ABSTRACTS FROM THE

SEVENTEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGICAL SOCIETY

1

Using in silico priors for pathway modeling

D.V. Conti (1), J.C. Figueiredo (1), A.J. Levine (1), C.M. Ulrich (2), J.N. Poynter (1), H.F. Nijhout (3), M. Reed (4), Y. Zheng (2), R.W. Haile (1)
(1) Dept. of Prev. Med., USC; (2) Cancer Prev. Research Program, Fred Hutchinson Cancer Research Center; (3) Dept. of Biology and (4) Dept. of Mathematics, Duke University

Folate-mediated one-carbon metabolism is thought to play an important role in colorectal cancer. However, results from conventional analyses have been equivocal. We can improve upon these analyses by incorporating into the statistical analysis the underlying biologic knowledge. We propose to use a hierarchical modeling framework to investigate genetic variants and compare these findings to a stochastic search variable selection algorithm. The hierarchical modeling approach uses a first-stage conditional logistic regression to independently estimate the association between each polymorphism and risk. In contrast, the model selection approach allows for joint modeling by searching over the entire model space to identify single polymorphisms and combinations that may be influential. Common to both approaches is a second-stage model regressing the first-stage estimates on prior covariates specifying gene-specific characteristics. To demonstrate these approaches, we analyze 524 SNPs genotyped on 28 genes involved in the folate and DNA repair pathways in the Colon Cancer Family Registry. In silico priors are obtained using a mathematical model of one-carbon metabolism that uses folate enzyme kinetics and regulatory mechanisms to simulate the impact of genetic and nutritional variation on the entire pathway. We present results from the applied analysis and discuss potential advantages and limitations to each approach.

2

The use of genome-wide eQTL associations to identify novel genetic pathways involved in complex traits

J.L. Min (1), J.M. Taylor (1), J.B. Richards (3), T. Watts (2), J. Broxholme (1), F. Pettersson (1), K.R. Ahmadi (3), I. Ragoussis (2), A.P. Morris (1), T.D. Spector (3), L.R. Cardon (1), K.T. Zondervan (1)
(1) Bioinformatics & Statistical Genetics, WTCHG, University of Oxford, UK, (2) Genomics Laboratory, WTCHG, University of Oxford, UK, (3) Twin Research Unit, King's College London, UK

Despite recent successes of genome-wide association studies in complex traits, many associations between clinical phenotypes and genetic variants will remain difficult to uncover because of phenotypic heterogeneity. The use of downstream biological phenotypes may provide a more powerful approach. Gene expression levels are highly variable and heritable, and are known to be strongly associated with genetic variants. This study investigates the association between 44 quantitative metabolic phenotypes and 19,828 gene expression levels in 299 twins, followed up by targeted SNP association analysis in a replication set of 2277 female twins. Expression profiling was conducted in lymphoblastoid cell lines from 154 female twin pairs from the St. Thomas' UK adult twin registry, and 57 unrelated CEU HapMap individuals, using the Illumina Sentrix Human-6 version 2 BeadChip. We found 956 probes correlating with one or multiple traits. Genome wide association analysis between 900,651 non-redundant SNPs and these probes in HapMap individuals identified 5 probes with association signals in cis and 314 probes in trans. Replication of these signals in other eQTL studies was obtained for 20% of cis signals and 8% of the trans signals. SNPs associated with these probes are currently being genotyped in 2277 twins.

3

Hierarchical Modeling of Pathway-Based Candidate Genes and Gene-Environment Interactions

H.E. Volk (1), F. Gilliland (1), D. Diaz-Sanchez (3), D.V. Conti (1,2)
(1) Dept. of Preventive Medicine, (2) Zilkha Neurogenetic Institute, Keck School of Medicine, Univ. of Southern California, USA; (3) Environmental Protection Agency, USA

We evaluate the ability of the hierarchical model to identify common pathway and gene-family effects using data from a single-blind, randomized placebo controlled study of environmental exposures. Multiple candidates were evaluated for effect on biomarker response upon exposure to allergen, diesel exhaust, or both conditions. Prior covariates were incorporated into the higher levels of the model based on pathway and gene-family to summarize these effects. Specifically, 20 phenotypes ranging from IL-8 to INF-γ are categorized into inflammation and allergic response phenotypes. Likewise, genes are categorized into oxidative stress and inflammatory related genes. Interactions with the three exposures and summary estimates of interactions were also obtained. We were able to identify pathway and gene-family specific responses to environmental exposures and summarize main and interaction effects. Our results show specificity of immune response for the allergic and inflammatory measures for allergen and diesel exhaust exposures, respectively. Differences in model fit and comparability of information across models will be discussed. These methods may be useful in evaluation of large-scale

Published online in Wiley Interscience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20367

pathway-based candidate gene studies and demonstrate the utility of incorporating prior covariates into hierarchal analyses.

4

SNPs to pathways - making biological sense of GWA results

P. Holmans
Cardiff University, UK

Genome-wide association (GWA) studies are a promising way of detecting associations between SNPs and complex traits. However, relatively few SNPs have p-values sufficiently small to give conclusive evidence of association. Conversely, there are usually several hundred SNPs with moderately significant p-values (p~10-3-10-4). These will likely contain several false-positives, but may also contain genuine effects of small magnitude. The presence of a greater than expected number of associated SNPs in genes of similar biological function gives a degree of confidence that the associations are genuine (even if none is individually very significant) as well as giving an insight into the biological processes underlying the disease.
We present a method where a list of significantly associated genes is generated (genes containing SNPs with a p-value for association less than a pre-defined threshold). Gene Ontology (GO) categories are tested for over-representation on this list (relative to the rest of the genome) allowing for varying numbers of SNPs per gene. Correction for testing multiple non-independent GO categories is performed using bootstrapping. The method is demonstrated using datasets from the Wellcome Trust Case Control Consortium (WTCCC) study.

5

Candidate Epistasis: Generating putative gene-gene interactions for an analysis of a whole-genome association study of multiple sclerosis

W.S. Bush, S.M. Dudek, J.L. Haines, and M.D. Ritchie
Vanderbilt University, Center for Human Genetics Research

Examining epistasis in WGA studies is a difficult challenge. Exhaustively enumerating every possible SNP combination is computationally and statistically prohibitive. Non-exhaustive strategies include evaluating epistasis in a set of SNPs that have statistically significant main effects, and evaluating epistasis between SNPs in genes where there is evidence of biological interaction. In this work, we illustrate a simple bioinformatics approach for generating and ranking multi-SNP models of multiple sclerosis (MS) susceptibility, using data sources implying biological interaction of molecules, sources implying gene relationship to disease, and literature-based information in an approach similar to "genomic convergence". We constructed putative gene-gene interactions based on the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, the Database of Interacting Proteins (DIP), the Protein Families Database (PFAM), the Gene Ontology (GO), and Netpath. We also constructed putative disease-associated genes, using the Genetic Association Database (GAD), previous linkage screens of MS families, three studies of MS gene expression, and other candidates from literature-based sources. Using these sets, we generated a list of multi-locus models for evaluation in WGA data indexed by the number of data sources that support the model. Applying this information to WGA studies provides an unprecedented opportunity to analyze genetic data in the context of decades of research on systems biology and prior genetic studies of a phenotype.

6

Joint analysis of 591 SNPs in the IGF pathway and prostate cancer risk: results from the Breast and Prostate Cancer Cohort Consortium (BPC3)

S. Lindstrom (1), P. Kraft (1)
(1) Program in Molecular and Genetic Epidemiology, Harvard School of Public Health

Technological advances have enabled researchers to genotype large numbers of SNP markers. These SNPs are often tested for association with disease risk individually, but joint analysis of multiple SNPs may be more powerful if marginal effects are small and interaction effects exist (e.g. due to local LD structure or epistasis). The ability to detect joint effects depends on several factors including allele frequency, number of causal loci and sample size. The relative power of different analysis strategies will differ depending on the (unknown) true penetrance model, but all multilocus analysis strategies face practical obstacles, including missing data and computational burden. We present the results from four different statistical methods for testing the association between disease risk and multiple genetic variants, as applied to data on 591 SNPs in 24 IGF-pathway genes among 6,600 prostate cancer cases and 7,500 controls from the BPC3. These methods included single-SNP analysis, forward stepwise logistic regression (using a permutation test to preserve validity) and kernel machine procedures. We performed the analysis gene-by-gene as well as on an overall pathway level. We found no associations with SNPs in the IGF pathway and prostate cancer risk beyond a single SNP in IGF1, but note that multi-locus analyses did not make use of over 25% of the available samples which had a high proportion of missing and non-imputable SNPs. We gratefully acknowledge the BPC3 investigators for access to data.

7

Optimising the power of genome-wide association studies by using publicly available reference samples to expand the controls group

J.J. Zhuang (1), K. Zondervan (1), F. Nyberg (2,6), C. Harbron (3), A. Jawaid (4), L.R. Cardon (1,5), B.J. Barratt (4), A.P. Morris (1)
(1) Wellcome Trust Centre for Human Genetics, Oxford Univ., UK
(2) Epidemiology, AstraZeneca R&D, Mölndal, Sweden
(3) Statistical Sciences, AstraZeneca, UK
(4) Research and Development Genetics, AstraZeneca, UK
(5) Fred Hutchinson Cancer Research Center, Seattle, USA
(6) Institute of Environmental Medicine, Karolinska Institute, Stockholm, Sweden

Genome-wide association (GWA) studies have proven to be successful in identifying a number of novel genetic loci contributing to complex human diseases. However, they have also highlighted that many potential loci of modest effect remain undetected due to the need for study samples of many thousands of individuals. Large-scale international initiatives, aim to facilitate discovery of modest-effect genes by making genome-wide data publicly available. These resources can be designed to improve the detection of disease genes by allowing disease data sets to be combined at the level of raw data for the purpose of pooled analysis. In principle, genetic data on samples from these studies could be used to increase the power of a GWA study *via* judicious use as genetically-matched controls for other traits. This strategy is not without potential problems. We present simulations to demonstrate that naïve application of this strategy can greatly inflate the false positive error rate in the presence of population structure. To remedy this problem, we present a novel application using model selection methods to define axes (major components) of genetic variation, and to identify those which are disease associated. These axes are then included as covariates in the GWA association analysis to correct for population structure between the added samples and the core case-control set. Simulations demonstrate that the false positive error rate of this approach is robust even to moderate and realistic population structure, but can result in appreciable increases in power over standard analysis of samples from the original GWA study data alone.

