancer risk. For never smokers we identify *MTHFR* and *MTRR*, and no nutritional variables.

123

Polymorphisms In Oxidative Stress-Related Genes, Radiotherapy, And Overall Survival In Breast Cancer Patients - A Replication Study

Petra Seibold (1) Per Hall (2) Nils Schoof (2) Heli Nevanlinna (3) Tuomas Heikkinen (3) Jianjun Liu (4) Dario Greco (3) Peter Schmezer (1) Odilia Popanda (1) Dieter Flesch-Janys (5) Jenny Chang-Claude (1)
(1) German Cancer Research Center (2) Karolinska Institute (3) Helsinki University Central Hospital (4) Genome Institute of Singapore (5) University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf

As radiotherapy (RT) leads to increased formation of reactive species, we hypothesized that SNPs in oxidative stress-related candidate genes might modify efficacy of treatment and therefore overall survival (OS) of breast cancer patients.
The breast cancer patients were aged 50-74 at diagnosis 2001-2005 and recruited in a German population-based case-control study (MARIE). Follow-up of vital status by end of 2009 was carried out through population registries. 109 SNPs in 22 candidate genes were genotyped in 1,639 patients. After excluding patients with *in situ* or previous tumour(s), 1,410 were included for analysis (199 events). Cox proportional hazards models were used to assess log-additive effect of individual SNPs with OS as well as interaction with RT, adjusted for prognostic factors. We attempted validation of our results in two independent studies from Scandinavia with available genome-wide SNP data: HEBCS (Helsinki Breast Cancer Study, n=702 / 272 events) and SASBAC (Singapore and Swedish Breast Cancer Study, n=795 / 195 events).
Eight of the SNPs in *MT2A*, *NQO1*, *PRDX1, PRDX6*, *TXN*, which were significantly associated with OS in MARIE ($p<0.05$), were genotyped in the Scandinavian studies but did not replicate. Of the SNPs showing significant interaction with RT in MARIE, one SNP in *TXN* replicated in SASBAC but not in HEBCS.
We did not find clear evidence for association between oxidative stress-related genetic variants and breast cancer survival after replication.

124

Phospholipase A2G1B Polymorphisms and Risk of Colorectal Neoplasia

Clare Abbenhardt (1) Elizabeth E Poole (2) Liren Xiao (3) Karen Curtin (4) Rachel L Galbraith (5) David Duggan (6) Li Hsu (5) Karen W Makar (5) Richard J Kulmacz (7) John D Potter (5) Betty J Caan (8) Lisel Koepl (5) Anna E Coghill (5) John Muehling (6) David Taverna (6) Christopher S Carlson (5) Marty L Slattery (4) Cornelia M Ulrich (9)
(1) National Center for Tumor Diseases/German Cancer Research Center, Heidelberg, Germany (2) Fred Hutchinson Cancer Research Center, Seattle, WA, USA,Channing Laboratory, Department of Medicine, Brigham and Women's Hospit (3) Fred HUtchinson Cancer Research Center, Seattle, WA, USA (4) University of Utah, School of Medicine, Department of Medicine, Salt Lake City, UT, USA (5) Fred Hutchinson Cancer Research Center, Seattle, WA, USA (6) Translational Genomics Research Institute, Phoenix, AZ, USA (7) University of Texas Health Science Center at Houston, Houston, TX, USA (8) Kaiser Permanente Medical Research Program, Department of Research, (9) National Center for Tumor Diseases/German Cancer Research Center, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Phospholipase A2 (*PLA2G1B*) catalyzes the release of fatty acids from dietary phospholipids. Some polyunsaturated fatty acids removed from the intestinal lumen are precursors to eicosanoids, which are linked to inflammation and colorectal carcinogenesis. We evaluated the association of *PLA2G1B* tagSNPs with colorectal neoplasia risk in 3 study populations.
A linkage-disequilibrium (LD)-based tagSNP selection algorithm (r^2=0.90, MAF=4%) identified 3 tagSNPs in *PLA2G1B.* We genotyped the SNPs on the identical Illumina platform in 3 case-control studies of colon cancer (1424 cases/1780 controls), rectal cancer (583 cases/775 controls), colorectal adenomas (485 cases/578 controls). For gene-level associations, we conducted principal components (PCA) and haplotype analysis. Multiple logistic regression was used for single SNPs, adjusting for age, sex, study site.
Two *PLA2G1B* variants were statistically significantly associated with reduced risk of rectal cancer (rs5637, 3702G>A Ser98Ser, p-trend=0.03 and rs9657930, 1593C>T p-trend=0.01) and PCA was significant (p=0.02). For a third we observed statistically significant interactions with NSAID use. LD between the SNPs was modest ($r^2<0.6$). Associations with colon tumor mutation subtypes (*TP53*-positive, *KRAS2*-positive) were observed.
The results suggest that genetic variability in *PLA2G1B* may affect susceptibility to rectal cancer and may be involved in colon tumors characterized by *TP53* and *KRAS2* mutations.

125

Glutathione Peroxidase (GPX) Tagsnps: Associations With Rectal Cancer But Not With Colon Cancer

Ulrike Haug (1) Elizabeth M Poole (2) Liren Xiao (3) Karen Curtin (4) David Duggan (5) Li Hsu (6) Karen W Makar (6) Ulrike Peters (6) Richard J Kulmacz (7) John D Potter (6) Lisl Koepl (6) Bette J Caan (6) Marty L Slattery (4) Cornelia M Ulrich (1)
(1) National Center for Tumor Diseases / German Cancer Research Center (DKFZ), Heidelberg, Germany (2) Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA (3) Fred HUtchinson Cancer Research Center, Seattle, WA, USA (4) University of Utah, School of Medicine, Department of Medicine, Salt Lake City, UT, USA (5) Translational Genomics Research Institute,Phoenix,AZ,USA(6) Fred Hutchinson Cancer Research Center, Seattle,WA,USA (7) University of Texas Health Science Center at Houston, Houston, TX, USA

Purpose: We investigated tagSNPs in *GPX1-4* in relation to colorectal neoplasia in three independent study populations.
Methods: A linkage-disequilibrium (LD)-based tagSNP selection algorithm (r^2?0.90, MAF?4%) identified 21 tagSNPs. We used an identical Illumina platform to genotype *GPX* SNPs in three population-based case-control studies of colon cancer (1424 cases/1780 controls), rectal cancer (583 cases/775 controls), and colorectal adenomas (485 cases/578 controls). For gene-level associations, we conducted principal components analysis (PCA); multiple logistic regression was used for single SNPs. Analyses were

adjusted for age, sex, and study center and restricted to Caucasians. Analyses of cancer endpoints were stratified by molecular subtypes.
Results: Without correction for multiple testing, one polymorphism in *GPX2* and three polymorphisms in *GPX3* were associated with a significant risk reduction for rectal cancer at ?=0.05, specifically for rectal cancers with *TP53* mutations. The associations regarding the three polymorphisms in *GPX3* remained statistically significant after adjustment for multiple comparisons. The PCA confirmed an overall association of *GPX3* with rectal cancer (p=0.03). No other statistically significant associations were observed.
Conclusion: These data provide preliminary evidence that genetic variability in *GPX3* contributes to risk of rectal cancer but not of colon cancer.

126
PTGS1 And PTGS2 Polymorphisms, Fatty Acid Intake, And Risk Of Colon And Rectal Cancer
Nina Habermann (1) Martha L Slattery (2) Elizabeth M Poole (3) Liren Xiao (4) Rachel L Galbraith (4) David Duggan (5) Richard J Kulmacz (6) Bette J Caan (7) Lisel Koepl (4) Anne Coghill (4) Karen W Makar (4) Abbie Lundgreen (4) John D Potter (4) Cornelia M Ulrich (1)
(1) DKFZ / NCT (2) University of Utah(3) Harvard Medical School / Harvard School of Public Health(4) Fred Hutchinson Cancer Research Center(5) Translational Genomics Research Institute(6) University of Texas Health Science Center at Houston(7) Kaiser Permanente Medical Research Program

Inflammation plays a major role in colorectal cancer. Prostaglandin-endoperoxide synthase (PTGS) 1and 2 convert n-3 and n-6 fatty acids into prostaglandins which have anti- and pro-inflammatory effects, respectively. We examined potential interactions between polymorphisms in *PTGS1* and *PTGS2* and fat and fatty acid (FA) intake in relation to colon and rectal cancer risk. We genotyped SNPs in *PTGS1* and *PTGS2* in two independent study populations. We investigated potential interactions with intake of total fat, saturated FA (SFA), monounsaturated FA (MUFA), polyunsaturated FA (PUFA), arachidonic acid (AA), eicosapentaenoic acid (EPA), and docosahexaenoic acid (DHA) on risk of colon or rectal cancer. We observed several significant interactions between *PTGS1* and *PTGS2* genotypes and fat and FA intake. Few interactions were consistent across fat variables and fell into three general categories of associations: a) individuals with the variant allele and low intake of total fat, SFA, or MUFA were at increased risk for rectal cancer; b) individuals with the variant allele and high EPA and DHA intake were less likely to develop colon cancer; and c) individuals with the wildtype genotype and high EPA and DHA intake had stronger protection from colon cancer. No interactions were observed for AA. Our study provides evidence that genetic variability in *PTGS1* and *PTGS2* and intake of total fat, SFA, or MUFA and EPA or DHA may interact in causing rectal and colon cancer, respectively.

127
Whole-Genome Detection of Deletions Associated with Glioma in a Case-Control Study Using SNPs
Chih-Chieh Wu (1) Robert Yu (1) Long Ma (1) Georgina Armstrong (1) Yanhong Liu (1) Ching Lau (2) Melissa Bondy (1) Sanjay Shete (1)
(1) Dept Epidemiology, M. D. Anderson Cancer Center (2) Texas Children's Cancer Center, Baylor College of Medicine

Deletion copy number variations have been known to influence many microdeletion syndromes and are frequently observed in patients with certain neuron-developmental disorders and cancers. Glioma accounts for ~80% of malignant primary brain tumors. The increased risk in glioma is particularly high for the probands' siblings with standardized incidence ratio=2.56. There are about 22,000 patients with malignant brain tumors diagnosed in the US in 2010. A recent genome-wide association study identified five susceptibility loci for glioma. The effect of these loci collectively accounts for only 7-14% of the excess familial risk of glioma. We recently developed a genome-wide statistical method for detecting disease-associated deletion variants using high-density SNP genotype data. In this study, we used this method to analyze Illumina Hap550K SNP genotypes from a genome-wide association study of glioma, which consists of 1248 glioma patients and 2243 control individuals. Our method was designed to detect statistically significant evidence of a deletion at individual SNPs for SNP-by-SNP analyses, and then combines the information between neighboring SNPs for cluster analyses. Our preliminary SNP-by-SNP analysis outcome identified 58 significant SNPs over the whole genome at a nominal significance level of 10^{-7}. Subsequently, we are performing cluster analyses and Penn CNV analyses.

128
Further Studies Of The Genetic Architecture Of Lung Cancer
Chris I Amos (1) Mala Pande (1) Alexander Li (1) Xifeng Wu (1) Margaret Spitz (1) Stacy Lloyd (1)
(1) UT MD Anderson Cancer Center

We genotyped 1681 nonsmall cell lung cancer cases and 1235 controls for 2168 markersselected as most significant in a prior GWAS of 1154 lung cancer cases and 1137 controls.The most significant markers in the combined analysis were in the alpha3/5/B4 region,in the HLA region and near hTERT.We found suggestive associationson chromosomes 10 near TMEM23 ($p=7.7 \times 10^{-6}$), on chromosome 14 in SLC24A4 ($p=9 \times 10^{-6}$) andon chromosome 19 near exosome subunit 5 (rs10853751, $p=2 \times 10^{-5}$). To further characterize association near hTERT we performed dense mapping of 215 markers across the TERT and CLPTM1L genes.The two SNPs rs370348 ($P=1.6\times10^{-6}$), and rs4975538 (OR=1.18, P=0.005) showed independent associations. We also conducted pathway-based analysis for the GWAS data using an approach developed by Luo et al., (Eur J Hum Genet. 18:1045-53,2010) that adjusts for correlations within and among geneswithin pathways.The following KEGG defined pathways were significant: Metabolism of xenobiotics by cytochrome P450 ($p=5.2 \times 10^{-5}$), PPAR signaling pathway ($p=1.7 \times 10^{-4}$), Basal cell carcinoma ($p=1.8 \times 10^{-4}$) and Glutathione metabolism ($p=7.3 \times 10^{-4}$). While these results suggest that additional loci influence lung cancer risk, further confirmations using meta-analyses from multiple research

groups are underway to further evaluate and replicate findings.