8

Multiple Testing in Genomewide Association Studies: How Bad Really is the Bonferroni Correction?

J.P. Lewinger (1), D.J. Duggan (2), D.M. Taverna (2), W.J. Gauderman (1), D.O. Stram (1), D.C. Thomas (1)
(1) Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA
(2) Translational Genomics Research Institute, Phoenix, AZ, USA

Commercially available SNP panels for genomewide association studies (GWAS) are selected to maximize the coverage of common variants with minimal redundancy. However, linkage disequilibrium (LD) between panel SNPs can remain high and it is thus widely believed that adjusting for multiple testing using a Bonferroni correction is highly conservative. To empirically determine the required multiple testing adjustment for six commercially available GWAS panels, Illumina HumanHap300, HumanHap550, HumanHap650Y and Human1M and Affymetrix SNP5.0 and SNP6.0, we resampled HapMap haplotypes to create null datasets that preserve the LD structure of the HapMap samples. We estimated the null distribution of the maximum of 1 degree of freedom tests for the SNPs in each of the panels. The upper tail of these distributions provides the critical values required to control the global type I error rate. In the tail, these distributions are very well approximated by the maximum over an 'effective number' of independent tests ranging between 40% and 80% of the actual number of SNPs in the panels. Thus, when employing a Bonferroni adjustment, one is effectively 'overcorrecting' for hundreds of thousands tests. However, this overcorrection results in negligible power losses for very high or very low powered studies. The most conservative scenario occurs for studies powered at close to 50% which loose less than 5 percentage points of power.

9

Beyond the results of genome-wide association studies

D.F. Schwarz, A. Ziegler, I.R. König
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Germany

Recently, a number of genome wide association (GWA) studies identified and validated novel single nucleotide polymorphisms (SNPs) strongly associated with complex diseases [1,2]. Despite this obvious success, this poses only a stopover en-route to understanding the genetic and biological background of complex diseases. In this presentation, we focus on two possible ways to progress further.
First it is important to better understand the biological processes which might be associated with the identified genomic region for a specific disease. For this, we developed the tool SNPtoGO [3] that links biological processes in the gene ontology (GO) database with given SNP information. A GO term's dominance is characterized by the ratio of the number of observed appearances in a particular set of SNPs versus the number of expected appearances for a random selection. Second, it is not only interesting to identify associated SNPs and regions, but also to develop prediction models able to correctly classify patients and controls, and SNP interaction lists. Given the known constraints of classical multivariate regression analyses in the setting of GWAs, we developed the machine learning tool Random-Jungle (RJ). This is an efficient generalized implementation of RandomForests. Specifically, a RJ can be a collection of any desired decision trees.
Our two approaches show a stride towards the understanding of underling biological processes and thus serve to build on the results from classical GWAs.

Reference:
[1] N Engl J Med. 2007;357(5):443-53
[2] Circulation. 2008;117(13):1675-84
[3] Bioinformatics. 2008;24(1):146

10

Robust optimal receiver operating characteristic (ROC) curve for predictive genetic tests

Q. Lu, X. Zhu, S. Won and R.C. Elston
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106

Current ongoing genome wide association studies represent a powerful approach to uncover many unknown genetic causes for common complex diseases. The discovery of these genetic variants offers an important opportunity for early disease prediction, prevention and individualized treatment. We here introduce a novel approach to combine multiple genetic variants for disease prediction. The approach is derived based on the optimality theory of the likelihood ratio rule. This theory simply shows that the receiver operating characteristic (ROC) based on likelihood has maximum performance at each cutoff point and the area under its ROC curve (AUC) is

highest among all approaches. Through simulations and real data application, we compared the proposed approach with the commonly used logistic regression and the classification tree. The three approaches have similar performance if we know the underlying disease model. However, for most of the common diseases, we have little prior knowledge of the disease model, in which situation the new approach has an advantage over the logistic regression and the classification tree approaches. We apply the new approach to Type 1 diabetes data and the AUC of the test is estimated to be 0.730, which is improved over logistic regression and relatively high for common diseases.

11

Kernel Based Adaptive Cluster (KBAC): A Powerful Method to Detect Associations for Complex Traits due to Rare Variants in the Presence of Gene x Gene and Gene x Environment Interactions

D.J. Liu (1,2), S.M. Leal (2)
(1) Department of Statistics, Rice University, Houston, TX 77005
(2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030

Recent studies have demonstrated that common diseases can be due to multiple functional variants with allele frequencies ranging from rare to common. Both gene x gene and genes x environment interactions can play a role in disease susceptibility. Although genome-wide association studies using tagSNPs are a powerful approach for detecting common variants, they are underpowered to detect associations with rare variants due to indirect mapping. The development of cost-effective sequencing technologies enables the detection of rare variants for use in association studies. Although methods used for the analysis of common variants are applicable to sequence data, their performance is poor for analyzing rare variants and their power may be further reduced in the presence of interactions. We developed a novel method, Kernel Based Adaptive Cluster (KBAC) to carry out direct association studies of rare variant data. This method can be used for gene mapping with or without the presence of gene x gene and gene x environmental interactions. Significance for the KBAC method can either be determined empirically through permutation or using the derived asymptotic distribution when the sample size is sufficiently large. It is demonstrated through extensive simulations motivated by real data (e.g. Hirschsprung's Disease, Breast Cancer) that the KBAC method has superior performance than univariate and multivariate tests for detecting associations with or without interactions. The KBAC method is powerful and robust even in the presence of misclassification error where either non-causal variants are included in or casual variants are excluded from the analysis. The KBAC method can be applied to the analysis of sequence data from either whole genomes or candidate genes.

12

Crystal Ball, Magic 8 Ball, or Both? Empirical examples of the promise and limits of genetic risk prediction for prostate cancer and type 2 diabetes

P. Kraft (1), S. Wacholder (2), M. Cornelis (3), F. Hu (3), R.B. Hayes (2), D.J. Hunter (1), S. Chanock (4)
(1) Prog. in Mol. and Gen. Epi., Harvard School of Public Health, (2) Div. of Can. Epi. and Gen., NCI, (3) Dep. of Nut., HSPH, (4) Core Geno. Fac., NCI

Summary genetic risk scores based on common, independent alleles are strongly associated with risk of prostate cancer and diabetes. Using data from the Cancer Genetic Markers of Susceptibility prostate cancer genome-wide association scan and a nested case-control study of type 2 diabetes based in the Health Professionals' Follow-up Study, we show that despite large relative differences in risk between extremes of genetic risk scores for these diseases, the scores add little to the discriminatory power of standard risk factors for prediction of disease development (as measured by sensitivity and specificity). We discuss why measures of absolute risk are also important in the context of predictive genetic testing and how they can be estimated from case-control data using external-population or underlying-cohort rates. The absolute five-year risk for men with high genetic risk scores for these two diseases is generally low, although a small subset of men who have traditional risk factors (e.g. a family history of disease) and high genetic risk scores are at notably high risk. These results–and the lack of a clear mortality-reducing primary intervention for prostate cancer–suggest that broad genetic risk screening for these diseases currently has little clinical utility. Eventual accuracy of predictive genetic testing will be limited by (knowledge of) the genetic architecture of disease.

13

Combining information from multiple genes to evaluate clinical validity in nuclear families with affected offspring: autism spectrum disorder as example

A. Ziegler (1), J. Carayol (2), F. Tores (2), F. Rousseau (2), J. Hager (2)
(1) Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Germany, (2) IntegraGen SA, France

Genetic tests for complex diseases are clinically useful only if a number of disease genes is combined using case-control or cohort studies. For early onset diseases like autism spectrum disorder (ASD), only family rather than case-control data is available. Using nuclear families, we explored the effect of combining information from the four genes EN2, PITX1, SLC25A12 and ATP2B2 for risk assessment of ASD. For this purpose, we developed a novel approach for evaluating clinical validity using family data with affected offspring only.

Two-hundred-twenty-six nuclear families were taken from the Autism Genetic Resource Exchange (AGRE) repository. The first sibling to a proband and both parents were genotyped. A single SNP was chosen for each gene to avoid haplotyping. We estimated odds ratios for the number of risk alleles (risk score). Sensitivity and specificity were estimated, and an receiver operating characteristic was constructed for the number of risk alleles. The increase per risk allele was 1.35 (95% confidence interval (CI): 1.16-1.58), and the area under the curve (AUC) was 0.59 (95% CI: 0.57-0.61). With the four genes risk model, the AUC is in the same order as the AUC from comparable case-control studies in other

complex diseases. With the risk score approach we are able to provide measures for judging clinical validity from affected offspring nuclear family data.

14

Whole genome analysis of copy number variants in lung cancer etiology

C.I. Amos (1), E. Lu (1), C.G. Lambert (2), G.F. Rudy (2), J.E. Grover (2), I.H. Lake (2), X. Wu (2), M.R. Spitz (1)
(1) Department of Epidemiology, U.T. M.D. Anderson Cancer Center, (2) Golden Helix Inc, Bozeman, MT

As a part of our ongoing research to identify genetic factors influencing risk for lung cancer, we performed genetic analysis of 1154 lung cancer cases and 1137 controls, matched for smoking behavior, sex and age to the lung cancer cases using an Illumina Hap300K BeadArray. To identify potential copy number variants, we obtained the log R ratios and used the Copy Number Analysis Module (CNAM) of Golden Helix' SNP & Variation Suite to compare the log R ratios in cases to that in controls, so identifying areas suggesting copy number changes. To control for variability among subjects that may relate to sample processing and ethnic substructure among Caucasians, we applied principle components analysis to the correlations among individuals for the log-R intensities. Results from the principal components analysis showed only 8 eigenvalues exceeding 1.0. Conservatively correcting for 10 eigenvectors, we found highly significant associations of log-R intensities with two genomic regions encompassing the DAD1 gene on chromosome 14q ($p=3 \times 10\text{-}30$), and the TARP gene on chromosome 7p ($p=1 \times 10\text{-}20$). Both of these regions contain repetitive elements relating to T-cell immune responses, raising possibilities that rearrangements in these regions affect cancer development. Several other regions also displayed highly significant results including LAPTM5 ($p=6 \times 10\text{-}9$, chromosome 1p), DISC1 ($p=2 \times 10\text{-}8$, chromosome 1q42), and LOC 51236 ($p=9 \times 10\text{-}8$, chromosome 8q).

15

Methods and Discoveries Drawn from Twenty Whole Genome Copy Number Variation Studies

C.G. Lambert, J.E. Grover, I.H. Lake, G.M. Linse, G.F. Rudy
*Golden Helix, Inc.

With over a dozen ongoing collaborations with leading research organizations & access to a wealth of whole genome studies across multiple platforms, we have seen specific themes emerging. One of the most persistent & challenging issues has been batch effect correction. With high genotyping call rates mostly unaffected by plate effects, the majority of research groups have insufficiently randomized cases & controls on plates or borrowed controls from other experiments. Unfortunately, copy number variation (CNV) studies are very sensitive to batch effects. We have employed PCA-based approaches to mitigate the enormous confounding of CNV association studies by batch effects & population stratification. This presentation will focus on methods of data processing, whole genome association analysis & the challenges of CNV segmentation using the various Wellcome Trust case/control studies (~2000 cases, ~1500 common controls) on Rheumatoid Arthritis, Bipolar Disorder, Coronary Artery Disease, Crohn's Disease, Hypertension, Type I Diabetes, & Type II Diabetes. Further, in looking at over 20 different studies, we have seen particular regions of chromosome 7 & 14 persistently associated across a majority of diseases, regardless of genotyping platform. We also found 30-40% of CNV associations are confirmed by past studies, whereas 60-70% represent novel findings. Corroborated CNV associations of note in the Wellcome Trust studies are regions in or near CHL1, PPP1R12B, SLC8A1, SMAD6 in Coronary Artery Disease, CHL1 in Bipolar Disorder, BTNL2 in Rheumatoid Arthritis, PTPRD in Hypertension, and GGTL4 in Type I Diabetes.

16

Modeling linkage disequilibrium in DNA sequence data for improved polymorphism discovery and genotype calling

P. Scheet
Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

The advent of high-throughput DNA sequencing technologies promises to allow detailed surveys of human genetic variation. However, current technologies are prone to nontrivial error rates. Furthermore, the depth of sequenced reads for a given individual is not uniform and may leave some loci with little or no coverage of mapped reads. These factors contribute to difficulties in calling individual genotypes or verifying whether individual loci are polymorphic. The dependence of alleles at nearby loci (linkage disequilibrium; LD) provides a "built-in redundancy" of sequenced loci. Here, I adapt a model for LD among tightly-linked SNP markers to DNA sequence data for the purpose of explicitly taking account of this dependence, thereby allowing one to "borrow strength" from nearby markers when making single-marker inferences. To take advantage of available information about the quality of the sequenced reads, I fit the model directly to likelihoods for the true genotypes and compare this approach to using as input the allelic counts or called genotypes. I apply the method to simulated data as well as to data from The 1000 Genomes Project to demonstrate that the incorporation of LD information leads to an improvement in polymorphism discovery and obtaining accurate probabilities of individual genotypes.

17

Increased power for candidate gene screening by classification of polymorphisms through evolutionary conservation

A. Thomas (1), S.V. Tavtigian (2)
(1) Genetic Epidemiology, University of Utah, Utah, USA
(2) IARC, Lyon, France

To assess whether a candidate gene has an effect on a disease phenotype, we can sequence the gene in a set of cases and controls and evaluate the differences in the profile of polymorphisms or mutations seen in these groups. In the absence of a smoking gun mutation, such as a mutation that

creates a premature stop codon early in the gene's sequence, such a case-control association study often has little statistical power due to the low frequency of likely deleterious miss sense variants. Improving power by subselecting variants that appear more frequently in cases clearly introduces biases that require further samples to remove. We propose to increase power by a prior grouping of variants that is independent of the sample. Specifically, we assess whether a variant is likely to be deleterious by comparing its deviation from the canonical protein sequence with the variation seen at that site across a broad range of species. The approach is illustrated by analysis of the profiles of miss sense mutations seen in a large set of DNA sequences for the BRCA1 and BRCA2 breast cancer susceptibility genes. The method is statistically robust and unbiased.

18

New Haplotype Sharing Method for Genome-Wide Case-Control Association Studies Implicates Gene for Parkinson's Disease

Andrew S. Allen (1), Glen A. Satten (2)
(1) Duke University Department of Biostatistics and Bioinformatics, (2) Centers for Disease Control and Prevention

The large number of markers considered in a genome-wide association study (GWAS) has resulted in a simplification of analyses conducted. Most studies are analyzed one marker at a time using simple tests like the trend test. Methods that account for the special features of genetic association studies, yet remain computationally feasible for genome-wide analysis, are desirable as they may lead to increased power to detect associations.
Haplotype sharing attempts to translate between population genetics and genetic epidemiology. Near a recent disease-causing mutation, case haplotypes should be more similar to each other than control haplotypes. We give computationally simple association tests based on haplotype sharing that can be easily applied to GWASs while allowing use of fast (but not likelihood-based) haplotyping algorithms and properly accounting for the uncertainty introduced by using inferred haplotypes. We also give haplotype sharing analyses that adjust for population stratification.
Applying our methods to a GWAS of Parkinson's disease, we find a genome-wide significant signal in a biologically-plausible gene that is not found by single-snp methods. Further, a missing-data artifact that causes a spurious single-SNP association on chromosome 9 does not impact our test.

19

Fast and Robust Tests of Association for Untyped SNPs in Case-Control Studies

M.P. Epstein (1), A.S. Allen (2), S. Griffiths (3), F. Dudbridge (3), G.A. Satten (4)
(1) Dept. of Human Genetics, Emory Univ., USA, (2) Dept. of Biostatistics, Duke Univ., USA, (3) MRC Biostatistics Unit, Cambridge, UK, (4) CDC, USA

Case-control association studies of complex diseases typically genotype and analyze a set of tagSNPs that capture the genetic variation within a region of interest. However, recent literature has proposed novel methods for analyzing untyped SNPs, which can assist in signal localization and permit cross-platform comparison of results from different studies. Such methods typically extrapolate information on the untyped SNP in the test sample using the observed tagSNP data coupled with external linkage-disequilibrium (LD) information on all SNPs from a detailed catalogue of human genetic variation (such as HapMap). Using this logic, we propose a novel efficient-score procedure for testing untyped SNPs in case-control association studies. Our approach is simple to implement and easily allows for covariates. However, the main strength of our approach is that it is robust to misspecification of external LD patterns as well as to haplotype-phase ambiguity in genotype data. As a result, our method is robust to an inappropriate choice of the reference sample and further requires only a fraction of the computation time required by other methods to test untyped SNPs in a genomewide association scan. At the same time, our method appears to have similar power compared to popular hidden-Markov methods for testing untyped SNPs. We illustrate our approach with an application to a first-stage genomewide association study of Parkinson's Disease.

20

Quantifying the effects of imputation on the power, coverage and cost-efficiency of genomewide SNP platforms

C.A. Anderson (1), F.H. Pettersson (1), J.C. Barrett (1), J.J. Zhuang (1), I. Ragoussis (1), L.R. Cardon (2), A.P. Morris (1)
(1) The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, United Kingdom, OX3 7BN
(2) Fred Hutchinson Cancer Research Center, Seattle, WA 98109–1024

Genotype imputation is potentially a zero-cost method for bridging gaps in coverage and power between genotyping platforms. Here, we quantify these gains in power and coverage using 1,376 population controls from the 1958 British Birth Cohort genotyped by the Wellcome Trust Case-Control Consortium using the Illumina HumanHap-550 and Affymetrix SNP-array-5.0 platforms. Direct genotypes from SNPs featuring on the Illumina HumanHap-300 or Affymetrix SNP array-5.0 were used separately to impute genotypes for SNPs featuring solely on the Illumina HumanHap-550. We contrasted the direct and imputed genotypes at these SNPs using partial least-squares projection to latent structures discriminant analysis. Approximately 50% of genotypes at SNPs exclusively on the HumanHap-550 can be accurately imputed from direct genotypes on the SNP array-5.0 or Illumina HumanHap-300. This approximately halves differences in coverage and power between the platforms. When the relative cost of currently available genome-wide SNP platforms is accounted for, and finances are limited but sample size is not, the highest powered strategy in European populations is to genotype a larger number of individuals using the HumanHap-300 platform and carry out imputation. Platforms consisting of around 1 million SNPs offer poor cost-efficiency for SNP association in European populations.