129

COX-2 (PTGS2) Promoter Variant Increases Risk Of Rectal Cancer

Karen Makar (1) Elizabeth M Poole (1) Karen Curtin (2) Dominique Scherer (3) Liren Xiao (1) AE Coghill (1) Sarah E Kleinstein (1) David Duggan (4) Richard J Kulmacz (5) Li Hsu (1) Christine Rimorin (1) Bette J Caan (6) John D Potter (1) Martha L Slattery (2) Cornelia M Ulrich (1)
(1) Fred Hutchinson Cancer Research Center, Seattle, USA(2) University of Utah, School of Medicine, Department of Medicine, Salt Lake City, USA(3) National Center for Tumor Diseases, Heidelberg, Germany(4) Translational Genomics, Phoenix, USA(5) University of Texas Health Science Center at Houston, Houston, USA(6) Kaiser Permanente Medical Research Program, Department of Research, Oakland, USA

COX-2 (encoded by *PTGS2*) is a central enzyme in prostaglandin synthesis, which has been unequivocally linked to colorectal carcinogenesis. No non-synonymous single nucleotide polymorphisms (SNPs) in *PTGS2* are known in Caucasians, however, a SNP in the promoter region (-765 G>C; rs20417) is associated with reduced expression of COX-2 and reduced concentrations of C-reactive protein (CRP). In this study, we used a case - unaffected sibling control design to genotype -765G>C in a large case-control study of colorectal cancer (Colon Cancer Family Registry, CCFR; n=3811). Subsequently, we replicated the observed associations in two case-control studies of colon (n=1420 cases / 1780 controls) and rectal cancer (n=583 cases / 775 controls). The -765 G>C homozygous variant was associated with a greater than 4-fold increased risk of rectal cancer in the CCFR population with a significant trend across variant alleles (odds ratio OR CC vs. GG=4.88; 95% confidence interval CI=1.54-15.44; OR GC vs. GG=1.36; 95% CI 0.95-1.94; p=0.01). This association replicated in a second case-control study of rectal cancer with a 2-fold increased risk (p=0.05). No significant associations were observed for colon cancer. These results suggest that a promoter variant with demonstrated functional impact specifically affects rectal cancer risk. Further follow-up on the underlying biologic mechanisms is needed.

130

Systematic Meta-Analysis for Common Low Penetrance Genes in Colorectal Cancer

Zahra Montazeri (1) Evropi Theodoratou (2) Julian Little (1) Harry Campbell (2)
(1) Ottawa University (2) The University of Edinburgh

To identify genetic variance influencing colorectal cancer (CRC), we apply systematic review and meta-analysis for a set of genes which were identified in GWAS studies. To date, 15 common genetic variants influencing risk of CRC have been reported from genome wide association studies for colorectal cancer. We identified and extracted data from published and unpublished studies up to 31 March 2011. We carried out meta-analysis to derive summary effect estimates for 18 polymorphisms in 15 genes; and obtained summary crude odds ratios and 95% CI for two additive models and one dominant model for variants that were identified from GWAS. We applied either the fixed effect model or in the case of heterogeneity the random effect model. Heterogeneity was quantified by calculating the Q statistic; we also calculated the I^2 heterogeneity metric. In order to detect a statistical significant effect, power of each meta-analysis was estimated. In assessing the credibility of genetic association, we consider the Venice criteria and the Bayesian False Discovery Probability (BFDP). We classified the genetic association in 3 categories as positive, less credible positive and negative. Furthermore, we applied the model free meta-analysis approach for those SNPs that were identified as "positive" The goal of this research is to find the genetic associations on CRC and interpreted appropriately to direct future research efforts.

131

Plasma MicroRNAs in Breast Cancer Detection

Cheryl L Thompson (1) Li Li (1) Rom S Leidner (1)
(1) Case Western Reserve University

Circulating microRNA (miR) levels have been proposed as a biomarker for cancer detection due to their role in cancer and stability in the circulation. A miR-based test for colorectal cancer is in clinical trials, but the field is much less advanced in breast cancer. We interrogated the expression level of 1145 miRs using the Illumina miR microarray from the plasma of 18 breast cancer patients prior to tumor resection, 17 patients post-resection and 20 controls (matched to cases on age and race). We excluded 245 miRs due to low expression across all samples. Differences in expression levels between pre-resection cases and controls were assessed via a pooled t-test, and 36 were differentially expressed between cases and controls ($p<0.01$). Using a ratio-normalized miR "signature" of 4 miRs, we are able to correctly classify 100% of controls and 50% of cases. In the post-resection patients, 90% returned to baseline levels. Furthermore, the expression is highly correlated with stage (lowest in in-situ cases and highest in stage 3/4). Further signature modeling using 6 miRs yields a test with 94% sensitivity and 95% specificity, and holds up to permutation testing ($p<0.01$). Validation in a larger independent population is warranted, and is underway. Our data provides compelling evidence of the potential of miRs as a minimally invasive screening test for breast cancer. The high specificity suggests utility in combination with mammography to increase accuracy.

132

Comparison of Methods for Evaluating the Predictive Benefit of Genetic Information for Prostate Cancer Risk

Paul J Newcombe (1) Brian H Reck (2) Jeilin Sun (3) Greg T Platek (4) Claudio Verzilli (1) A. Karim Kader (3) Seong-Tae Kim (3) Tao Jin (3) Zheng Zhang (3) S. Lily Zheng (3) Vincent E Mooser (5) Lynn D Condreay (2) Colin F Spraggs (1) John C Whittaker (1) Roger S Rittmaster (6) Jianfeng Xu (3)
(1) Genetics Division, GlaxoSmithKline Stevenage, SG1 2NY, UK (2) Genetics Division, GlaxoSmithKline, Research Triangle Park, NC, 27709, USA (3) Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA (4) Oncology Biostatistics, GlaxoSmithKline, Research Triangle Park, NC 27709,

USA (5)Genetics Division, GlaxoSmithKline, King of Prussia, PA, 19406, USA (6) Oncology Medicines Development, GlaxoSmithKline, Research Triangle Park, NC 27709, USA

We present the most comprehensive assessment to date of the predictive benefit of genetics for prostate cancer (PCa) in addition to clinical variables, using genotype data for 33 GWAS replicated markers in 1593 Caucasian men from the placebo arm of the REduction by DUtasteride of prostate Cancer Events (REDUCE(R)) trial. Furthermore, we conducted a detailed comparison of three techniques for incorporating genetics into clinical risk prediction. The first method was a standard logistic regression model which included separate terms for the clinical covariates and for each of the genetic markers, but ignores a substantial amount of external prior information concerning genetic effect sizes. The second and third methods investigated two possible approaches to incorporating meta-analyses of these external genetic effect estimates - one via a weighted PCa risk score based solely on the meta analysis estimates, and the other incorporating both the current and prior data via informative priors in a Bayesian logistic regression model.
All methods demonstrated a slight improvement in predictive performance upon incorporation of genetics. The two methods which incorporated external information showed the greatest increase in area under the receiver-operator curve (ROC) from 0.61 to 0.64. The value of our methods comparison is likely to lie in observations of performance similarities, rather than differences, between three approaches of very different resource requirements.

133

Ethanol Metabolizing Genes and Risk of Head and Neck Cancer: Preliminary Report

Jeffrey S Chang (1) Jenn-Ren Hsiao (2) Tung-Yiu Wong (3) Sen-Tien Tsai (2) Chun-Yen Ou (2) Hung-I Lo (2) Cheng-Chih Huang (2) Wei-Ting Lee (2) Jehn-Shyun Huang (3) Ken-Chung Chen (3) Yi-Hui Wang (1) Ya-Ling Weng (1)
(1) National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan (2) Department of Otolaryngology, Medical College and Hospital, National Cheng Kung University, Tainan, Taiwan(3) Department of Stomatology, Medical College and Hospital, National Cheng Kung University, Tainan, Taiwan

Head and neck cancer (HNC), including cancers of the oral cavity, pharynx, and larynx, is the fifth most common cancer in the world. One of the major risk factors of HNC is alcohol drinking; however, most drinkers do not develop HNC, suggesting a role of genetics. This analysis examines the association between HNC risk and two functional single nucleotide polymorphisms (SNPs) of two ethanol metabolizing genes (*ADH1B*: rs1229984, Arg48His and *ALDH2*: rs671, Glu504Lys). HNC patients and frequency-matched (by age and sex) non-cancer controls were recruited from the department of otolaryngology and department of stomatology. Information on alcohol consumption was ascertained by interviews. Genotyping was performed using TaqMan based real-time PCR. Unconditional logistic regression was performed with data of 88 HNC cases and 81 controls to estimate HNC risk associated with the two SNPs adjusted for age and sex. The results showed that those who carried at least one *ADH1B* 48Arg allele were twice as likely to develop HNC compared to homozygous 48His individuals [Odds ratio (OR)=1.97; 95% confidence interval (CI): 1.04-3.72]. The positive association between *ADH1B* 48His and HNC risk was more prominent among those who were heterozygous for *ALDH2* Glu504Lys (OR=3.19; 95% CI: 1.17-8.74) and among never drinkers (OR=5.51; 95% CI: 1.32-23.06). This study shows a significant contribution of *ADH1B* Arg48His to the development of HNC, which is modified by *ALDH2* Glu504Lys and alcohol drinking.

134

Effect Of Reproductive Factors And Body Mass Index On The Mutation Localization-Specific Risk Of Breast Cancer In The French National BRCA1/2 Carrier Cohort (GENEPSO)

Julie Lecarpentier (1) Catherine Nogues (2) Emmanuelle Mouret-Fourme (2) Dominique Stoppa Lyonnet (3) Christine Lasset (4) Olivier Carron (5) Jean-Paul Fricker (6) Laurence Gladieff (7) Laurence Faivre (8) Hagay Sobol (9) Paul Gesta (10) Marc Frenay (11) Elisabeth Luporsi (12) Isabelle Coupier (13) GENEPSO study (14) Rosette Lidereau (15) Nadine Andrieu (1)
(1) INSERM U900 Institut Curie-Inserm/Mines ParisTech (2) Institut Curie, Hopital Rene Huguenin(3) INSERM U509, Service de Genetique Oncologique, Institut Curie, Universite Paris-Descartes(4) Centre Leon Berard, Departement de Sante Publique(5) Institut de Cancerologie Gustave Roussy, Service d'Oncologie Genetique(6) Centre Paul Strauss, Service d'Oncologie, Departement de Biologie(7) Institut Claudius Regaud, Service d'Oncologie Medicale(8) Hopital d'Enfants, Service de Genetique Medicale(9) Institut Paoli-Calmettes, Departement d'Oncologie Genetique(10) C.H.R. Georges Renon, Pole Oncologie(11) Centre Antoine Lacassagne, Unite d'Oncogenetique(12) Centre Alexis Vautrin(13) Hopital Arnaud de Villeneuve, CHU Montpellier, Service de Genetique medicale et Oncogenetique(14) Collaborating Centers(15) INSERM U735 and Laboratoire d'Oncogenetique, Hopital Rene Huguenin

Mutations in BRCA1/2 confer a high risk of breast cancer (BC), but the magnitude of this risk varies according to various factors. Although still controversial, there are data to support the hypothesis of allelic risk heterogeneity.
We assessed variation in BC risk according to reproductive factors and body mass index (BMI) by location of mutation in homogenous risk region of BRCA1/2 on 990 women by using a weighted Cox-regression model. Homogenous risk regions have been previously defined as lower risk of BC for central regions of BRCA1 and BRCA2 and as increased risk of BC in the C-terminal region of BRCA2 (Lecarpentier et al. 2011, in revision)
Among the studied factors, we found a decrease risk of BC associated with having more than 3 fulltermpregnancies (HR=0.5, $p<10^{-3}$) and an increased risk of BC associated with having a BMI less than 18.5 (HR=2.1, $p<10^{-3}$) and no variation in risk according to mutation location. However, a possible variation in BC risk associated with menopausal status according to the location of the mutation in BRCA1 was suggested. Menopause was associated with an increased BC risk for mutations outside the central region (HR=2.4) and with a decreased risk for mutations in the central region (HR=0.5) ($P_{het}=0.04$). There is no obvious biological process to explain this variation however, taking

together environmental/lifestyle modifiers and location of mutations might be important in the clinical management of BRCA mutation carriers.

135

Comparison Of Count Models Regarding Their Ability To Capture The Relationship Between Genomic Instability, Methylation And Expression In Human Hepatocellular Carcinoma

Miriam Kesselmeier (1) Thomas Longerich (2) Robert Geffers (3) Matthias Ganzinger (1) Justo Lorenzo Bermejo for the SFB/TRR77 Consortium "Liver Cancer-From Molecular Pathogenesis to Targeted Therapies" (1)
(1) University Hospital Heidelberg, Institute of Medical Biometry and Informatics(2) University Hospital Heidelberg, Institute of Pathology(3) Helmholtz Center for Infection Research, Department of Cell Biology and Immunology

A better understanding of the relationship between chromosomal alterations and gene expression may facilitate the identification of prognostic biomarkers for human hepatocellular carcinoma (HCC).
The SFB/TRR77 Consortium "Liver Cancer-From Molecular Pathogenesis to Targeted Therapies" has generated a unique collection of patients with array-based information on comparative genomic hybridization (aCGH), gene expression and gene methylation. We have applied standard and tailored count models to examine the dependence between genomic instability and the methylation and expression of around 600 genes previously proposed to define expression-based subclasses of HCC. Instability was measured by the number of chromosome arms with an aCGH-based gain or loss of genetic material.
The 60 investigated tumor samples showed 4 to 35 instable arms (median 14). The data was better accommodated by negative binomial and robust Poisson regression than by standard Poisson regression. A forward variable selection based on Akaike's information criterion included the methylation of a single gene or, alternatively and with a poorer fit, the expression of three different genes, in the best models of genomic instability. A weak relationship between methylation and expression was found for these four genes (Spearman's rank rho from -0.34 to 0.19).
Modeling details and results based on alternative parameterizations of the genomic instability will be discussed during the meeting.