21

Preserving Candidate Regions in the Era of Genome-Wide Association via Stratified False-Discovery Rate Control Improves Power

L. Sun (1,2,3), Y.J. Yun (4), S.B. Bull (4,1), A.D. Paterson (3,1), D. Waggott (4)
(1) Pub Health Sci, (2) Statistics, U. of Toronto, (3) Sickkids, (4) Samuel Lunenfeld Research Institute, Toronto, Canada

A central issue in Genome-Wide Association (GWA) studies is how to assess statistical significance taking into account the inherent large-scale multiple hypothesis testing. To improve power, we emphasize the importance of utilizing available prior (e.g. linkage results) and/or auxiliary information (e.g. minor allele frequency) via a Stratified False Discovery Rate (SFDR) control method (Sun et al, 2006). The essence of SFDR is to prioritize the genome based on the prior information, and maintain the power of interrogating candidate regions within the GWA design framework. These candidate regions can be defined as, e.g., the linkage-peak regions and/or candidate genes. In addition, we theoretically unify the SFDR approach and the Weighted FDR method (Roeder et al, 2006), and we perform extensive simulation studies to compare them. The empirical results confirm our theoretical derivation and show that when the prior is informative, both methods improve power with similar performance. However, when the prior is uninformative, unlike WFDR, SFDR does not lose power. Finally, we demonstrate the utility of the methods by applying them to two GWA studies, WTCCC and an on-going GWA study of data from the Illumina 1M chip in 1,304 DCCT/EDIC probands for association with diabetic retinopathy, using previous GW linkage results as prior information. In both studies, we identify some cases in which the SFDR method appears to increase power.

22

Ordered-subset regression analysis is a useful approach to detect heterogeneity of association: 17q21 SNPs and early-onset asthma

E. Bouzigon (1), E. Corda (1), H. Aschard (1), M.H. Dizier (2), N. Chateigner (1), F. Kauffmann (3), M. Lathrop (4) & F. Demenais (1)
INSERM (1) U794, (2) U535, (3) U780, Paris, (4) CNG, Evry, France

A genome-wide association study identified genetic variants at 17q21 locus associated with asthma in British and German children. To elucidate further the relationship of this locus with disease, we investigated whether the effect of these variants differ according to age-of-onset of asthma in 372 French EGEA families. We first replicated association of asthma with 11 of 36 SNPs typed at 17q21 locus ($p < 0.01$). We performed an ordered-subset regression analysis (OSRA) to identify the cut-off point to classify patients into early-onset asthma and late-onset asthma. For each ordered age-specific subset, we computed **Δ**-LRT, the difference between the likelihood ratio test (LRT) statistic for association in that subset and the baseline LRT from the whole sample. We found that **Δ**-LRT was maximum at ≤ 4 years of age for all 11 markers ($10^{-4} \leq p \leq 0.02$ using permutations). This age was thus used to define early-onset and late-onset asthma. Subsequent association analysis, using a likelihood-based method (LAMP program), showed highly significant association with early-onset asthma ($p < 10^{-5}$ for 4 SNPs), while no association was found with late-onset asthma ($p \leq 0.002$ against homogeneity).This study shows that, in the context of association studies, OSRA is a useful approach to identify genetic heterogeneity. It also demonstrates that increased risk conferred by 17q21 variants is restricted to early-onset asthma. Funded by: EC-FP6 (GABRIEL, GA2LEN)

23

Modeling Age Variation in QTL Effects Leads to Substantially Improved Linkage Evidence for Blood Pressure in the HyperGEN Study

G. Shi, C.C. Gu, A.T. Kraja, D.K. Arnett, R.H. Myers, J.S. Pankow, S.C. Hunt, D.C. Rao
Washington Univ. in St. Louis (GS, CCG, ATK, DCR), USA, Univ. of Alabama at Birmingham (DKA), USA, Boston Univ. (RHM), USA, Univ. of Minnesota (JSP), USA, Univ. of Utah (SCH), USA

This study presents genome-wide linkage analysis of blood pressure phenotypes using a recently developed variance components method incorporating age variation in QTL and polygenic effects. The analysis was motivated by our prior work and animal models. Linkage analysis was carried out for systolic blood pressure (SBP) and diastolic blood pressure (DBP) in the Hypertension Genetic Epidemiology Network. Previous linkage analyses of hypertension in these data yielded a maximum LOD score of 2.08 on chromosome 2 [Rao et al., 2003], and conventional variance component linkage analysis of SBP yielded a LOD score of 2.17. The new method resulted in substantially improved linkage evidence, with several significant linkage peaks with maximum LOD scores ranging between 3.01 and 4.61. We note that 14 of these signals are replicated in the literature. We also investigated false positive rate by this method and are satisfied that, for the most part, the empirical and nominal type I error rates are nearly identical. In conclusion, this investigation provides ample evidence about the promise of our methodology and that QTL effects on blood pressure seem to vary by age. The vastly improved genetic linkage results presented here should help in identifying the specific genetic variants that explain the observed results.

24

Things to think about before performing GW analysis: an experience with the Affymetrix 6.0 SNP Array

M. de Andrade (1), E.J. Atkinson (1), W.R. Bamlet (1), S. Maharjan (1), M.E. Matsumoto (1), S.L. Kardia (2)
(1) Div. of Biostatistics, Mayo Clinic, USA, and (2) Dept. of Epidemiology, University of Michigan, USA

In an era of genome-wide association analyses, researchers are facing the challenge not only of analyzing a large volume of data but also of processing the genotype data and creating appropriate workflows. We present our experiences in working with the Affymetrix 6.0 Genome-Wide Human SNP Array including the workflows that we have created. In

particular, we focus on the issues prior to genotype extraction, assessment of genotype accuracy, and automation of the workflows recognizing that the Affy 6.0 data will eventually be used for analysis using a wide range of study designs. We will share our experience working with Birdseed1 and 2 using individual plate and all plates to generate genotype call, examining replicate samples and the challenges with quality control measures in sibships. We will present our workflow and initial results using 900 samples of hypertensive sibships from Rochester, MN.

25

Differential Bias in Genotype Calls between Plates due to the Effect of a Small Number of Lower DNA Quality and/or Contaminated Samples

A. Pluzhnikov (1), J.E. Below (1), A. Tikhomirov (1), A. Konkashbaev (1), C. Roe (1), D. Nicolae (1), N.J. Cox (1)
(1) Dept. of Medicine, The Univ. of Chicago, USA

Quality control (QC) issues in genotype calling are becoming increasingly important in genome wide association studies. We examined the effect of a small number of samples that had lower than average DNA concentration, DNA fragmentation, contamination, or other quality issues not detected by standard QC, on genotype calling of other samples in the study using data from the Genetics of Kidneys in Diabetes (GoKinD) collection of more than 1600 unrelated probands with Type I diabetes. All samples were typed using the Affymetrix Genome-Wide Human SNP Array 5.0 platform, and genotypes were called by plate using the Birdseed v.2 algorithm to minimize the amount of missing data. We detected 8 problematic samples displaying unusual patterns of relatedness and high levels of heterozygosity and showed that, for a number of markers, these samples cluster together, thus altering the allele calls for other samples on the plate and leading to significant differential bias in allele frequencies between plates that eventually resulted in false-positive associations in the GoKinD data. We discuss ways to detect this kind of differential bias between plates, and to correct it depending on the availability of the raw intensity (.cel) files.

26

Bias-Corrected Effect Estimators for Genome-Wide Association Studies

R. Pahl (1), T.T. Nguyen (1), A. Hinney (2), B. Greene (1), J. Hebebrand (2), H. Schäfer (1)
(1) Institute of Medical Biometry and Epidemiology, Philipps-University, Marburg, Germany, (2) Dept. of Child and Adolescent Psychiatry, University of Duisburg-Essen, Essen, Germany

Naive estimates of the genetic effect size such as odds ratios are often reported for the top markers in genome wide association studies, but are known to exaggerate the true genetic effects. We propose a bootstrap method to correct for the selection bias. While the likelihood-based method recently proposed by Zoellner and Pritchard (2007) is focused on an individual marker for which genome wide significance was reached, the bootstrap method generates bias-reduced point and interval estimates for the whole list of top markers. In our Monte Carlo simulations based on a two-locus genetic model with OR=1.275 and OR=1.39, the bootstrap removed between 40% and 100% of the bias of the naive log odds ratio estimator, depending on the sample size and the position in the sorted list of top SNPs. Our bootstrap procedure also allows to calculate simultaneous confidence intervals (CIs) for a specified number of top markers. With nominal confidence level of 95% we found simultaneous coverage of about 90% for sample sizes not smaller than 1000. In a GWAS on early onset obesity, the naive odds ratio estimate for the top marker in the FTO gene was sized down from 1.66 to 1.47. An OR of 1.42 was found in an independent replication study.

References:

[1] S. Zoellner, J.K. Pritchard, 2007. Am J Hum Genet 80:605–615.

27

Impact of correcting for population structure by adjusting for global and local ancestry

S.J. Kang (1), H. Lyon (2), B. Tayo (3), C. Chiang (2), T. Feng (1), R. Cooper (3), J. Hirschhorn (2,4), X. Zhu (1)
(1) Case Western Reserve Univ. (2) Children's Hospital, Harvard Med Sch (3) Loyola Univ. Med Center (4) Broad Inst.