136

Mixed Effects Cox Models For Gene Set Analysis In Lung Cancer

Marianne Huebner (1) Terry Therneau (2)
(1) Statistics, Michigan State University and Mayo Clinic(2) Biostatistics, Mayo Clinic

Adding random effects to proportional hazard models enables direct analysis of survival endpoints and their association to both clinical covariates and gene expression data. Mixed effects Cox models allow for rich gene interactions while adjusting for clinical covariates. An exciting aspect of this method is the ability to use a patterned covariance matrix that captures available data on gene-gene interactions within pathways allowing for both close interaction and negative feedback. The mixed effects model performed reliably well for simulation scenarios of gene set sizes from 10 to 100 and sample sizes from 100 to 300 even when a percentage of coefficients have opposite signs. We evaluated the power of the method and the finite-sample probability mass at zero under a range of scenarios. This methodology was applied to a non-small lung cancer dataset. There were 592 unique gene sets with sizes ranging from 9 to 470 with a median of 26. Results from the mixed effect Cox model with two parameters (random effects variance and correlation) were compared to those applying the maxmean statistic (Ann Appl Stat (2007) 1:107-129). For this collection of gene sets the largest values for both methods correspond to chromosomal regions, which is possibly an indicator for copy number variation. The mixed effects Cox models were fit using the coxme function within the R statistical system. http://cran.r-project.org.

137

Trade-Off In The Effects Of The APOE Polymorphism On The Ages At Onset Of CVD And Cancer: Insights From The Genetic Stochastic Process Model

Konstantin G. Arbeev (1) Alexander M. Kulminski (1) Svetlana V. Ukraintseva (1) Liubov S. Arbeeva (1) Igor Akushevich (1) Irina V. Culminskaya (1) Deqing Wu (1) Anatoliy I. Yashin (1)
(1) Duke University

In our recent study (Aging Cell 10 (3): 533-541, 2011), we demonstrated trade-off in the effects of apolipoprotein E (APOE) e2/3/4 polymorphism on ages at onset of cardiovascular diseases (CVD) and cancer in participants of the Framingham Heart Study (FHS). In this study, we applied the genetic stochastic process model (J Theor Biol 258 (1): 103-111, 2009) to data on incidence of CVD and cancer and longitudinal measurements of physiological indices in FHS to investigate possible mechanisms responsible for the observed trade-off. We performed joint analyses of data on genetic (carriers vs. non-carriers of the APOE e4 allele) and non-genetic (those for whom genetic data were not collected) subsamples of FHS. We found that carriers and non-carriers of the APOE e4 allele differ in shapes and age dynamics of various aging-related characteristics, such as physiological "norms" (i.e., the age-specific values of indices minimizing risks of onset of the diseases), aging-related decline in stress resistance (associated with the narrowing of the U-shape of the risk as a function of a physiological index) and adaptive capacity, and mean allostatic trajectories (i.e., the trajectories of the indices that organisms are forced to follow by the process of allostatic adaptation). We conclude that the differences in these characteristics and associated aging-related processes may contribute to the observed differential effects of the APOE e4 allele on risks of the two diseases.

138

Aetiological Role Of Folate Deficiency In Congenital Cardiovascular Malformation: Evidence From "Mendelian Randomisation" And Meta-Analysis

Chrysovalanto Mamasoula (1) Tomasz Pierscionek (1) Darroch Hall (1) Ana Topf (1) Julian Palomino Doza (1) Thahira Rahman (1) Angeline Tan (1) Jamie Bentham (2) Shoumo Bhattacharya (2) Caroline Cosgrove (2) David Brook (3)

Javier Granados Riveron (3) Frances A Bu'Lock (4) John O'Sullivan (4) Christopher Wren (5) Judith A Goodship (1) Heather J Cordell (1) Bernard Keavney (1)
(1) Institute of Genetic Medicine, Newcastle University UK(2) Department of Cardiovascular Medicine, Oxford University, UK(3) Institute of Genetics, Nottingham University, UK(4) University Hospitals of Leicester NHS Trust, Leicester, UK(5) Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle, UK

The existence of a causal relationship between lower levels of plasma folate and congenital cardiovascular malformation (CVM) remains contentious. We present a genetic approach using "Mendelian randomization" to determine the causality of folate in CVM risk. We compared genotype frequencies at the (MTHFR) C677T SNP in 1186 CVM cases and 4168 controls. The TT genotype at MTHFR C677T is known to be associated with lower activity of MTHFR and plasma folate, and higher levels of plasma homocysteine. We placed our results in the context of a meta-analysis including 3069 CHD cases and 7271 controls. We conducted sensitivity analyses to examine folate fortification of flour as a potential source of heterogeneity. The primary genotyping data in 1186 cases and 4168 controls revealed a trend towards increased risk with the TT genotype, but this did not reach statistical significance (OR 1.15 [95% CI 0.94-1.40]). Combination of our primary data with previous studies, however, revealed association in the complete dataset (OR 1.45 [95% CI 1.12-1.89]; p=0.005). Studies conducted in countries with mandatory folate fortification showed no effect of C677T genotype on CHD risk, whereas studies conducted in countries without mandatory fortification showed a significant effect of genotype. The absence of a genetic association in studies performed in countries practicing folate fortification suggests that fortification largely abrogates the risk of CHD attributable to folate deficiency.

139

Nonalcoholic Fatty Liver Disease Predicts Coronary Heart Disease Independent Of Overall Obesity And Insulin Resistance In Non-Diabetics In The NHLBI Family Heart Study

Mary F. Feitosa (1) Alexander P. Reiner (2) Mary K. Wojczynski (1) Kari E. North (3) John J. Carr (4) Ingrid B. Borecki (1)
(1) Div Statistical Genomics, Washington Univ Sch Medicine, St Louis, MO(2) Dept Epidemiology, Univ of Washington, Seattle, WA(3) Dept Epidemiology, Univ of North Carolina, Chapel Hill, NC(4) Dept Radiology, Wake Forest University School of Medicine, Winston Salem, NC.

NAFLD is associated with obesity, insulin resistance (IR), and type 2 diabetes (T2D), which are known risk factors for CHD; however, the underlying biologic mechanisms of these risk factors in the development of CHD are unclear. CT measured liver attenuation is inversely related to the amount of fatty liver (FL), and represents a non-invasive measure of steatosis. Alanine aminotransferase (ALT) is a NAFLD inflammation marker. We investigated whether NAFLD using two definitions, FL and ALT (?40U/L: affected), were associated with higher CHD risk (self report of MI, PTCA, or bypass) in ~2,300 subjects. Heavy drinkers (men:?21 drinks/wk, women:?14 drinks/wk) and subjects that tested positive for hepatitis-C were excluded. Using a linear mixed model and adjusting for gender, age, centers, smoking, and alcohol consumption, FL (p=0.008) and ALT (p=0.013) were individually associated with CHD. These associations were attenuated when we further adjusted for HOMA-IR, T2D and BMI. In the multivariate model, HOMA-IR ($p<0.0001$) and T2D ($p<0.0001$) were predictors of CHD. Since T2D is associated with CHD, we employed a similar multivariate model in non-T2D. We found that ALT (p=0.0195, OR=2.241, 95%CI: 1.139-4.409) and HOMA-IR (p=0.0008, OR=1.464, 95%CI: 1.172-1.828) were the only significant predictors of CHD. This finding suggests that: a) while simple steatosis was not associated with CHD in presence of IR; b) ALT was a predictor of CHD independently of IR and BMI in non-T2D.

140

DNA Repair Enzyme Genes and Congenital Heart Defects

Sadia Malik (1) Mario A Cleves (1) Stewart L Macleod (1) Stephen W Erickson (1) Charlotte A Hobbs (1) National Birth Defects Prevention Study (2)
(1) University of Arkansas Medical Sciences(2) NBDPS

Background: Maternal smoking in pregnancy has been associated with congenital heart defects (CHDs). DNA repair enzymes act as a primary defense mechanism against mutagenic exposures secondary to tobacco by-products. We hypothesize that specific genetic polymorphisms in genes that encode DNA repair enzymes increase the risk of non-syndromic CHDs among offspring prenatally exposed to tobacco.
Methods: Questionnaire and biologic data were obtained from infants who were prenatally exposed to tobacco smoke and their parents,all of whom were participants of the National Birth Defects Prevention Study. DNA was genotyped for single nucleotide polymorphisms (SNPs) in DNA repair enzyme genes using an Illumina Golden Gate 384 Custom SNP panel.Infant and maternal case-control data was used to estimate the association between CHDs and each SNP.
Results: DNA Samples from 244 infant-maternal dyads with congenital heart defects and 465 infant-maternal dyads with no birth defects were genotyped. We identified multiple fetal SNPs within the NEIL2 gene and one SNP from the XPC and XRCC1 gene that were associated with CHD. Maternal case-control analysis also revealed an association with maternal SNPs in the NEIL1, ERCC1, XPC and GTF2H1 genes and CHD in infants.
Conclusion: We have identified polymorphisms in DNA repair enzyme genes that are associated with increased risk of CHD-affected pregnancies in a subgroup of women who smoke during pregnancy.
Funding: 5K08HL090494-03 from NHLBI/NIH.

141

DC-SIGN Gene Promoter Polymorphisms In Coronary Artery Disease And IVIG Treatment Outcomes In Hispanic Kawasaki Disease (KD) Patients

Sadeep Shrestha (1) Howard W Wiener (2) Aditi Shendre (2) Michael A Portman (3)
(1) Dept of Epidemiology, University of Alabama at Birmingham(2) Dept of Epidemiology, University of Alabama at

Birmingham(3) Dept of Pediatrics, University of Washington, Seattle Children's Hospital

Dendritic cell-specific intracellular adhesion molecule-3-grabbing nonintegrin (DC-SIGN), encoded by a member of the CD209 gene family, has been identified as a specific receptor for sialylated Fc, the component responsible for the anti-inflammatory action of intravenous immunoglobulin (IVIG), the principal therapy for Kawasaki Disease (KD). We tested 5 SNPs in the promoter of DC-SIGN, in a cohort of 427 KD patients to examine the association with IVIG refractoriness (IVIG-R) - defined by AHA guidelines, and coronary artery disease (CAD) - dilation (Z-score>2.5) or aneurysm persisting>6 weeks after appropriate IVIG treatment (2 gm/kg). A case-control approach was performed, separately for Caucasians, Asians, and Hispanics as determined by the principal component analysis (PCA) of 155 Ancestry Informative Markers (AIMs), to examine the differential distribution of alleles and genotypes according to IVIG-R and CAD. Two SNPs (rs2287886 and rs4804804) showed significant genotype association ($p<0.02$) with IVIG-R among Hispanics (49 responders and 19 non-responders); rs2287886 and rs4804803 showed both allelic and genotype associations with CAD also in the Hispanic populations (31 CAD and 44 non-CAD, $p<0.004$). In our study, the genetic associations with variants in DC-SIGN gene are specific to Hispanic population, of primarily Mexican and Central American origin, but no associations in other ethnic groups.

142

Genetic and Epigenetic Analysis of Neonatal and Early Childhood Phenotypic Outcomes in a Community-Based Longitudinal Cohort in Memphis, TN: The CANDLE Study

Frances Tylavsky (1) Collin Hovinga (2) Laura Murphy (1) Carolyn Graff (1) Frederick Palmer (1) Fridtjof Thomas (1) Vicki Park (1) Pamela Connor (1) Eszter Volgyi (1) Ronald Adkins (1) Julia Krushkal (1)
(1) University of Tennessee Health Science Center (2) University of Texas, Austin School of Pharmacy

As molecular analysis techniques continue to rapidly develop and advance, molecular measurements, such as changes in DNA methylation and gene expression changes, can serve both as predictors of health outcomes and intermediate molecular phenotypic outcome variables. Conditions Affecting Neurocognitive Development and Learning in Early Childhood (CANDLE) Study is designed to follow 1,500 pregnant women (68% black, 60% below the poverty level) and their children living in Memphis/Shelby County, TN from the second trimester into early childhood. The study is constructed to investigate molecular associations with the following: demographic, health and nutritional information from mother and infant; infant cognitive, language, socioemotional, and adaptive outcomes at 12, 24 and 36 months; periodic screening for specific disabilities, maternal mental health and intellectual functioning; and maternal-infant interactions influencing child development across urban and suburban neighborhoods. The study collects extensive anthropometric,psychosocial, socioeconomic, and environmental measures during fetal and postnatal development. We have collected a rich data set of molecular genome-wide single nucleotide polymorphisms, copy number variants, DNA methylation, and gene expression profiles on a subset of CANDLE newborns and their mothers and are currently analyzing the links between molecular variants and pregnancy and early childhood outcomes.