The impact of using local ancestry estimates to correct for local population structure has not been well studied. The global population structure may reflect demographic history, while local structure may be driven by demography, fluctuation in admixture in certain regions, and natural selection. We compared the global and local population structure ancestry estimates by examining 735 African Americans and 857989 SNPs (Affymetrix 6.0 platform) after quality control. We performed multidimensional scaling analysis and calculated correlations between the global ancestry estimated using all available markers in the genome and the local ancestry estimated using markers in local regions, which are defined as 10Mb intervals. We picked several regions that previously showed linkage evidence to obesity and performed association analysis with BMI by adjusting for the global and local ancestry. Our results show that adjusting for local ancestry increases the association evidence in some regions with previous linkage evidence. We also observed some association evidence became weaker after adjusting for local ancestry, possibly due to false positive or driven by local population structure. This study suggests potential benefit by adjusting for local ancestry in association studies possibly in identifying loci where admixture signals contribute to the evidence of association and possibly in eliminating false positives due to local variation in population structure

28

Genome-wide association scan identifies a novel susceptibility locus for Psoriasis on 6p21.3, independent of HLA-C

B.-J. Feng (1), R. Soltani (1), A.M. Bowcock (2), R.P. Nair (3), J.T. Elder (3), A.B. Begovich (4), G.R. Abecasis (3), K.C. Duffin (1), D.E. Goldgar (1), G.G. Krueger (1)
(1) Univ. of Utah, USA

(2) Washington Univ., USA
(3) Univ. of Michigan, USA
(4) Celera, USA

Abstract: A genome-wide association study was performed to investigate the genetic contribution to the development of psoriasis. Samples in the initial scan include 1384 cases and 1414 matched controls, which were genotyped for 450,652 SNPs by Perlegen Biosciences as part of the GAIN (the Genetic Association Information Network) initiative. The most significant finding was for SNP rs12191877 ($p=8 \times 10^{-54}$, OR=2.8), which is located 13kb centromeric of HLA-C, and is in high linkage disequilibrium with HLA-C. Using the detailed study of Bakker P et al. (Nat Genet, 2006) and the genotypes of 4 SNPs, we imputed in all samples the HLA-C*0602, the consensus risk allele for psoriasis, with error rate of $<0.5\%$ (2 out of 412 genotypes) among a subset of the samples who have been serotyped. Another proposed psoriasis risk allele, HLA-C*1203, was also imputed. The association of the imputed HLA-C*0602 allele with disease risk was much stronger than the original SNP rs12191877 ($p=3 \times 10^{-62}$, OR=3.7). Adjusting for HLA-C, one additional association signal emerged in the 6p21.3 region, rs2073048 ($p=3 \times 10^{-6}$, OR=1.5), which is located 1.1Mb centromeric of HLA-C. This association remained significant in an analysis using only the HLA-C*0602-negative individuals ($p=4 \times 10^{-5}$, OR=1.5), and using the HLA-C*0602-negative and HLA-C*1203-negative individuals ($p=3 \times 10^{-4}$, OR=1.5), demonstrating that there may be an additional susceptibility gene on 6p21.3, which acts independently of HLA-C to confer risk of psoriasis.

29

Haplotype Association Analyses: Power Gain from Phasing Under the Alternative Hypothesis?

Ryan Abo, Nicola J. Camp
Dept. of Biomedical Informatics, Univ. of Utah, USA

Haplotype association analysis requires estimation of haplotypes from the observed unphased genotypes. The current norm is to phase all individuals together, and the estimation is made under the null hypothesis that haplotypes frequencies do not differ between cases and controls. If association does exist, however, haplotype frequencies will differ between cases and controls. Thus, phasing separately may gain power. We investigated this using hapMC, a haplotype analysis software that uses a Monte Carlo (MC) approach. This analysis uses the estimated maximum likelihood estimates (MLE) for haplotype pairs. hapMC generates simulated null genotypes for all individuals and MLE haplotype pairs for the null data are used to calculate null statistics to create a null distribution.
Using linkage disequilibrium structure from HapMap data we simulated a four-locus haplotype to tag a single disease SNP (dSNP) ($r^2=0.8$) and a sample size of 2,000 cases and controls using SimuPOP[1]. The genetic models included allele frequencies from 0.01-0.15, low to high relative risk (RR=1.2-4.0), and a sporadic rate of 0.01. A null model (RR=1.0) indicated both tests had correct type 1 error rates. For common dSNPs (≥ 0.05) and a low RR (1.2), there was no power difference. For a rarer dSNP (~ 0.01) and more moderate RR (~ 2.0) we observed a loss of power when phasing under the alternative hypothesis (0.449 vs. 0.51 when phased together).
This counter-intuitive finding is due to a lack of impact in the likelihood estimates for individuals' haplotype pairs despite substantial differences in haplotype frequencies between cases and controls. Hence, although test statistics were more often greater when cases and controls were phased separately, the variance in the null distribution was also greater leading to lower significance. Our results indicate that phasing under the alternative hypothesis does not provide an increase in power.

[1]Peng, B. and C.I. Amos, Bioinformatics, 2008. 24(11): p. 1408–9

30

Extensive Parent-Of-Origin Genetic Effects on Fetal Growth

R. Adkins (1), J. Krushkal (2), G. Somes (2), J. Fain (3), J. Morrison (4), C. Klauser (4), E.F. Magann (5)
(1) Dept. Pediatrics, Univ. TN Health Science Center, USA, (2) Dept. Prev. Med., Univ. TN Health Science Center, USA, (3) Dept. Mol. Sci., Univ. TN Health Science Center, USA, (4) Dept. Ob/Gyn, Univ. MS Medical Center, USA, (5) Dept. Ob/Gyn, Naval Medical Center at Portsmouth, USA

Epigenetic effects have recently been recognized as playing a very significant role in several normal and pathological phenotypes. Imprinting, the silencing of either the paternally or maternally inherited allele, is one of the most pervasive and consistent epigentic mechanisms across species and individuals. The majority of imprinted loci are involved in fetal growth regulation, and several defects in the epigenetic regulation of these genes are associated with extremes of fetal growth. We surveyed patterns of SNP variation in imprinted loci in a cohort of African-American mother-newborn pairs selected using stringent inclusion/exclusion criteria intended to enrich for the genetic component of fetal growth regulation. All association analyses were adjusted for admixture using a suite of ancestry informative SNP markers. By inferring haplotypes within the imprinted loci in mothers and newborns, we could unambiguously infer the parental origin of haplotypes and associated alleles in the majority of newborns. We found very significant parent-of-origin effects in the insulin, H19 and GNAS genes that were completely consistent with their known patterns of imprinting. In the case of the insulin polymorphisms, a consistent trend was also observed for newborn IGF-II levels with respect to parental origin of haplotypes.

31

Significant Linkage Evidence for a Pelvic Floor Predisposition Gene on Chromosome 9

K. Allen-Brady (1), P. Norton (2), J. Farnham (1), L. Cannon-Albright (1)
(1) Departments of Biomedical Informatics and (2) Obstetrics/Gynecology, University of Utah, Salt Lake City, Utah, USA

Predisposition factors for pelvic floor disorders, including pelvic organ prolapse, stress urinary incontinence, overactive bladder and hernias are not well understood. We assessed potential genetic causes of pelvic floor disorders in sister pairs who underwent surgical repair for either prolapse or stress urinary incontinence. We genotyped 80 affected women from 34 families using the Illumina 1 million SNP marker set. Parametric linkage analysis using general dominant and recessive models was performed using the Markov Chain, Monte Carlo linkage analysis method, MCLINK, and a pre-selected set of low linkage disequilibrium markers. Significant genome-wide evidence for linkage was identified on chromosome 9q21 with a HLOD score of 4.16 under a recessive model. Eighteen of our pedigrees had at least nominal evidence for linkage on a by-pedigree basis at this region. Suggestive linkage evidence was also found on chromosomes 3, 7, 10, 15, 19, and X. Our results provide evidence for a genetic basis of pelvic floor disorders.

32

Epistatic Interactions between ADIPOQ and LIPC Influence Insulin Sensitivity Response to Fenofibrate Therapy

P. An (1), M.F. Feitosa (1), D. Warodomwichit (2), D.K. Arnett (3), P.N. Hopkins (4), R.J. Straka (5), J.E. Hixson (6), J.M. Ordovas (2), M.A. Province (1), I.B. Borecki (1)
(1) St. Louis, MO; (2) Boston, MA; (3) Birmingham, AL; (4) Salt Lake City, UT; (5) Minneapolis, MN; (6) Houston, TX

We performed association tests in the GOLDN study to test whether candidate gene variants act interactively to influence insulin sensitivity (IS) response to fenofibrate therapy. IS at baseline and after 3-weeks of once daily 160 mg fenofibrate therapy were estimated in 170 families. The response was estimated using growth curve model of IS before and after the therapy period correcting for the exact number of days taking drug, age, sex, BMI, field center, and the baseline IS. Analysis of 115 selected SNPs within 29 candidate genes was carried out using mixed sandwich estimator approach to account for familial dependencies and logic regression. Significant marginal SNP associations after Bonferroni correction ($p < .0004$) included variants in the ADIPOQ promoter, APOA5 promoter, LIPC exon 3, MTP exon 6, and PDZK1 intron 1. Best combination of interacting SNPs was found with 3 ADIPOQ SNPs (promoter, rs17300539, rs266729; intron 2, rs1501299), and LIPC intron 1 rs1973028. The presence of minor alleles of the 3 ADIPOQ SNPs and homozygous minor allele of the LIPC SNP were associated with increased IS response. Follow-up tests allowing for interactions between rs17300539 and rs1501299 (p=.002), and rs266729 and rs1973028 (p=.030) were significant. In conclusion, the variants at the ADIPOQ promoter, intron 2, and LIPC intron 1 interact and influence IS response to fenofibrate therapy.

33

Stochastic Model for Joint Analysis of Genetic and Non-Genetic Data from Longitudinal Studies of Aging, Health and Longevity

Genet. Epidemiol.

K.G. Arbeev, A.I. Yashin, I. Akushevich, A.M. Kulminski, L.S. Arbeeva, L. Akushevich, S.V. Ukraintseva
Center for Population Health and Aging, Duke Univ., USA

Many longitudinal studies of aging collect genetic information only for a sub-sample of participants of the study. We present a stochastic model for studying such longitudinal data in joint analyses of genetic and non-genetic sub-samples. It includes major concepts of aging known to date: age-specific physiological norms, allostasis and allostatic load, stochasticity, decline in stress-resistance and adaptive capacity. The approach allows for estimating all these concepts together, even if such mechanisms are not directly measured in data (which is typical for longitudinal data available to date). The model takes into account dependence of longitudinal indices and hazard rates on genetic markers and permits evaluation of all these characteristics for carriers of different alleles (genotypes) to address questions concerning genetic influence on aging-related characteristics. The method is based on extracting genetic information from the entire sample of longitudinal data consisting of genetic and non-genetic sub-samples. Thus it results in a substantial increase in the accuracy of estimates of genetic parameters compared to methods that use only information from a genetic sub-sample. Such an increase is achieved without collecting additional genetic data. Simulation studies illustrate the increase in the accuracy in different scenarios for datasets structurally similar to the Framingham Heart Study. Applications of the model to different data and further generalizations are discussed.