143

Association Between Serum Uric Acid And Adiposity Markers: Mendelian Randomization Using SLC2A9 Variants

Tanica Lyngdoh (1) Philippe Vuistiner (1) Pedro Marques-Vidal (1) Valentin Rousson (1) Gerard Waeber (2) Peter Vollenweider (2) Murielle Bochud (1)
(1) Institute of Social and Preventive Medicine (IUMSP), University of Lausanne, Switzerland (2) Department of Medicine, Internal Medicine, CHUV, Lausanne, Switzerland

High serum uric acid (SUA) is known to be associated with an increased risk of adiposity. We examined the causal association of SUA with adiposity markers using *SLC2A9* variants as instruments in a Mendelian randomization approach in 2630 men and 2955 women in Lausanne. Single nucleotide polymorphisms (*rs7442295* in men and *rs7669607* in women) within *SLC2A9* gene were used as instrumental variables. Adiposity markers included weight, body mass index (BMI), waist circumference and fat mass. We observed highly significant positive associations between SUA and weight, BMI, waist and fat mass in an ordinary least square regression (? coefficient [95%CI]=3.38 [3.24,4.42], 1.24 [1.04,1.43], 3.65 [3.12,4.18]and 2.19 [1.84,2.54]in men and 6.38 [5.71,7.05], 2.45 [2.19, 2.70],6.38 [5.73, 7.03] and 4.67 [4.17,5.16] in women, respectively, all $p<0.001$). However, using genetic variants within *SLC2A9* in a two-stage least square regression, SUA explained by *rs7442295* and *rs7669607* was not significantly associated with any adiposity marker (estimates close to zero and often even negative) in both genders. Our findings provide no evidence for a positive causal effect of SUA on adiposity markers. These results are compatible with a possible reverse causality (i.e. elevated adiposity leading to hyperuricemia) or a common cause shared by hyperuricemia and elevated adiposity (e.g. fructose intake) or failure to fulfill the assumptions underlying Mendelian randomization.

144

UGT1A1 And Serum Bilirubin In American Indians: The Strong Heart Family Study

Phillip E. Melton (1) Karin Haack (1) Harald H. Goring (1) Sandra Laston (1) Jason G. Umans (2) Elisa T. Lee (3) Richard R. Fabsitz (4) Richard B. Devereux (5) Lyle G. Best (6) Jean W. MacCluer (1) Laura A. Almasy (1) Shelley A. Cole (1)
(1) Texas Biomedical Research Institute(2) MedStar Health Research Institute(3) University of Oklahoma Health Sciences Center(4) National Heart, Lung, and Blood Institute(5) Weill Cornell Medical College(6) Missouri Breaks Industries Research, Inc

Serum bilirubin is an important antioxidant thought to be protective against cardiovascular disease and certain types of cancer. Genetic variation within the promoter

of uridine diphosphate glucuronosyltransferase (*UGT1A1*) on chromosome 2q appears to be responsible for differences in serum bilirubin levels in European populations. However, no study has investigated genetic regulation of serum bilirubin levels in American Indians (AIs). We conducted a conditional linkage analysis of AIs in the Strong Heart Family Study in an attempt to replicate the chromosome 2q bilirubin quantitative trait locus (QTL). Statistical analyses were carried out with 3,484 AI participants recruited from three areas in the US: Arizona, Oklahoma, North and South Dakota. Variance components linkage analysis detected a QTL for bilirubin on chromosome 2q in the combined centers (LOD=6.61, p=4.24e$^{?6}$) and in Oklahoma, alone (LOD=5.65, p=4.57e$^{?5}$). After adjustment using conditional linkage for the *UGT1A1* promoter variant, the LOD score dropped to 1.10 in the combined sample and to a LOD score of 3.32 (p=0.02) in Oklahoma, indicating this known polymorphism is not completely responsible for the QTL in AIs. Suggestive QTLs were also detected in the Dakotas on chromosome 10p12 (LOD=2.18) and in the combined centers (LOD=2.24) on chromosome 10q21. We replicated a serum bilirubin QTL on chromosome 2q in AIs implicating *UGT1A1* but further genotyping is warranted to identify additional functional polymorphisms.

145

A Major Pleiotropic Locus On Chromosomal Region 11p15 Controls Mycobacteria-Triggered TNF Production

Aurelie Cobat (1) Eileen G Hoal (2) Erwin Schurr (1) Alexandre Alcais (3)
(1) McGill Center for the Study of Host Resistance, McGill University, Montreal, Canada(2) DST/NRF Centre of Excellence for Biomedical TB Research, Stellenbosch University and MRC, Tygerberg, South Africa(3) Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U980, Paris, France

Tuberculosis (TB) is a chronic infectious disease caused by *M. tuberculosis* (*Mtb*) that remains a major public health challenge worldwide, with 9 million new cases and 2 million deaths each year. Tumour necrosis factor (TNF) is a key immune regulator of TB resistance as is shown by the highly increased risk of TB disease among individuals receiving TNF-blocker therapy. We determined the extent of TNF production after stimulation with BCG or BCG plus interferon g (IFN-g) in 134 nuclear families from an area hyperendemic for TB in South Africa. Bivariate linkage analysis of TNF production after stimulation by BCG alone and BCG plus IFN-g was conducted using the new MLB-QTL method[1] extended to multivariate analysis, following an approach proposed by Mangin et al[2]. We detected a major locus ($p < 10^{-5}$) on chromosomal region 11p15 that controls TNF production after stimulation by both BCG alone and BCG plus IFN-g. Interestingly, this major locus overlaps the *TST1* locus that controls T-cell-independent resistance to *Mtb* infection.[3] This multivariate approach can easily be extended to the genetic analysis of other complex multivariate traits, as illustrated in.[4]

1. Cobat A., et al. 2011. *Genet Epidemiol* 35:46-56.
2. Mangin B., et al. 1998. *Biometrics* 54:88-99.
3. Cobat A., et al. 2009. *J. Exp. Med.* 206:2583-2591.
4. Bouzigon E., et al. 2007. *Hum Genet.* 121:711-719.

146

Association of Adiponectin Plasma Levels in African Americans with SNPs from Two Arrays

Sarah G Buxbaum (1) Solomon Musani (2) Matthew Allison (3) Jorge Kizer (4) Susan Redline (5) James G Wilson (2) Aurelian Bidulescu (6)
(1) Jackson State University(2) University of Mississippi Medical Center(3) University of California, San Diego School of Medicine(4) Weill Cornell Medical College(5) Division of Sleep Medicine, Harvard Medical School(6) Morehouse School of Medicine, Cardiovascular Research Institute and Department of Preventive Medicine and Community Health

Introduction: Adiponectin is a biomarker of systemic inflammation and insulin resistance and a putative predictor of subclinical and clinical cardiovascular disease. Four cohorts with adiponectin measurements that included African Americans were analyzed.
Methods: Using both the Affymetrix 6.0 (Affy) and the ITMAT Broad-CARe (IBC) SNP arrays, we performed a meta-analysis that included up to 2487 African-Americans for association with log-Adiponectin blood plasma level adjusted for age, sex and body mass index. Samples were from the Candidate gene Association Resource cohorts and included data from Jackson Heart Study, the Multi-Ethnic Study of Atherosclerosis, the Cleveland Family Study and the Cardiovascular Health Study (IBC data only).
Results: The strongest evidence for association for the Affy-based analysis (N=2375) adjusting for global ancestry was for rs4686807 at 3q27.3, p=6.01 $\times$ 10^{-7}. Further adjustment for local ancestry modestly increased the strength of association (1.69 $\times$ 10^{-7}). The most strongly associated SNP in the IBC array (N=2487) was at 6p21.1 with a p-value of 1.37 $\times$ 10^{-5}.
Conclusion: These analyses confirmed a known locus for adiponectin and identified a potential novel locus. The association at 3q27.3 is of interest because the closest gene in RefSeq is ADIPOQ, 21.1 kbp away. The association at 6p21.1 is within *TREM2*. These findings require replication in other cohorts and a meta-analysis is ongoing.

147

Novel Functional Variants For Serum Uric Acid And Total Serum Bilirubin Levels In An Irish Population

Cheryl D. Cropp (1) Yoonhee Kim (1) Anne M. Molloy (2) James L. Mills (3) Peader N. Kirke (4) John M. Scott (2) Lawrence C. Brody (5) Alexander F. Wilson (1) Joan E. Bailey-Wilson (1)
(1) Inherited Disease Research Branch, NHGRI, NIH(2) School of Immunology and Biochemistry, Trinity College, Dublin, Ireland(3) Division of Epidemiology, Statistics, and Prevention Research, NICHD, NIH (4) Child Health Epidemiology Unit, Health Research Board of Ireland, Dublin, Ireland(5) Genome Technology Branch, NHGRI, NIH

The catabolic products of purine and hemoglobin metabolism can directly cause disease and serve as pathologic markers of several disease states. We present results from a GWAS of the purine breakdown product, serum uric acid (URIC, mmol/L) and the major hemoglobin breakdown product, total serum bilirubin (TBIL, umol/L).

Genotyping was performed with the Illumina 1M HumanOmni1-Quad chip. After quality control 2232 unrelated, healthy individuals from Trinity College Dublin and 757,533 SNPs were retained for association testing that was performed with simple univariate linear regression, assuming an additive genetic model (PLINK v1.0.7). Locus-specific heritability (h^2) was calculated with R.
For URIC, we replicated previously reported SNPs in *SLC2A9* (rs6449213, h^2=10%), *WDR1* (rs717615, h^2=3%), and *ABCG2* (rs2199936, h^2=2%) at genome-wide significance levels (p-value?5e-08). The most significant results found were novel variants in *SLC2A9*: rs13111638 (p-value=4e-24, h^2=9.5%) in an intronic region, and two coding SNPs, rs10939650 and rs13113918 (p-value=2e-22, 1e-21; h^2=4.4%, 5.6% respectively). For TBIL, strong *UGT1A* signals were found including rs887829 (p-value=4e-156, h^2=28%). Two non-synonymous SNPs in *LOC339766* (rs6431631 & rs1500480, p-value=3e-09, 2e-08 respectively) were significantly associated with TBIL.
In conclusion, common variants associated with URIC and TBIL were replicated and novel coding genetic variants were found in *SLC2A9* and *LOC339766*, respectively.

148

Gene-Set Testing On Autosomes And Sex Chromosomes For Complex Phenotypes: An Application In Follow-Up Data For Rheumatoid Arthritis

Roula Tsonaka (1) Jeanine J Houwing-Duistermaat (1)
(1) Medical Statistics and Bioinformatics, Leiden University Medical Center

In association studies a gene-set analysis can be more powerful than separate locus analyses. In the literature various approaches have been proposed to test associations between gene-sets and simple phenotypes. However, development of gene-set-based tests for complex phenotypes, such as longitudinally measured responses, has received little attention. In this work, motivated by a follow up study on patients with Rheumatoid Arthritis (RA), we propose a two-stage approach to test the joint effect of SNPs at a gene-set on the disease progression. In particular, in the first stage we use a mixed-effects model with a general random-effects structure to capture the correlations between the SNPs, and in the second stage we test for gene-set effects using the empirical Bayes estimates of the random effects as covariates in the model for the longitudinal phenotype. The advantage of this approach is its broad applicability, since it can be used for any phenotype and can be implemented with standard statistical software. Moreover, it can be easily extended to test associations on the X chromosome. Finally, our approach takes advantage of the correlation between the SNPs which can considerably increase the power of the tests. In particular in the RA study, using our approach we have found a statistically significant gene-set effect (p=0.020), whereas no association could be established using a separate locus analysis after multiple testing correction.

149

The Ano3/Muc15 Locus Is Associated With Eczema In Family Samples Ascertained Through Asthmatics

Marie-Helene MH Dizier (1) Patricia P Jeannin (1) Anne-Marie AM Madore (2) Jorge J Esparza (3) Myriam M Moffatt (4) Eve E Corda (1) Florent F Monier (1) Isabella I Annesi-Maesano (5) Jocelyne J Just (6) Isabelle I Pin (7) Francine F Kauffmann (8) William W Cookson (4) Young-Ae YA Lee (3) Catherine C Laprise (2) Mark M Lathrop (9) Emmanuelle E Bouzigon (1) Florence F Demenais (1)
(1) Inserm U946, Paris, France(2) Universite du Quebec, Chicoutimi, Canada(3) Charite Universitatsmedizin Berlin, Germany(4) National Heart Lung Institute, Imperial College, London, UK(5) Inserm U707, Paris, France(6) Centre de diagnostic et traitement de l'asthme, Hopital Trousseau, Paris, France(7) Departement de Medecine Aigue Specialisee, CHU Michallon, Grenoble, France(8) Inserm U1018, Paris, France(9) CEA-CNG, Evry, France

A previous genome-wide linkage scan in 295 families of the French Epidemiological study on the Genetics and Environment of Asthma (EGEA) reported strong evidence of linkage of 11p14 to eczema. Our purpose was to conduct fine-scale mapping of the 11p14 region to identify the genetic variants associated with eczema. Association analyses were conducted in the EGEA discovery dataset using two statistical methods for internal validation: the family based association method (FBAT) and logistic regression. Replication of the EGEA findings was sought in French Canadian (SLSJ study) and UK (MRCA study) family samples, which similarly to EGEA, were ascertained through asthmatic subjects. We also tested for association in two German samples ascertained through subjects affected with eczema. We found significant association of eczema with 11p14 SNPs in EGEA ($p=10^{-4}$ for rs1050153 using FBAT, that reached the multiple testing-corrected threshold of 1.3×10^{-4}; p=0.003 using logistic regression). Pooled analysis of EGEA, SLSJ and MRCA samples showed strong improvement in the evidence for association ($p=6\times10^{-6}$ for rs293974, $p=3\times10^{-5}$ for rs1050153, $p=6\times10^{-5}$ for rs15783). No association was observed in the eczema-ascertained samples. The significant SNPs are located within the overlapping *ANO3* and *MUC15* genes. Further investigation is needed to confirm and better understand the role of *ANO3/MUC15* locus in eczema and its relationship with respect to asthma.