34

Pleiotropic and sex-specific effect of IL9 SNPs on two asthma phenotypes using a combination of univariate FBAT statistics applied to principal components

H. Aschard (1), E. Bouzigon (1), E. Corda (1), M.H. Dizier (2), M. Lathrop (3), F. Demenais (1)
(1) INSERM U794, Paris, (2) INSERM U535, Villejuif, (3) CNG, CEA, Evry, France

A bivariate linkage analysis, based on a principal components (PC) approach, led to detect a sex-specific pleiotropic QTL on 5q31 linked to two asthma-related traits: a measure of lung function (FEV1) and a measure of allergen polysensitization (SPTQ) in 295 EGEA families. High evidence for linkage was found, in the male sample, by combining linkage test statistics obtained with PC1 and PC2 (p=7x10-9). To identify the genetic variants associated with these traits, we applied univariate FBAT to FEV1, SPTQ and the two PCs using 24 SNPs belonging to five candidate genes (IRF1, IL13, IL4, IL9, CD14) in the 5q31 region. We also carried out bivariate association analysis by combining univariate FBAT statistics applied to the two PCs, thus extending our linkage approach to association. Univariate FBAT analysis showed evidence for association of rs2069882 of IL9 gene with FEV1 (p=0.0006), SPTQ (p=0.007), PC1 (p=0.004) and PC2 (p=0.001) only in males. Association signals were also found with two other IL9 SNPs. Bivariate association analysis increased evidence for association with the 3 SNPs (p ranging from 4x10-4 to 8x10-5). This study shows that combination of univariate FBAT statistics is a simple and powerful approach to detect genes with a pleiotropic effect.

35

Impact of HLA-DPB1 mismatches on Hematopoietic Stem Cell Transplantation (HSCT)

Y. Balavarca (1), Ludajic (2), Bickeböller (1), Pohlreich (3), Kouba (3), Dobrovolna (3), Vrana (3), Fischer (2), Fae (2), Kalhs (2), Greinix (2)
(1) Univ. of Goettingen, Germany, (2) Medical Univ. of Vienna, Austria, (3) Institute of Hematology and Blood Transfusion, Prague, Czech Republic

The goal of our study was to investigate the role of HLA-DPB1 allele matching in the pair patient/donor on HSCT outcomes: acute and chronic graft versus host disease (aGvHD, cGvHD), transplant related mortality (TRM), relapse, and overall survival (OS). It was also our interest to evaluate the association of polymorphic amino acids mismatches of DPB1 molecule with the HSCT outcomes. The study included 161 patients with unrelated donors who were HLA- A, B, C, DRB1, DQB1 matched at the allelic level, from the centers of Prague and Vienna. All the statistical analyses were adjusted by significant clinical risk factors. The results show that DPB1 allele mismatches were significantly associated with an increased incidence of aGvHD and worse OS. In addition, a mismatch at amino acid position 69 significantly increased the risk for aGvHD and TRM. Risk factors for aGvHD also included amino acid mismatches at positions 8, 9, 35, 76 and 84. This is, to our knowledge, the first report showing a clinically significant association of single amino acid mismatches on HSCT endpoints. Furthermore, the grouping of allelic mismatches into permissive and non-permissive categories and their association with transplantation endpoints proved relevant for TRM but not for the other outcomes.

36

Using Partial Least Squares regression for genetic association studies

Amina Barhdadi, Marie-Pierre Dubé
Montreal Heart Institute, Université de Montréal

In genetic association analyses, genetic markers such as single nucleotide polymorphisms (SNPs) are increasingly being used in prediction of disease outcome. Candidate gene and genome-wide association studies involve a large number of correlated SNPs that exceed the number of patients in the study. Some form of dimension reduction is then required to obtain useful parameter estimates. To have a parsimonious regression model that account for linkage disequilibrium (LD) among markers and that allows the regressors to capture the majority of the variation in the SNP genotypes, we propose the partial least squares regression (PLS) method. PLS is a recent technique that generalizes and combines features from principal component analysis and multiple regression. If the vector of outcome is denoted by y and the SNPs genotype matrix is denoted by X, then PLS regression searches for a set of components (latent vectors) that performs a simultaneous decomposition of X and y with the constraint that these components explain the maximum of covariance between X and y.
Here, comparison of PLS approach performance to principal components regression and other regression based methods was evaluated through simulation studies. Moreover, the PLS method was applied to an association study of depression in cardiac patients and confirmed a previously identified association between a SNP in the VWF gene and depression using a candidate gene approach.
PLS is a powerful approach for analyzing high-dimensional data involved in genetic association studies.This method is characterized by high computational and statistical efficiency

37

A framework to discover complex pathways from observational data

J.W. Baurley, D.V. Conti, W.J. Gauderman, D.C. Thomas
Dept. of Preventive Medicine, Univ. of Southern California, USA

The etiology of common diseases involves a network of complex biological interactions, genetic and environmental. Unraveling these complex interactions has been a challenge in epidemiologic research. We introduce a pathway modeling framework that stochastically discovers a class of plausible pathways from observational data, and allows estimation of both the net effect of the pathway and the types of interactions occurring among genetic or environmental risk factors. Each discovered pathway structure links combinations of observed variables through intermediate latent nodes to a final node, the outcome. Biologic knowledge can be readily applied in this framework as a prior on pathway structure to give preference to more biologically plausible models.
Data was simulated for binary inputs of which only a subset was involved in the pathway. Our algorithm was then used to recover the pathway from the simulated data. The posterior distributions of inputs, pair-wise and higher order interactions, and topologies were obtained by Markov Chain Monte Carlo (MCMC) methods. The evidence in favor of a particular pathway or interaction was summarized using Bayes factors. Our method can correctly identify the risk factors and interactions involved in the simulated pathway. We demonstrate our framework on an asthma case-control dataset with polymorphisms in 12 genes.

38

When LHisA meets Nathan: improved marker selection and power in genomewide haplotype-sharing analysis

L. Beckmann (1), S. Knüppel (2), K. Rohde (2), M. Guedj (3)
(1) Dept. of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany
(2) Dept. of Bioinf, Max Delbrück Center for Molecular Medicine, Berlin, Germany
(3) Ligue Nationale contre le Cancer, programme Carte d'Identité des Tumeurs, Paris, France

The aim of our study is to analyze the applicability of genomewide haplotype analysis in case/control data. Haplotype reconstruction is limited by linkage disequilibrium between markers, thus haplotype reconstruction of whole chromosomes is neither suitable nor feasible. Also, multiple testing of correlated data is an issue in genomewide analysis.

We present a two-stage design to overcome these limitations by combining a local score-based algorithm to identify genomic regions, with a haplotype sharing-based association test. Stage 1: under the assumption that high single point test statistics accumulate around a disease susceptibility locus, a local score statistic is applied to identify genomic regions (Guedj et al. Stat Appl Genet Mol Biol. 2006(5)). Stage 2: within the identified regions, association tests are performed using the Mantel statistics based on haplotype sharing (Beckmann et al. Hum Hered. 2005(59):67-78). Finally, we control for multiple testing and correlation between markers.
To evaluate our approach we drove simulations based on 674 SNPs localized on chromosome 19 from a population of 580 controls. For comparisons we applied a sliding-window approach for haplotype sharing analysis. Additionally, single-point p-values were computed with an unbiased and exact test based on alleles, for which p-values were corrected using a permutation procedure.

39

Localization of a Dominant Genetic Susceptibility Factor in Familial Malignant Mesothelioma

J.E. Below (1), A. Pluzhnikov (1), K. Aquino-Michaels (2), M. Nasu (2), V. Paz (1), B. Mossman (3), H. Pass (4), J.R. Testa (5), M. Carbone (2), N.J. Cox (1)
(1) The University of Chicago, Chicago, IL, (2) University of Hawaii, Honolulu, Hawaii, (3) University of Vermont, Burlington, VT, (4) New York University, New York, NY, (5) Fox Chase Cancer Center, Philadelphia, PA

Malignant mesothelioma (MM) is the primary, yet rare cancer of the pleura, and is generally found in populations with asbestos exposure[1]. The prognosis of MM patients is poor due in part to a prolonged latency period (frequently greater than 25 years) from initial exposure to clinical diagnosis[1], and the highly malignant nature of the cancer. Upon diagnosis the median survival time is 10 months[2], and early detection remains the best approach for reducing morbidity and mortality. An American family with no known asbestos exposure, but a markedly high rate of MM has been identified. All available family members have been genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Prior to linkage analyses, familial relationships were checked and corrected using PLINK[3]. Parametric linkage analyses based on a 0.2 cM SNP map assume a rare dominant model with age-dependent liability classes modeling the expected change in penetrance for different age groups. Preliminary findings indicate that a region on chromosome 6 shows genome-wide significance for linkage (LOD score >3.4). We report details of these analyses, and the results of sequencing and copy number analysis in the disease-linked region on chromosome 6.