150

Genetic Susceptibility To Language Development And Attention-Deficit/Hyperactivity Disorder

Sophie Tezenas du Montcel (1) Cathy L Barr (2) Michel Boivin (3) Bruno Falissard (4) Ginette Dionne (3)
(1) Universite Pierre et Marie Curie - Paris 6(2) Toronto Western Research Institute(3) Universite Laval(4) INSERM

Attention-deficit/hyperactivity disorder (ADHD) is the most prevalent psychiatric disorder emerging during childhood and is frequently associated with language deficits. It has been shown that early hyperactivity and/or inattention (H/I) symptoms negatively affect language development, which in turn contribute to H/I symptoms by early school-age. Genetic background has been shown to contribute both to ADHD and to language development specifically one candidate region on chromosome 6. We examined the genetic influence on the developmental course of the association between language skills and H/I using genetically informative data from the Quebec Newborn Twin Study (QNTS). Participants were twins assessed on language skills and H/I symptoms at 18 and 30, 5, 7 and 8 years old. Three candidate genes on chromosome 6

(DCDC2, VMP, KIAA0319) were genotyped for both twins and their parents when available.
Cross-lagged designs model the development and the interaction of phenotypes across time. Twin studies enable to estimate the relative contribution of genes, common and unique environments on phenotypic variance. We modified the cross-lagged models to fit twin data and the twin model to test for the influence of the candidate genes by taking into account the exact percentage of genes shared between DZ twins. Results show that the developmental course of the association between language skills and H/I is under the influence of the candidate genes on chromosome 6.

151
Gene-Based Association Study For Specific Language Impairment
Patrick D Evans (1) Eric Gamazon (1) J. B. Tomblin (2) Dan Nicolae (1) Nancy J. Cox (1)
(1) The University of Chicago(2) The University of Iowa

The acquisition and use of language is found in all cultures and is often considered to be one of the hallmark features of humans. Despite this universal feature of language among humans, some are more skilled language learners and users than others. Those children who present with particularly poor of language skills are at much higher risk for academic, social and later occupational limitations and are therefore viewed as presenting with language impairment (LI).
We performed a genome-wide association study (GWAS) on 429 children of European-descent from the Iowa. These results were used to perform a gene-based test that incorporates only functional SNPs. This includes missense, nonsense, brain eQTL data, and SNPs that are found within evolutionarily conserved regions. This test allows for reduced multiple-test correction instead of having to correct for all SNPs used in a traditional GWAS.
This analysis was performed for the SLI phenotype. Several genes as well as conserved elements with unknown function were identified using this method. The top hit, ABCG8, is a ATP-binding cassette. Other genes implicated are: LIPI, DPY19L2P1, FAM49B, and KCNJ6.

152
Methodological Approaches To Evaluate Teratogenic Risk Using Birth Defect Registries: Advantages And Disadvantages
Fernando A Poletta (1) Juan A Gili (1) Emanuele Leoncini (2) Mastroiacovo Pierpaolo (2) Eduardo E Castilla (1) Lopez Camelo S Jorge (1)
(1) ECLAMC (Latin American Collaborative Study of Congenital Anomalies) at CEMIC - CONICET(2) Headquarters of the International Clearinghouse for Birth Defects Surveillance and Research, Rome, Italy.

Different approaches have been used in case-control studies to estimate the maternal exposure to medications and birth defects risk. But the performance of each design using birth defect surveillance programs has not yet been reported. The aim was to evaluate the scope and limitations of three case-control approaches to assess the teratogenic risk on birth defects in mothers exposed to antiepileptic, insulin or acetaminophen.

Were studied 110,814 healthy newborns and 58,514 babies born alive with birth defects, registered by the ECLAMC during 1967-2008. Four controls by case and three different control groups were used: healthy newborns (HEALTHY), malformed newborns (SICK), and exposed cases only (ECO).
With HEALTHY design antiepileptics was associated with 14 birth defects; insulin with 11; and acetaminophen with 28. For SICK, antiepileptic showed association only with spina bifida; insulin with 4 birth defects, and no association was observed with acetaminophen. Using ECO, antiepileptic was associated with spina bifida only; insulin with 3 birth defects; and no association was found with acetaminophen.
The HEALTHY method showed high rate of false-positive results. The SICK and ECO methods does not estimated the true population risk except under certain assumptions. However, SICK is useful to determine the specificity of the teratogenic agent, whereas the ECO could be a good approach to estimate the etiological heterogeneity of the defect under study.

153
Complex Modalities Of Gene Action On Phenotypes With Post Reproductive Manifestation: The Case Of Genetic Trade-Off
Alexander Kulminski (1) Irina Culminskaya (1) Svetlana Ukraintseva (1) Konstantin Arbeev (1) Anatoli Yashin (1)
(1) Duke University

Progress in unraveling the genetic origins of healthy lifespan is tempered, in part, by a lack of replication of effects, which is often considered a signature of false positive findings. We convincingly demonstrate that the lack of genetic effects on phenotypes with post reproductive manifestation can be due to trade-offs in the gene action. We focus on the well-studied apolipoproetin E (*APOE*) e2/3/4 polymorphism and on lifespan and ages at onset of cardiovascular diseases (CVD) and cancer, using data on 3,924 participants of the Framingham Heart Study Offspring cohort. Kaplan-Meier estimates show that the e4 allele carriers live shorter lives than the non-e4 allele carriers (log rank=0.016). The adverse effect was attributed to the poor survival of the e4 homozygotes, whereas the effect of the common e3/4 genotype was insignificant. The e3/4 genotype, however, was antagonistically associated with onsets of CVD and cancer predisposing to an earlier onset of CVD and a later onset of cancer compared to the non-e4 allele genotypes. This trade-off explains the lack of a significant effect of the e3/4 genotype on survival; adjustment for it in the Cox regression model makes the detrimental effect of the e4 allele highly significant (p=0.002). This trade-off is likely caused by the lipid-metabolism-related (for CVD) and non-related (for cancer) mechanisms. An evolutionary rationale suggests that genetic trade-offs should not be an exception in studies of senescent traits.

154
Heritable Late Life Phenotypes and Inter-Chromosomal Linkage Disequilibrium in the Human Genome
Alexander Kulminski (1) Irina Culminskaya (1)
(1) Duke University

Studies of non-humans show that loci on non-homologous chromosomes can be in linkage disequilibrium (LD). Such LD is often observed in populations with different phenotypic structure. This work explores whether the phenomenon of inter-chromosomal LD can be associated with complex, polygenic phenotypes of late life and be caused by intrinsic bio-genetic mechanisms in the human genome. The analysis is based on an original two stage approach, which employs phenotype-based pre-selection of SNPs for the analyses of LD, and focuses on 9,274 genotyped participants of the Framingham Heart Study (FHS). The results document remarkably strong and extensive LD among SNPs at loci on multiple non-homologous chromosomes genotyped using two independent (Affymetrix 50K and 500K) arrays. The analyses provided compelling evidences that the observed inter-chromosomal LD was unlikely generated by population or family structure, mis-mapping, mis-genotyping, or any factor of stochastic origin. The analyses show that this LD is associated with complex heritable phenotypes of poor health. The inter-chromosomal LD was observed in parental and offspring generations of the FHS participants. These findings suggest that the observed inter-chromosomal LD can be caused by intrinsic biogenetic mechanisms which can be associated with favorable or unfavorable epistatic evolution. This phenomenon highlights a challenging role of genes and gene networks in regulating complex phenotypes of late life in humans.

155

Imputation-free Meta-Analysis with YAMAS

Tim Becker (1) Markus Leber (2) Dmitriy Drichel (1) Christine Herold (1) Manuel Mattheisen (2) Christian Meesters (2)
(1) German Center for Neurodegenerative Diseases, DZNE, Bonn, Germany(2) IMBIE, University of Bonn, Germany

A main purpose of imputation prior to MA is to unify the available marker panels and to avoid loss of SNPs that are not present in all studies. With YAMAS - Yet Another Meta-Analysis Software - we avoid such loss without the need to impute data. By using reference data (HAPMAP, 1000 Genomes) users are enabled to analyze all SNPs that are present in at least one study: LD-information is used to find substitute makers for those missing. For each SNP that is missing, the marker from the study with largest r^2, according to HAPMAP (1000 Genomes) data, with the missing marker is chosen as a "proxy SNP". Furthermore, based on the reference haplotype frequencies, "proxy alleles" of a SNP and its proxy-SNP are identified and MA of a particular SNP is done one the effect estimates of a SNP and its proxy-SNP.
MA with our algorithm is not quite as powerful as MA with imputation, due to a smaller effective marker panel. In general, however, the power loss is moderate and the YAMAS approach is more robust when reference and study sample are not ethnically matched. Thus, the proxy-algorithm can be recommended as a standard MA-approach for samples whose ethnicities are represented in either 1,000 Genomes or HAPMAP. We also present meta-analysis results for Parkinson Disease that identify a region that goes undetected with the imputing/MA-pipeline.
In summary, MA with YAMAS is an easy alternative, yielding ad hoc results and thereby giving an incentive to follow-up analysis.

156

A New R Package For The Calculation Of The Exact CDF Of Q And I^2 For Meta-Analyses

Michael Preuss (1) Andreas Ziegler (1)
(1) Institut fur Medizinische Biometrie und Statistik

The random effects (RE) model for meta-analyses is an essential part of the evidence-based medicine [1]. The main advantage of this model is that the between study variation $?^2$ is taken into account. This kind of heterogeneity can be observed because of sampling errors, the different definition of the dependent variable of the individual studies or due to population stratification [2]. Usually, the estimation of $?^2$ is carried out with the moment-based DerSimonian and Laird estimator [2]. The calculation of $?^2$ includes the Cochran Q-statisic, which is in particular an important component to calculate the heterogeneity estimator I^2 [3]. The well known advantages of I^2 are the independency of the number of studies as well as the intuitive interpretation towards the Cochran Q-statistic and $?^2$ [4], respectively. Biggerstaff and Jackson published an article in 2008 [5] in which the exact cumulative distribution function of the Cochran Q-statistic and the cumulative distribution function of I^2 are derived under the assumption of the RE model. This work considers the derivation of the exact I^2 under the RE assumption and updates simulations from Mittlbock and Heinzl [6]. In addition an own implemented R package is presented which provides a user-friendly way to calculate the exact I^2 measure as well as the exact RE estimator

157

A Simulation Pipeline For Genetic Disease Models

Hansjorg Baurecht (1) Thomas Augustin (2) Stefan Wagenpfeil (3) Konstantin Strauch (4) Paul A Scheet (5)
(1) Department of Dermatology, Allergology and Venerology, University Hospital Schleswig-Holstein, Kiel(2) Department of Statistics, Ludwig-Maximilians Universitat Munchen(3) Institute of Medical Statistics and Epidemiology, Technische Universitat Munchen(4) Institute of Genetic Epidemiology, Helmholtz Zentrum Munchen - German Research Center for Environmental Health, Neuherberg(5) Department of Epidemiology, University of Texas, MD Anderson Cancer Center, Houston

The advent of next-generation sequencing and generation of large data sets of population genetic data have allowed interrogations of association between phenotype and increasingly rare forms of genetic variation. To detect association with extremely rare variants some form of aggregation is necessary to combine the effects of variants from multiple affected individuals; for example, one way this is done is on the genic level. Evaluation of these procedures may involve simulated data, and these simulations typically assume independence among alleles at variant sites within genes. While this may be a reasonable approximation for some extremely rare variants, linkage disequilibrium at moderate frequencies may affect association detection. Here we aim to address this phenomenon in data simulation. We develop an R package to simulate case-control data using externally generated haplotype data, such as those from a coalescent simulator. We explore various ways to partition population attributable risk (PAR), which is used for specifying the disease model, among non-independent variant

sites to maintain an overall PAR. Our simulation pipeline provides a flexible tool for simulating a varierty of scenarios for validating statistical methods in the context of next-generation sequencing data, genomewide association studies or even candidate genes.

158
Massively Parallel Model Selection For Re-Sequencing Studies Using GPU Clusters
Gary K Chen (1)
(1) Division of Biostatistics, Department of Preventive Medicine, University of Southern California

It is common to conduct genomic studies across millions of SNPs genotyped/imputed on thousands of individuals. Ideally one would include all SNPs into a single model, but this is normally infeasible as 1) the number of covariates far exceed observations and 2) large matrix inversions are computationally intractable. The LASSOmodel overcomes these two obstacles and produces sparse models (i.e. some values of ? are exactly zero) enabling easy interpretation. The standard algorithm for fitting the LASSO is cyclic coordinate descent(CCD) which"cycles" through all variables, updating each ?one at time. CCD however, can be painfully slow for datasets with an extremely large number of variables with non-zero regression coefficients. An obvious strategy to improve speed is parallelize across samples at each variable update, but improvement soon hits an upper bound once the number of processor cores exceeds samples. A more scalable strategy is to parallelize across variables since graphics processing units (GPUs) expose thousands of cores on a single desktop machine. We present algorithms that scalesacross the number of variables. Specifically, we partition the large design matrix across multiple GPUsthat are synchronized across a cluster. We encounter speedups of over 300 fold using a single desktop machine with two nVidia Tesla C2050 GPU devices.