Reference:
[1] J.R. Testa and S.C. Jhanwar. *Textbook of pleural diseases*, 120-130 (Arnold, London, New York, 2003)
[2] S. Emri, et al. 2001. Prognostic significance of flow cytometric DNA analysis in patients with malignant pleural mesothelioma. Lung Cancer 33: 109–14
[3] S. Purcell, 2007. PLINK v0.991

40

Conditioning on Hardy-Weinberg equilibrium in order to optimize genome wide association studies

S. Boehringer (1), J. Hebebrand (2), H. Holzmann (3)
(1) Institut für Humangenetik, Universitätsklinikum Essen, Essen, Germany, (2) Klinik für Kinder- und Jugendpsychiatrie und Psychotherapie, Rheinische Kliniken Essen, Essen, Germany, (3) Institut für Stochastik, Universität Karlsruhe, Germany

In current genome wide association studies (GWAs) based on a case-control design, single nucleotide polymorphisms (SNPs) are typically evaluated for an association test and a Hardy-Weinberg-Equilibrium (HWE) goodness-of-fit test. SNPs are then excluded from analysis based on a HWE cutoff to avoid false positives. We have therefore established a conditional genotype based test that conditions the Pearson Chi-Square test in the 3x2 contingency table on the HWE statistic in the control group. We develop the asymptotic theory and derive the asymptotic distribution which turns out to be the convolution of a central Chi-Square and a scaled, non-central Chi-Square distribution, each with one degree of freedom. We show by simulations, that our test is more powerful than the unconditional Pearson test under HWE in the control group. Moreover, an important additional advantage is a better ranking of SNPs in GWAs as HWE is accounted for in computing p-values of SNP association. We demonstrate this effect on a data set in an obesity study. In conclusion, our test makes separate HWE testing superfluous by providing a unified framework and strictly improves on the standard procedure in terms of power and interpretability, thereby making replication more cost effective and improving subsequent fine mapping.

41

Bayesian Inference of Multilocus Models in Genome-wide Association Studies

L. Briollais (1), J. Liu (1), A. Dobra (2), H. Massam (3)
(1) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Canada, (2) Dept. of Stat., University of Washington, U.S.A., (3) Dept. of Stat. and Math., York University, Canada

The actual paradigm to analyze GWAS data is to perform an exhausting testing of all single-SNP associations with the response. Besides the challenging issue related to finding an appropriate significance level, there is no guarantee that the selected subset of SNPs has good prediction properties. Alternatively, one can fit multi-locus models but this raises several problems. Theoretically, modeling discrete mutli-way data is quite challenging and computationally, investigating a huge number of these models is also very difficult. Following the work of Massam and Dobra (2008), we propose to use graphical models within a Bayesian framework to analyze GWAS data. Graphical models provide a general probabilistic framework for making inference and representing the knowledge that we have about complex structured data. For computational issues, we used an efficient search algorithm, MOSS (Dobra and Massam, 2008), which can perform variable selection by moving quickly into the regions of interest of the model space. Another advantage of this approach is the possibility to include informative priors into

the model search. We will illustrate the interest of our new method through an application to the CGEMS breast cancer data and show how it can provide a general framework to find the best combinations of SNPs that can predict breast cancer.

42

Detecting Identity By Descent for Complex Gene Mapping

Sharon R. Browning, Brian L. Browning
Department of Statistics, The University of Auckland, New Zealand

Identity by descent (IBD) mapping has been proposed as a powerful way to map genes underlying complex diseases. Case individuals from a disease with a genetic basis will tend to be more related than average, and in particular will tend to share material inherited IBD from common ancestors at the locations of disease susceptibility variants. Thus, the approach relies on detecting small segments of IBD sharing between individuals. Existing methods for detecting IBD do not adequately account for linkage disequilibrium (LD) between closely spaced genetic markers, such as those in genome wide association mapping SNP panels. We propose a new method for detecting segments of IBD between pairs of individuals. Our method accounts for LD by means of the localized haplotype cluster model. This model forms the basis for our earlier work on multilocus association mapping and haplotype phase inference, and has computational advantages that enable it to be applied successfully on a genome-wide scale. Our method allows for detection of smaller tracts of IBD than can be found using existing methods, for improved power to detect regions harbouring disease-susceptibility variants.

43

Pleiotropic Effects in Thrombotic-Related Traits Using Bivariate Variance Components Models

A. Buil (1,2), A. Martinez-Perez (2), J.C. Souto (2), J. Fontcuberta (2), J.M. Soria (1,2)
(1) Unitat de Genòmica de Malalties Complexes, IR-HSCSP Barcelona, Spain, (2) U. d'Hemostasia i Trombosi, HSCSP Barcelona, Spain

The major pathways in hemostasis (coagulation and fibrinolysis) are complex enzymatic systems composed of multiple inter-related proteins. Variation in several of these proteins has been associated with risk of thrombosis. On the other hand, the variation of the levels of these proteins has a significant genetic component. In our study we looked for evidence of shared genetic effects (pleiotropic effects) among several of these proteins.
We used the sample from the Genetic Analysis of Idiopathic Thrombophilia (GAIT) Project that included 399 individuals grouped in 21 families. In each individual we measured 50 quantitative traits related with the hemostatic systems. We analyzed the data using bivariate variance component models to estimate the genetic and environmental correlations between pairs of traits. The estimated genetic component of this correlation is a measure of the shared genetic effects between the traits, that is, a measure of pleiotropy. First we estimated the genetic correlation between each of the 50 quantitative traits and thrombosis. Fifteen of the traits gave a significant genetic correlation with thrombosis. Then, we estimated the genetic correlations among all of the possible pairs of the 15 selected traits. From this matrix of genetic correlations we extracted two clusters of traits that were genetically correlated among them. In summary, we detected two clusters of quantitative traits that share pleiotropic effects among them and with thrombosis.

44

Sampling ancestries at a hidden disease locus conditional on data from surrounding genetic markers

K.M. Burkett, B. McNeney, J. Graham
Department of Statistics and Actuarial Science, Simon Fraser University,Vancouver, B.C., Canada

The association of genetic variability with disease outcomes reflects the latent genetic ancestries giving rise to the sample's genetic variability. These ancestries, or genealogies, contain information about which sequences carry the disease-predisposing variant. Two loci separated by a recombination event have different parental chromosomes so in general there will be multiple correlated genealogies along a chromosome. Though the genealogies of a sample of sequences from unrelated individuals are typically unknown, the marker data does provide some genealogical information. Incorporating genealogies informed by the marker data into genetic association methods, in a manner that accounts for their uncertainty, therefore requires methods that model their distribution conditional on the observed marker data. However, since the ancestry space is highly complex for even a small number of sequences, it is necessary to sample ancestries from their distribution. This presentation describes our implementation of a Markov Chain Monte Carlo sampler, outlined in Zöllner and Pritchard (Genetics, 169: 1071-92, 2005), that samples genealogies compatible with observed sequence data from unrelated individuals. Each sampled genealogy is for a specified focal point along the sequence. We will apply the sampler to simulated data and discuss how it can be used to understand what the data from surrounding genetic markers tells us about the ancestry at a putative disease locus.

45

Shared Genomic Segment Analysis in Prostate Cancer and Melanoma

Z. Cai (1), A. Thomas (1)
(1) Dept. of Biomedical Informatics, University of Utah, USA

Background: Simulation analysis (Thomas et al, 2008) has shown that long runs of loci that share a common allele identically by state (IBS) could be used to localize hypothesized predisposition genes. Long runs of IBS genomic segments indicate underlying regions shared identically by descent (IBD), and IBD regions shared among related individuals with a common phenotype would in turn become potential candidates for a predisposition gene affecting the phenotype. Shared genomic segment analysis provides the means for identifying IBS genomic regions.
Objective: To identify regions containing potential genes predisposing to cancer using sets of affected individuals distantly related in extended pedigrees.

Methods: We look for long runs of loci at which individuals from extended prostate cancer and melanoma pedigrees share an allele IBS. Sixty individuals from the HapMap CEPH dataset were used as a control group. The genomic anomalies found were examined using tests to detect departures from Hardy-Weinberg equilibrium, extent of linkage disequilibrium and amount of allele heterozygosity.
Results: We found that there are long runs of genomic segments on chromosomes 5 and chromosome 18, shared among both Utah prostate cancer and melanoma pedigrees. These are also shared among European HapMap CEPH individuals. On the other hand, individuals from African and Asian populations genotyped in the HapMap project do not share alleles in these regions.
Conclusions: There is heterozygous sharing of a genomic region on chromosomes 5 and 18 that is peculiar to Europeans. We are still exploring reasonable explanations to our finding.

Reference:
[1] A. Thomas, N.J. Camp, J.M. Farnham, K. Allen-Brady, L.A. Cannon-Albright, 2008. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. Ann Hum Genet. 72(Pt 2): 279–87.

46

Score test for age at onset linkage analysis of selected nuclear families

A. Callegaro, H.C. van Houwelingen and J.J. Houwing-Duistermaat
Dept. of Medical Statistics and Bioinformatics, S-5-P LUMC, 2300 RC Leiden, The Netherlands

We propose to use a four dimensional additive model for genetic linkage analysis of nuclear families taking into account available information on age at onset. For testing we derived the corresponding score statistic from the retrospective likelihood of the identical by descent status given the phenotypic data.
The new score statistic appeared to be a weighted version of the standard score statistic for affected sibling pair linkage analysis (mean test). The weights depend on marginal survival and variance and correlation parameters of the frailty distribution. Analogously to Merlin-Regress [1] linkage analysis for quantitative traits, we propose to use estimates from registries and twin studies. One of the properties of the statistic is that it has the right type I error even when the model for age at onset is not correct. Further the test-statistic can also be used when parents are not genotyped. To illustrate our new statistic we apply it to the age at onset data from the 12th Genetic Analysis Workshop [2] where data have been generated using a complex model of seven genes influencing the liability and the age at onset of a common disease. We conclude that the power of linkage analysis of age at onset data may be increased when the new statistic is used instead of a statistic which does not use the available information on age at onset of the parents.

References:
[1] Genetic Epidemiology (2001) 21S, 332–338.
[2] American Journal of Human Genetics (2001) 68, 1527–1532.