159
Forward-time Simulation of Linkage Disequilibrium across Two Populations using GenomeSIMLA
Carrie C. Buchanan (1) Eric S. Torstenson (1) William S. Bush (1) Marylyn D. Ritchie (1)
(1) Vanderbilt University

Simulation studies are often used to assess the power and Type I error rate of an analysis technique. Such studies have been critical for developing and applying sophisticated GWAS data analysis approaches. Additionally, the assembly of the human genome haplotype map has greatly influenced disease gene mapping and association studies in the last decade. Blocks of linkage disequilibrium are mainly defined by recombination hot spots which directly contribute to genetic diversity and evolutionary processes. We incorporate recombination frequencies calculated from International Haplotype Map Consortium and 1000 Genomes Project data to guide forward-time simulation using GenomeSIMLA in a population-specific manner. There is no single accepted measure for assessing how closely our simulated populations resemble natural data from 1000 Genomes Project. However, we adopted a simple statistical framework to measure frequency distribution of alleles and two measures of linkage disequilibrium for comparison between 1000 Genomes Project low coverage pilot data and the simulated data for CEU and YRI independently.

160
Multifactor Dimensionality Reduction 3.0: Open-Source Software for Systems Genetics
Peter Andrews (1) Jason H Moore (1)
(1) Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH USA

Multifactor dimensionality reduction (MDR) was designed as a nonparametric and genetic model-free approach to identifying gene-gene interactions in genetic studies of common human diseases. The kernel of the MDR algorithm uses constructive induction to combine two or more polymorphisms into a single predictor that captures interaction effects. This general approach has been validated in numerous simulation studies and has been applied to a wide-range of different human diseases. We describe here version 3.0 of the open-source MDR software that has been made freely available since February of 2005. Over this time period MDR has been downloaded more than 30,000 times. This new version of MDR has been significantly updated to allow users carry out a systems genetics analysis by inferring and characterizing large networks of gene-gene interactions. Here, the vertices of the network represent the single-nucleotide polymorphisms in the data while the edges or connections among vertices represent the synergistic gene-gene interactions that exceed some predefined threshold. We report the degree distribution, motifs, centrality, modularity and other measures of complexity. We also allow users to filter their list of SNPs according to the structure of the network. These new features in the MDR software move beyond models of several SNPs to the inference of large interacting networks of SNPs enabling a systems genetics approach to complex disease.

161
EMIM: Estimation Of Maternal, Imprinting And Interaction Effects Using Multinomial Modelling
Richard Howey (1) Heather J Cordell (1)
(1) Newcastle University

We present a new computer tool, EMIM, for the estimation of parental and child genetic effects based on genotype data from a variety of different child-parent configurations. An accompanying tool, PREMIM, allows the extraction of child-parent genotype data from standard pedigree data files for subsequent use with EMIM. The use of genotype data from the parents as well as from the subject in question allows the estimation of complex genetic effects such as maternal genotype effects, maternal-foetal interactions or parent-of-origin (imprinting) effects. Two existing popular approaches are to use genetic data from affected offspring and their mothers (case/mother duos) along with an appropriate control sample, or else to use genetic data from affected offspring and both parents (case/parent trios) without use of controls. EMIM, however, uses a multinomial modelling approach which allows the simultaneous use of case/mother duos and case/parent trios together with additional child-parent genotype data. The optimal

child-parent genotype data to use with EMIM can be extracted using PREMIM from any pedigree file. These data can then be used by EMIM to estimate genetic effect parameters that are of interest to the user, incorporating chosen assumptions such as Hardy-Weinberg equilibrium. Together EMIM and PREMIM provide easy to use command-line tools for the analysis of pedigree data, giving unbiased estimates of the relative risks of parental and child genetic effects.

162

Confounding between Genomic Imprinting and Sex-specific Recombination Frequencies: Evaluation of Properties of the MOD Score-based Imprinting Test Statistic MOBIT in a Linkage Simulation Study

Markus Brugger (1) Mathieu Lemire (2) Michael Knapp (3) Konstantin Strauch (1)
(1) Institute and Chair of Genetic Epidemiology, Ludwig-Maximilians-Universitat, Munich, and Helmholtz Zentrum Munchen, Germany (2) Ontario Institute for Cancer Research, Toronto, Canada, (3) Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany

Genomic imprinting is a parent-of-origin effect apparent in many human diseases. With imprinting, the penetrance depends on the sex of the parent who transmits a certain allele. In the context of linkage analysis, imprinting is confounded with differences in recombination fractions between males and females. It has been proposed to apply sex-specific marker maps or to perform multipoint analyses using dense marker frameworks in order to avoid confounding.
Using the GENEHUNTER-MODSCORE software, it is possible to test for imprinting and to account for sex-specific recombination fractions; thus confounding between the two can be investigated. In an extensive simulation study, nuclear families with two or four affected children segregating an additive trait were analyzed using the MOD score-based statistic MOBIT to test for imprinting. Different sex-specific map ratios and marker scenarios were considered for the evaluation of type I error and power. We show that analyses assuming a sex-averaged map, albeit confounded in twopoint scenarios, lead to a more powerful imprinting test than their sex-specific counterparts when an adequate correction for the sex differences in recombination fractions is applied. However, extreme map ratios and sparse marker frameworks in multipoint analyses again lead to confounded results. Therefore, we propose to perform multipoint analyses using densely spaced markers and sex-averaged maps to efficiently discover new imprinted loci.

163

ESPRESSO: Algorithm for realistic power analysis and sample size calculation

Amadou Gaye (1) Paul Burton (1)
(1) University of Leicester

Background: According to the WHO, complex chronic diseases will count for about 57% of deaths worldwide by 2020. One of the main limitations of genetic association studies, in the investigation of such diseases, is the lack of power. Conventional approaches to estimate the sample size required to achieve adequate power do not take into account some complex elements.
Methods: The trait of interested and the environmental determinants are binary or quantitative and the genetic determinants are modeled as binary or additive. Assessment errors in both exposure and outcome and the impact of unmeasured aetiological determinants as well as heterogeneity in disease are included as parameters in the simulation. The data is analysed by logistic regression.
Results: The algorithm allows for the integration of some elements, so far disregarded, in the design of large scale biobanks and association studies across many bioclinical scenarios. The tool is available from CRAN and an interactive version can be run from the P3G website.
Conclusion: Given the vast cost and effort that is needed to establish and maintain a contemporary biobank or design and carry out large studies, a small loss of power can impact substantially on the balance of costs and benefits.It is, hence, crucial to assess the impact, on power, of errors. ESPRESSO can also be used to answer relevant scientific questions that have been identified as being critical for biobanks and genetic association studies.

164

Easygwa-S - A General Software For Stratified Genome-Wide Association Meta-Analyses

T W Winkler (1) F Kronenberg (2) I M Heid (1)
(1) Department of Epidemiology and Preventive Medicine, Regensburg University Medical Center, Regensburg, Germany(2) Department of Medical Genetics, Molecular and Clinical Pharmacology; Innsbruck Medical University, Innsbruck, Austria

Genome-wide association meta-analyses (GWAMAs), pooling numerous results from genome-wide association studies (GWAs) have been highly successful in identifying genetic loci. To increase power for stratum-specific genetic loci, stratified GWAMAs - e.g. stratifying by sex or other risk factors for the disease under study - are being conducted. Software tools to specifically handle stratified analyses were lacking.
We develop a general programming tool, named "EasyGWA-S", specifically tailored to stratified GWAMAs. It offers a scripting interface and software modules, (1) to easily perform and compare stratified GWAMAs including testing for strata-effects, e.g. by generating specific graphical summaries, extracting associated genetic loci, and testing for difference between strata (GWAMA-Stratified); (2) to model the whole workflow of consortia performing GWAMAs including study-specific GWAs, quality control (QC) on the study level, GWAMA, QC on the meta-analysis level, SNP selection and graphical presentation of results (GWAMA-pipeline); (3) to facilitate general data-handling for high-dimensional genetic datasets with the added complexity through stratified analyses, e.g. by manipulating data and extracting information (GWAMA-management).
In contrast to other available software tools for genome-wide analyses, EasyGWA-S focuses on stratified analyses and provides functions for each of the steps from the study-specific level to the meta-analysis results.

165
Comparison Of Statistics Of Genetic Association Regarding The Discrimination Between Causal Variants And Linked Markers
Justo Lorenzo Bermejo (1)
(1) University Hospital Heidelberg, Institute of Medical Biometry and Informatics

Most genome wide association (GWA) studies still rely on the common-disease common-variant hypothesis: polymorphisms are genotyped and their association with disease is investigated. The identified, indirect associations are assumed to reflect a shared inheritance of genotyped markers and linked causal variants.
Probability values and Bayes factors are the most common summary of results from GWA studies. Alternative statistics of genetic association include pseudo R^2 measures, the area under the receiver operating characteristic curve, the population attributable fraction and the attributable familial relative risk. We have used simulation to compare these statistics regarding their ability to differentiate between markers and causal susceptibility variants. Theoretical results were illustrated by established causal associations with age-related macular degeneration and by imputation of genotypes based on HapMap for a case control study of breast cancer.
Preliminary results suggest that, under high penetrance, the representation of genetic association by familial relative risks instead of probability values and Bayes factors may facilitate the separation of causal variants. Moderate to low penetrance variants seem to be best discriminated by Bayes factors. The relevance of these findings in the context of association studies which take advantage of public data repositories (the International Hapmap and the 1000 Genomes projects) will be delineated in the meeting.

166
Aggregating Information Within Loci When Testing The Genome For Associations
Niall J Cardin (1) Joel A Mefford (1) John S Witte (1)
(1) University of California, San Francisco

New sequencing technologies provide an avenue for assessing the impact of rare and common variants on complex diseases. Several methods have been developed for evaluating rare variants. These methods typically use weighted aggregation to combine rare variants within or across genes. Arbitrary frequency thresholds below which to aggregate alleles need to be defined. Also, effect sizes for each aggregated variant are essentially assumed to be the same, or a function of minor allele frequency. When these assumptions do not hold, performance of the test will be adversely affected.
We propose a hierarchical model to detect the joint signal from rare and common variants within a genomic region, while properly accounting for linkage disequilibrium between variants. Our hierarchy controls the scale, rather than the mean of the odds ratio distribution, thus we allow for both causative and protective effects. We use a novel approach to assess the evidence for association in a region, using approximate Bayes Factors. We simulate data under a wide range of disease models, with effects at common and rare SNPs. Overall our method had more power than any we compared to; at the 0.001 significance level: Step-Up [1], CMC [2] and Madsen and Browning [3] had relative powers of 51%, 50% and 41% compared to ours.
[1] Hoffmann T.J. et al, PLoS one, 5(11):e13584, 2010
[2] Li B. and Leal S.M., Am J Hum Genet, 83(3): 31-321, 2008
[3] Madsen B.E. and Browning S.R., PLoS Genet, 5(2):e1000384, 2009

167
Single- And Multi-Locus Association Tests Incorporating Phenotype Heterogeneity
Hatef Darabi (1) Keith Humphreys (1)
(1) Karolinska Institutet

Taking disease subtypes into account when testing for an association between genetic factors and disease risk may help to identify specific aetiologic pathways. One way to assess a genetic association, whilst accounting for heterogeneity, is to use polytomous regression. This approach only allows heterogeneity to be considered in terms of a single categorical variable. We describe an alternative and novel test of association which incorporates multivariate measures of categorical and continuous heterogeneity. We describe both a single-SNP and a global multi-SNP test and use simulated data to demonstrate the power of the tests when genetic effects differ across disease subtypes. Applying the tests to the study of genetic variation in the oestrogen metabolic pathway and its association with breast cancer risk and prognosticators strengthened our understanding that the modulation of aromatase activity can influence the occurrence of tumours, and their grade and size, in postmenopausal women.

168
A Nonparametric Approach to Population Based Association Tests
Sharon M Lutz (1) Wai-Ki Yip (2) John Hokanson (1) Nan Laird (2) Christph Lange (2)
(1) University of Colorado at Denver, (2) Harvard School of Public Health

In population-based genetic association studies, the standard approach is to modelthe phenotype of interest as a function of the o?spring genotype. We propose an alternative approach based on conditional score-tests that treats the genetic informationas the random variable and conditions upon the phenotypic information. The flexiblestructure of the approach enables the straight-forward application to standard andcomplex phenotypes. By treating the phenotype data as deterministic, the validity ofthe approach does not depend on the correctness of any assumptions about the phenotype. This makes the approach especially suitable for complex phenotypic modelsand the analysis of secondary phenotypes in studies that applied ascertainment conditions to the primary phenotype. If both phenotypes are correlated, the ascertainmentconditions can cause a perturbation of the distribution of the secondary phenotypes.Based on theoretical considerations and on simulation studies, we show that our approach is robust against mis-speci?cation of phenotype assumptions and, at the sametime, achieves the same or higher power level as standard genetic association testsfor population-based designs.

169

Bayesian Approaches To Identifying Susceptibility Loci: A Simulation Study

Katie M O'Brien (1) Robert C Millikan (1) Steven R Cole (1)
(1) Gillings School of Global Public Health, University of North Carolina at Chapel Hill

Identification of genetic risk factors for complex diseases is limited by the almost universal use of frequentist statistical methods and multiple comparisons adjustments, as such approaches are underpowered and do not make use of all available information. By simulating a case-control study with 4000 observations and 50 single nucleotide polymorphisms (SNPs) with known effects, we explored the mean squared error (MSE), a measure of both accuracy and precision, for several alternative statistical methods, including hierarchical regression modeling and Bayesian analyses. The simulated SNPs had minor allele frequencies ranging from 4- 41%. We imitated haplotype blocks by inducing correlations between select adjacent SNPs. The simulation included four causal SNPs, with odds ratios ranging from 1.1-1.4. When all 50 SNPs were modeled simultaneously using frequentist methods and a log-additive genetic model, the average MSE for the four causal SNPs was 0.0127. A full Bayesian model with weakly positive priors resulted in an average MSE of 0.0123. Similar analyses on a subset of 600 observations showed further discrepancies between models, with average MSEs of 0.0778 and 0.0553 for the frequentist and full Bayes models, respectively. Hierarchical regression models also outperformed frequentist methods. Simulation studies such as this quantify the strengths and limitations of various Bayesian methods and provide guidance for their application in genetic epidemiology.

170

Genotype Imputation And Association Testing Using Data From The 1000 Genomes Project

Jian'an Luan (1) Jing-Hua Zhao (1) Daniel Barnes (2) Ruth JF Loos (1)
(1) MRC Epidemiology Unit, Cambridge, UK(2) Cancer Research UK Genetic Epidemiology Unit, Cambridge, UK

Recently available data from the 1000 Genomes Project has led to the prospect of approximately 11 million SNPs to be imputed for each individual in genome-wide association studies (GWAS). Compared to HapMap based imputations, the imputation based on the 1000 Genomes Project is still under extensive review. Using the imputation software IMPUTE, we have made a workflow of 1000 Genomes imputation considering things such as the number of SNPs available and the centromere of a chromosome. With respect to GWAS analysis, we have developed a simple yet novel procedure using Stata, based on GWAS analysis software SNPTEST. This allows a range of options for analysis including variable transformation, subgroup analyses and covariate adjustment. The use of this procedure does not require a great deal of statistical, computing or data management experience. We also make use of Linux clusters whenever available. As the interface is generic, it is straightforward to apply the same procedure for both imputed data based on HapMap and 1000 Genomes. The procedures have been used extensively in contributions to consortia on a number of traits. Furthermore, we have used the rich facility in Stata to compare various procedures for accounting for genotypic uncertainty including SNP dosage and probability weighting. We expect our work will be of interest to colleagues working on GWAS and believe it deserves more attention.

171

Imputation Accuracy In The MHC Region Based On 1000 Genomes Data

Nicole M Roslin (1) Andrew D Paterson (2) Dafna D Gladman (3) Proton Rahman (4) Mathieu Lemire (5)
(1) The Hosptial for Sick Children(2) The Hospital for Sick Children and University of Toronto(3) Toronto Western Research Institute and University of Toronto(4) Memorial University of Newfoundland(5) Ontario Institute for Cancer Research

Imputation is a valuable tool to estimate genotypes for SNPs which are not present on genome-wide chips. The accuracy of imputation methods relies on exploiting linkage disequilibrium (LD). It is unclear how well these methods work in the major histocompatibility (MHC) region, where LD patterns are complex and highly variable. We present measures of concordance between genotypes imputed in the MHC region and true genotypes, using 1000 Genomes data as a reference set. 192 white cases of psoriatic arthritis and 247 controls from Newfoundland and Labrador (NL), Canada, were genotyped for 2136 SNPs on Illumina's MHC Mapping and MHC Exon-Centric Panels. Genotypes were removed for 1665 SNPs; the remaining 451 SNPs approximate the set of SNPs which would be available in the MHC region on the Affymetrix Human 6.0 genome-wide chip. Genotypes from 283 samples of European descent from the 1000 Genomes Project (20100804 release) were used as a reference set to impute genotypes for the 1665 SNPs in the NL sample, using MaCH. Agreement between the true and imputed genotypes was quantified by Cohen's kappa (k). In general, the agreement was excellent: median(k)=0.97. 1631 (98.1%) SNPs had k>0.6, which can be interpreted as substantial to almost perfect agreement. Imputation was worse in the region surrounding HLA-B, a locus strongly associated with psoriatic arthritis, and one of the most polymorphic genes in the human genome.

172

A Computationally Fast Bayesian Semi-Parametric Algorithm For Inferring Population Structure And Adjusting For Case-Control Association Tests

Arunabha Majumdar (1) Sourabh Bhattacharya (1) Analabha Basu (2) Saurabh Ghosh (1)
(1) Indian Statistical Institute(2) National Institute of Biomedical Genomics

Genome-wide case-control association studies have been successful in identifying novel variants involved in complex disorders. However, the problem of population stratification remains a major limitation of such studies. While methods have been developed (e.g., Genomic Controls, STRUCTURE along with STRAT, EIGENSTRAT) to infer on population structure and correct for stratification, the estimation of the number of underlying subpopulations (K), which is of additional interest from an evolutionary perspective, has not been adequately addressed, except in STRUCTURE. An ad hoc approach of Bayesian deviance

adopted by STRUCTURE tends to overestimate *K* and may lead to reduced power in detecting association. We have developed a Bayesian semi-parametric model in the lines of Bhattacharya (2008) to estimate population structure which has been complemented by a summarization of the clustering data generated by the MCMC based on a "Central Clustering" method [Mukhopadhyay et al. (2011)]. Our approach has several advantages over STRUCTURE, the most prominent being a substantial reduction in computational time. Based on extensive simulations under a set-up of no admixture and unlinked set of markers, we find that our method provides more accurate estimates of *K* and the test for association using STRAT is marginally more powerful compared to STRUCTURE. We analyzed the Human Genome Diversity Panel data using our model and obtained very good clustering of the individuals in the panel.

173

Evaluation Of Power For Linkage Disequilibrium Mapping

Ryo Yamada (1) Chikashi Terao (2) Takahisa Kawaguchi (2) Maiko Narahara (1)
(1) Statistical Genetics, Medicine, Kyoto University (2) Human Disease Genomics, Medicine, Kyoto University

Genetic factors of various diseases have been studied with SNP chips and next generation sequencing technology is now being utilized for the purpose. Although polymorphic and potentially affecting markers distribute along the chromosomes with relatively even density in the studies with SNP chips and sequence data, the number of markers and their density as well as linkage disequilibrium pattern in each gene or locus varies substantially.
In this poster we propose a novel method which handles multiple marker tests for a set of categorical phenotypes in the context of geometric statistics so that we are able to estimate power when multiple tests are applied to identify association between a locus with multiple markers and a phenotype that might have multiple surrogating phenotype-related criteria. First of all, we introduced a method to define multiple tests in a higher dimensional space. Secondly, we studied power when one surrogating marker was applied with our geometric approach. Thirdly, we designed models of condition of multiple tests and evaluated the effect of relation between the power and pattern of linkage disequilibrium. The result suggested some loci with higher variation in linkage disequilibrium might have several times higher power than less variable loci.

174

Does Size Always Matter? A Simulation Study On The Impact Of Slightly Altered True Genetic Models

Carolin Putter (1) Karl-Heinz Jockel (1) Heinz-Erich Wichmann (2) Andre Scherag (1)
(1) Institute for Medical Informatics, Biometry and Epidemiology, University Duisburg-Essen, Essen, Germany(2) Helmholtz Zentrum Munchen, German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany

Genome-wide association studies revealed robust associations between single nucleotide polymorphisms and complex traits. As the proportion of the explained phenotypic variance is still limited for most of the traits larger and larger meta-analysis are being conducted to detect additional association signals. Here we investigate the impact of the study design and the underlying assumption about the true genetic effect in a bimodal mixture situation on the power to detect additional association findings. We performed simulations of quantitative phenotypes analysed by standard linear regression (n=130,000) and artificially dichotomized case-control data sets (5,000 pairs) from the extremes of the quantitative trait analysed by standard logistic regression. Using linear regression, markers with an effect in the extremes of the traits were almost undetectable even with steadily growing sample sizes. In this situation, analysing extremes by a case-control design of a much smaller sample size had better power to detect the associated marker. Our results indicate a) that it might be worthwhile to re-analyse dichotomized versions of the available meta analysis data sets to detect new loci while it b) offers an explanation for discrepant findings pertaining to quantitative phenotypes analysed by linear regression as compared to case-control approach analysing selected samples. A real-data example for body-mass-index in n=16,463 is provided to support the bimodal mixture assumption.

175

Performance Of Different Balancing Score Methods In Case-Control Genetic Association Studies

Amina Barhdadi (1) Marie-Pierre Dube (2)
(1) Montreal Heart Institute, Research Centre (2) Faculte de medecine, Universite de Montreal

The propensity score is a popular method to control for confounding in prospective studies. Its counterpart for the case-control design is the stratification score which is a retrospective balancing score. Here, we compare 3 balancing score methods for genetic studies: stratification (SB), matching (MB) and covariate adjustment (CAB) all applied to the balancing score. We simulated 1000 datasets of 4000 subjects with 3 continuous covariates and 3 binary covariates that were imbalanced between cases and controls. Genotypes were simulated for 10 SNPs of varying minor allele frequency (MAF) and odds ratios (OR). One SNP was simulated to be associated with one of the 3 binary covariates. The 3 methods were evaluated for their ability to identify the simulated genetic effect based on the mean relative bias. All methods have 100% power to detect the causal SNP with a MAF of .30. Type I error rate was .053, .049, .053 for CAB, SB and MB and the relative bias mean was .08±.07, .08±.06, .12±.09. With a MAF of .05, the power to detect the associated SNP was 49.9%, 52.8%, 16.4%, type I error rate was .014, .024, .010 and the relative bias mean was .58 ± .83, .58+.80, .62±.69 for CAB, SB and MB. Type I error rate at the simulated confounding SNP was .047, .045, and .055 for CAB, SB and MB.
The SB method performed better than other methods. Further investigations are underway to explore the performance of each method under strong confounding effect.

176

A Sequential Combined P-Value Test For Multiple Hypothesis Testing And Its Application In Significance Analysis In Genomic Studies

Huann-Sheng Chen (1) Shunpu Zhang (2) Ruth Pfeiffer (3)
(1) Division of Cancer Control and Population Sciences, National Cancer Institute(2) Department of Statistics, University of Nebraska(3) Division of Cancer Epidemiology and Genetics, National Cancer Institute

Two combined *p*-value tests (the truncated product method (TPM) of Zaykin et al. (2002) and the rank truncated product (RTP) test of Dudbridge and Koeleman (2003) have been proposed, and are widely used in genome-wide association analysis. Instead of combining all the *p*-values, these tests combine only a subset of the *p*-values. It is claimed that these two tests are capable of providing a list of hypotheses in which there is at least one hypothesis that is truly false. An obvious drawback of the TPM and RTP tests is that the size of the subset of the *p*-values being included in the test statistics needs to be pre-selected subjectively. We propose a new step-up combined *p*-value test which does not require selecting the *p*-values for the test statistic, has weak control of the family-wise error rate (FWER), and can provide individual statements on hypotheses when the global null hypothesis is rejected. In simulations we demonstrate that the proposed test enjoys both the advantages of the combined *p*-value test and the classical multiple tests, and has the most robust performance to departures from the independent assumption among the existing combined *p*-value tests discussed in this paper. More importantly, the proposed test has significantly higher power than the TPM and RTP tests. Finally, we apply the method to a real genomic data set and the results show that the proposed test correctly identifies the association between the test region and the disease status.

177

Analysis Of Microsatellite Markers: A Comparison Of Four Different Ways To Evaluate Association With Binary Outcomes

Anja Rudolph (1) Hong Shi (2) Rebecca Hein (1) Asta Forsti (2) Juan Sainz (2) Michael Hoffmeister (3) Kari Hemminki (2) Hermann Brenner (3) Jenny Chang-Claude (1)
(1) Division of Cancer Epidemiology, German Cancer Research Center(2) Division of Molecular Genetic Epidemiology, German Cancer Research Center(3) Division of Clinical Epidemiology and Aging Research, German Cancer Research Center

Microsatellites (MS) are characterized by being poly-allelic, depending on the different numbers of repeats that occur. In the literature, different ways of evaluating MS-disease associations can be found: (i) a global test considering all alleles at once, (ii) a permutation test, using the minimal overall allelic *p*-value, (iii) dichotomization of alleles, and (iv) a reference genotype approach, based on the most common allele.
The power of the global test is considered to be weak, since the simultaneous consideration of all alleles leads to a test with many degrees of freedom. The permutation test is preferable if single alleles show an association. If short alleles are differently associated than long alleles, a dichotomization can be useful, while a non-arbitrary cut-off should be used (e.g. median allele length). When applying the reference genotype approach, heterogeneous genotypes are taken together into one category "H". This does not allow for trend tests and any association with the "H"-category is non-informative.
Choosing a preferable way to evaluate associations of MS with binary outcomes is context dependent. We therefore compared the four methods by exemplary analysing a CA repeat in the *ESR2* gene in a German population-based case-control study on colorectal cancer (DACHS). The CA repeat was genotyped and analyzed in 1798 cases and 1810 controls. The study illustrates each method's strengths and weaknesses and gives recommendations for different situations.

178

Admixture Mapping By Graphical Modeling

Haley J Abel (1) Alun Thomas (1)
(1) Division of Genetic Epidemiology, University of Utah

We develop a method to apply graphical modeling to genome-wide admixture mapping. Our approach accounts for linkage disequilibrium and gives sensitive detection of ancestral origins along the haplotypes of admixed individuals. We first use Markov chain Monte Carlo to estimate, given a reference set of genotyped individuals from known populations, a graphical model representing the joint distribution of alleles and ancestry at all loci. This model is then extended to form a new graphical model in which ethnic origin varies along the chromosome according to a first-order Markov process. The genotypes of unphased, admixed individuals can then be phased, the haplotypes segmented according to ancestral population, and missing data imputed. For a test population of mixed European (CEU) and Yoruban (YRI) descent, simulated based on 20,000 loci from HapMap individuals, the ethnicity estimates at all loci correlate closely with the true simulated ethnic origins (R^2=0.95). For comparison, we obtained R^2=0.98 using the Hapmix software. Our approach has the advantage that it can be used for admixed individuals from 3 or more ancestral populations, albeit with reduced accuracy. Furthermore, our method allows for one-time estimation and storage of the reference graphical model, so that subsequent phasing, imputation, and admixture mapping on test populations can be performed rapidly and without re-estimating the model.

179

Evolutionary Genetics of Myoclonin1/EFHC1, a gene for Juvenile Myoclonic Epilepsy (JME).

Julia N Bailey (1) Reyna M Duron (2) Myabi Tanaka (2) Dongshei Bai (2) Antonio V Delgado-Escueta (2)
(1) Department of Epidemiology, UCLA; West/LA VA Epilepsy Center of Excellence, Epilepsy Genetics/ Genomics Laboratories, Los Angeles (2) West/LA VA Epilepsy Center of Excellence, Epilepsy Genetics/ Genomics Laboratories, Los Angeles, California, USA

Juvenile Myoclinic Epilepsy (JME) is one of the more common forms of epilepsy, accounting for approximately 25% of all idiopathic generalized epilepsies. JME is genetically heterogeneous. Several genes have been found for JME, most of them are rare and only prevalent in one or two large families. The notable exception is Myoclonin1/EFHC1, which our group localized in JME cases in a Hispanic cohort from Los Angeles, Mexico and Honduras.
Mutation: The biochemistry demonstrates JME can be due to mutations in EFHC1; a developmental gene involved in

cell division, apoptosis, neuroblast migration and synaptogenesis. We had originally shown that nine percent of JME singletons/sporadic and families from epilepsy clinics in the GENetics of the EpilepSieS (GENESS) consortium had heterozygous missense, nonsense and deletion/frameshifts mutations in Myoclonin1/EFHC1. To date, 17 mutations in EFHC1 have been identified and validated by knock out/knock in mouse models.
Migration: Similar and new heterozygous mutations in Myoclonin1/EFHC1 have also been described in Japan, Chile, Austria, Italy and Brazil. These mutations were not found in populations from Tennessee or the Neatherlands that were screened, and have not been reported in any of the JME genetics cohorts from the European or Australian consortiums. This pattern is consistent with both the gene migrating over the barring straight from Asian and coming across the water with the European Conquistadors.

180
Signatures Of Recent Positive Selection At The VKORC1 Gene Locus
Blandine Patillon (1) Pierre Luisi (2) Sabbagh Audrey (3) Genin Emmanuelle (4)
(1) Inserm UMRS-946, Univ Paris Diderot, Paris, France-Univ Paris Sud, Kremlin-Bicetre and Univ Paris Sud, Kremlin-Bicetre, France.(2) IBE-Institute of Evolutionary Biology, CEXS-UPF-PRBB, Barcelona, Spain.(3) Universite Paris Descartes, UMR 216, Paris, France.(4) Inserm UMRS-946, Genetic Variability and Human Diseases, Institut Universitaire d'Hematologie, Univ Paris Diderot, Paris, France

Different tests have been developed to detect signatures of natural positive selection on the human genome using SNP data. In this study, we will discuss these different tests using data from the Human Genome Diversity Panel in the *VKORC1* gene region. *VKORC1* is a major gene influencing individual response to oral anti-coagulants (OAC), such as warfarin. More specifically, the functional g.-1639G>A polymorphism (rs9923231) explains by itself one third of the wide interindividual variability in dose requirement. We found that this variant presents an unusual pattern of genetic differentiation at the worldwide level and occurs on a unique haplotype that reaches near fixation in all the East-Asian populations investigated while being rare in Africans. We applied both haplotype-based and allele frequency-based tests that all detected a strong selective sweep surrounding *VKORC1* in East-Asian populations. The selective phenomenon identified in East-Asia explains in part the important differences in OAC doses between human populations, which range from 3.5 mg in East-Asians to 6 mg in Africans for the warfarin daily-dose. Several genes involved in drug metabolism, such as *VKORC1*, show some important ethnic differences in allele frequencies that might be due to adaptive events. Documenting these differences and studying them in relation with natural selection through population genetic studies could help us better understand the genetic basis of drug response.

181
Genetic Structure And Admixture Of The Seychelles' Population
Pierre Bady (1) Georg Ehret (2) Conrad Shamlaye (3) Michel Burnier (4) Francois Mach (5) Fred Paccaud (6) Mauro Delorenzi (1) Pascal Bovet (6) Murielle Bochud (6)
(1) University Hospital (CHUV) and University of Lausanne, Switzerland(2) University of Lausanne and University of Geneva, Switzerland(3) Ministry of Health, Seychelles(4) Nephrology Division, University Hospital (CHUV) and University of Lausanne, Switzerland(5) Cardiology Division, University Hospital (HUG) and University of Geneva, Switzerland(6) Institute of Social and Preventive Medicine, University Hospital (CHUV) and University of Lausanne, Switzerland

Inhabited only since the 1770s, the population of the Seychelles, Indian Ocean, is composed of persons who came from East Africa (a large majority) and smaller numbers from Europe, India and China. We studied the genetic structure of 1034 people from the Seychelles (population-based survey, 2004) using 196'725 markers genotyped with the Illumina CardioMetaboChip. We used the ADMIXTURE program (43'598 markers) to estimate participants' ancestry as well as principal component and cluster analysis (22'317 markers) to compare the genetic structure of the Seychelles' population with that of populations from the "HapMap" project. The Seychelles' population is composed of three main sub-populations. About 69% of individuals clustered closer to African Americans (ASW) than to Africans from Kenya or Nigeria (MKK, LWK and YRI). The second cluster (about 21%) was closest to Europeans (CEU and TSI). The third, more widespread, cluster (about 10%) spread toward Asians (CHB=Han Chinese, JPN=Japanese and GIH=Gujarati Indians). Seychelles participants belonging to the "African" cluster had 71%, 14% and 15% African, European and Asian ancestry, respectively, and had a larger proportion of short haplotypes than Europeans or Asians. In conclusion, the genetic structure of the Seychelles' population is consistent with the demographic history of this country and will allow taking population substructure into account for future genetic association studies.

182
Modeling The History Of A Sample Of Genotypes
Fabrice Larribe (1)
(1) UQAM

To map a disease gene, pairwise statistics methods are commonly used, but these methods do not use all the available genetic information. To take into account the dependance between genetic markers, we have to use haplotypes. To make the most of the haplotype information, and to take into account the dependance between haplotypes, we have to model the ancestry of these haplotypes. We show how the unknown history of a set of haplotypes can improve the estimation of the location of a disease gene. We present the different models that have been used to describe the genealogy of a set a haplotypes, and we argue that the coalescent process with recombination should be an ideal model to work with. We present the challenges to use such a complex model (inference, size of the genetic data), but also the advantages to use a theoretically well founded model, with a lot of natural possible extensions to the basic coalescent with recombination (selection, ascertainment bias). We show how we can accommodate for complex factors such as incomplete penetrance, phenocopy, genotypes and several type of genetic markers. We present some examples of inference of the location of a disease gene using such a

model, and compare the estimation with simple methods commonly used.

183

A National Resource Combining Genetic And Phenotypic Data: The VA Genealogy Project

Lisa Cannon-Albright (1) Sue Dintelman (2) Tim Maness (2) Alun Thomas (1) Lawrence Meyer (3)
(1) University of Utah School of Medicine(2) Pleiades Data Systems(3) George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah

The Utah Population Data Base and the Icelandic genealogy, combining genealogic and phenotypic data, have proven valuable to genetic studies. We are creating a resource linking a genealogy of the United States to the Veteran's Administration population of 25 million retired service people.
We have built the genealogy of Utah (UT) and Massachusetts (MA) using genealogical data from public sources, extracted in common published formats. We used 13,000 data sources to build the current UT and MA genealogy. The genealogy currently includes 20 million individuals born after 1500. We have begun record linking to 2M VA patients using the VA in MA or UT. We have conservatively linked 30,000 (2%) of these patients to the genealogy.
An analysis of relationships for the 30,000 linked patients identified hundreds of clusters of related VA patients. Multiple clusters of related VA patients with cancer have been identified; an example colorectal cancer includes 8 patients with a common ancestor 9 generations back. We analyze the relationships of patients with phenotypes of interest to define relatedness and relative risks. The recent initiation of the VA Million Veterans Project, which will sample DNA for 1 million Veterans, will result in the largest genetic repository in the world. When combined with this VA genealogy, the resource will have unlimited potential for study of the genetics of health and disease.

184

Regression models for DNA-mixtures

Thore Egeland (1)
(1) Norwegian University of Life Sciences

In several applications there is a need to determine whether specific individuals have contributed to a DNA-mixture. Previously this has been consider mostly relevant to forensic problems like rape cases. In such cases the mixture may consist of DNA from a victim and several men. Determining whether a specified individual has contributed to the mixture or not is obviously of direct relevance for the court. Recently the relevance of mixture problems for GWA (Genome Wide Association) studies have been demonstrated. In this case the mixture arises from DNA pooling. Pooling allows allele frequencies in groups of individuals to be estimated based on fewer PCR reactions and genotyping assays than are used when genotyping individuals. Reports showing that the individuals contributing to a pool can be identified have been disturbing and led to the precautionary removal of large amounts of summary data from public access. Regression models designed to determine whether specific individuals contribute to a DNA mixture will be presented and illustrated based on simulations and real data.

185

Preprocessing Illumina DNA Methylation BeadArrays

Kimberly D Siegmund (1) Timothy Triche Jr. (1)
(1) University of Southern California

Variation in the epigenome, the distribution of DNA-related modifications and structural features that inform the packaging of the DNA, can confer a host of specialized functions to different cells with the same genome. DNA methylation is the most commonly studied epigenetic mark; its importance well-established in human development and disease. Presently, DNA methylation microarrays provide the most cost-effective means of high-throughput analysis. As with other types of microarrays that measure gene expression, genotype, or copy number variation, technical artifacts are a concern. Illumina's Bead Array technology for gene expression now has its own preferred data preprocessing pipeline, taking advantage of the hundreds of control probes for background correction and sample normalization. We describe the Illumina BeadArray technology for DNA methylation analysis, and present a novel Gamma-Gamma convolution model to correct for bias due to background fluorescence. Using data generated on the HumanMethylation27 BeadArray, we find that the Gamma-Gamma convolution model reduces bias in signal intensity and variation in probe signal across replicate samples better than competing approaches. Adaptations of the method for the recently launched HumanMethylation450 array will be discussed.

186

Automated Investigation Of Genotype Calling Using Angles And Tests For Unimodality

Arne Schillert (1) Michael Pfutzenreuter (1) Andreas Ziegler (1)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck

Incorrect genotype assignments during genotype can invalidate genome-wide association studies. Recently, we developed an algorithm to automatically detect failed genotype clusterings [1]. To this end, we plotted the contrast of the signal intensities against its sum and counted the number of points which were too close to a cluster for a different genotype. In this work, we substantially improve the algorithm by adding two features. First, we compute the angle of the first principal component because failed clusterings often result in tilted clusters. Second, we investigated tests for unimodality [2] to identify genotype clusters which actually consist of disjunct sub-groups as this hints at failed clusterings as well. To enable the automated analysis of all genotypes of a recent microarray experiment we transform the data massively. In particular, we convert the usually very large intensity file into many smaller chunks in DatABEL's data format [3]. The genotypes are stored as GenABEL objects. This allows fast access to different sets of SNPs and rapid evaluation. When run on a computer cluster with 50 nodes, the analysis of 3.000 individuals with 700.000 SNPs can be done over the weekend.
[1] Schillert et al. 2009, BMC Proc 3:S58
[2] Larkin 1979, Behav Res Meth Instr 11:467-468
[3] Aulchenko 2007, Bioinformatics 23:1294-1296