47

Identification of SNPs Explaining a Quantitative Trait Linkage Signal

M.-H. Chen (1), C.-T. Liu (2), Q. Yang (2), J. Dupuis (2)
(1) Dept. of Neurology and Framingham Heart Study, Boston Univ., (2) Dept. of Biostatics, Boston Univ., School of Public Health, USA

Once a linkage signal has been detected, the next step is often to identify genetic variants explaining the observed linkage evidence. For binary traits, one can compare IBD sharing among affected siblings from homozygote and heterozygote parents to test whether a single nucleotide polymorphism (SNP) explains a linkage signal using the Homozygote Sharing Tests (HST). In this study we show that the idea of decomposing IBD sharing from homozygote and heterozygote parents is also applicable to quantitative traits using theoretical derivations and computer simulations. We incorporate the idea of HST with three regression-based linkage approaches and with variance components (VC) linkage method to develop methods for identifying SNPs that explain a quantitative trait linkage signal. A simulation study was conducted to compare the newly proposed methods to a method implemented in the QTDT software. For identifying SNPs that partially explain a quantitative trait linkage signal, the method incorporating HST with VC was the most powerful, whereas for identifying SNPs that fully explain a quantitative trait linkage signal, the QTDT was the most powerful, at the cost of an inflated type I error rate under recessive models. Our study demonstrates the validity and flexibility of decomposing the IBD sharing from heterozygote and homozygote parents, an approach similar to the HST for binary traits, to identify SNPs explaining quantitative trait linkage evidence.

48

Handling Linkage Disequilibrium in Linkage Analysis with Dense Single Nucleotide Polymorphisms

K. Cho and J. Dupuis
Dept. of Biostat, Boston Univ., USA

As the density of genetic mapping escalates with rapid genotyping technical advances, it is pertinent to account for the underlying linkage disequilibrium (LD) in linkage analysis because ignoring LD results in spurious linkage results especially with ungenotyped parents. However, most multipoint linkage analysis methods assume linkage equilibrium and thus use incorrect estimates of haplotype frequencies when dense single nucleotide polymorphisms (SNPs) are considered where parental information is unavailable. Previous studies have observed logarithm-of-odds (LOD) score bias in affected sib-pair (ASP) analysis using dense SNPs. The resulting bias is due to apparent excess sharing of identity-by-descent (IBD) among sib-pairs when parents are ungenotyped. The effect of LD among dense

SNPs on quantitative trait linkage analysis has not been studied. In our study, we investigate the impact of LD among dense SNPs and theoretically demonstrate that an inflated estimate of IBD sharing causes inflation of a linkage statistic in ASP analysis. However, in quantitative trait linkage analysis, overestimation of IBD sharing estimate does not create the bias but reduces power. Based on our theoretical exploration, we propose and implement a two-step strategy to minimize the impact of LD on linkage results. We evaluate this two-step strategy through simulation and make recommendations on appropriate linkage analysis approaches and tolerable LD thresholds. Our research is pioneering in establishing a theoretical basis for the effect of LD on linkage statistics for both qualitative and quantitative traits.

49

Case-Control Association Testing in the presence of Unknown Relationships

Y. Choi (1), E.M. Wijsman (1,2,3), B.S. Weir (1,2)
(1) Dept. of Biostat, Univ. of Washington, USA, (2) Dept. of Genome Sci., Univ. of Washington, USA, (3) Division of Med. Genet, Dept. of Medicine, Univ. of Washington, USA

Genome-wide association studies result in inflated false positive results when unrecognized cryptic relatedness (CR) exists. A number of methods have been proposed for testing association between markers and disease with a correction for known pedigree-based relationships. However, in most case-control studies, relationships are generally unknown, yet the design is based on the assumption of at least ancestral relatedness among cases.
Here, we focus on adjusting CR when the genealogy of the sample is unknown, particularly in the context of samples from isolated populations where CR is more problematic. We estimate CR using maximum-likelihood methods and use a corrected chi-square test with estimated kinship coefficients, for testing in the context of unknown CR. Estimated kinship coefficients characterize precisely the relatedness between truly related people, but are biased for unrelated pairs.
The proposed test reduces substantially spurious positive results, producing uniform null distribution of p-values. Especially with missing pedigree information, estimated kinship coefficients can still be used to correct non-independence among individuals. The corrected test was applied to a real data set from a genetic isolate and created a distribution of p-value that was close to uniform. Thus the proposed test corrects the non-uniform distribution of p-value obtained with the uncorrected test and illustrates the advantage of the approach on real data.

50

Modeling Multistage Sampling of Family Data with Missing Information

Yun Hee Choi (1) and Laurent Briollais (1)
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

In family-based genetic studies, multistage sampling permits the allocation of resources to families that are most informative for a given objective while allowing population-based inference. A common problem in this design is the presence of missing genetic information among family members. To analyze multistage family designs, we propose a composite likelihood approach that we further extend to account for missing genetic information using an EM algorithm. We then show an application to a study of early-onset breast cancer among BRCA mutation carriers where data are collected from several cancer family registries using a multistage sampling.

51

Combined Genome-wide Linkage and Association Analysis of extended Utah prostate cancer pedigrees identifies significance at 8q12

G.B. Christensen, J. Farnham, N.J. Camp, L.A. Cannon-Albright
Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah

We performed genome-wide linkage and case/control association studies in 27 prostate cancer cases from 2 extended, informative, high-risk Utah pedigrees. All relationships between cases within pedigrees were more distant than first degree. Genotyping was performed with the Illumina 550k SNP set, after exclusion of 58,000 markers failing quality control. For controls, we selected caucasians from the Illumina Icontrol data set (n=1,579), also genotyped for the 550k SNPs.
Our initial screen for association included naive Fisher's Exact Test, ignoring the familial relationships between cases, under three models: dominant, recessive, and an allele test. 54 distinct markers were selected for secondary screening with a significance cut off of $p<1e-5$. Secondary screening was performed using Genie software, which included known relationships between cases. In the secondary screen 1 marker reached the genome-wide significance threshold of $p<3.4e-7$. This marker was on chromosome arm 8q12 (p=1e-7). In addition to providing the best overall GW association evidence, 5 of the top 8 associations from the secondary screening were also at 8q12; 9 SNPs in a 217kb region at chromosome 8q12 passed stage 1 screening. Other regions with markers reaching $p<3e-6$ included: 4 other markers at 8q12, 4p13, 2p25, 7p21, 17q22, and 21q21.
We also performed linkage analysis in the 2 pedigrees. We selected a set of 27,157 SNPs from the Illumina 550k set, with no evidence of LD, and used the Smith (1996) inheritance model. Two regions showed suggestive evidence of linkage; chromosome arm 2p (hetLOD=2.44) and chromosome arm 8q12-q21 (max hetLod=2.28). The SNPs showing significant evidence for association were in chr 8q12. This small study shows the power and synergistic utility of using both linkage and association analysis in high risk pedigrees. The 8q12 region identified as significant for prostate cancer predisposition has not been previously reported for linkage or association, but is recognized for LOH.

52

Case-control and family-based association studies on multiple candidate genes of hypertension and its endophenotypes, elevated triglyceride

C.M. Chung (1), H.B. Leu (2), J.W. Chen (2), W.H. Pan (1)
(1) Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, (2) Taipei Veterans General Hospital and Insititute of Clinical Medicine, National Yang-Ming University, Taipei, Taiwan

Plasma triglyceride (TG) is among major quantitative risk factors for hypertension, metabolic syndrome, stroke, and cardiovascular diseases (CVD). It is a well-established intermediate or sub-phenotype of CVD with heritability estimates ranging from 20% to 80%. Therefore, the purpose of this study is to find influential loci contributing to the etiology of hypertension through searching the quantitative loci of plasma TG. We used multiple candidate gene-based tag-SNP approach to identify quantitative trait loci of TG levels in 1337 individuals from 245 families of the young-onset hypertension study. 18 Tag-SNPs were selected from HapMap for APOA1/C3/A4/A5, *LPL*, *Ppar*γand *ACDC*, only two SNPs on *APOA5* gene were significantly associated with TG levels in a dose-dependent manner (GG > AG > AA for mean level of serum TG with p=0.00017 and $1.36x10^{-5}$ for rs2266788 and rs662799, respectively), using linear regression model with generalized estimating equations. We performed genotyping and TG measurements for additional 1023 pairs of young-onset hypertension and controls and was able to replicate the findings of these two SNPs ($p=3.05x10^{-7}$ and $3.67x10^{-11}$, respectively). In addition, we found that rs662799 of APOA5 was significantly associated with young-onset hypertension combined with elevated TG in the original study (OR=2.50, 95% CI=1.31- 4.76, p=0.0053) and in the replication study (OR=2.065, 95% CI=1.25-3.42, p=0.0047). Our study concludes that an *APOA5* locus but not those of *APOA1/C3/A4*, *LPL*, *PPAR*γ and *ACDC* was associated with TG level. Furthermore, the data suggest *APOA5* gene via TG metabolism may impact on the development of hypertension.

53

The Fundamentals of Allele Flipping in Association Studies

G.M. Clarke (1), L.R. Cardon (2)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, UK, (2) GlaxoSmithKline, Philadelphia, USA

Replication of initial findings in an independent sample is the gold standard for confirmation of a genuine disease-marker association. Replication is generally taken to mean association of the same SNP with the same direction of effect. Reports of allele flips, where an initial study finds an allele to be protective but a follow-up study finds it to be causative, are increasing. It is therefore of practical interest to determine when an allele flip is genuine, that is, when a high risk allele in an original study is, in fact, the low risk allele in a replication study. Allelic heterogeneity, locus heterogeneity, variation in environmental exposures and population differences are all examples of factors that can combine to create scenarios for a genuine allele flip. Instead of examining various scenarios in turn, we instead identify the common underlying parameters that must be affected in order to trigger an allele flip. We show that unless the sign of the mean of the distribution of the association test statistic varies between studies, the probability of observing an allele ?ip at a genuine causal locus in samples ascertained similarly from a common population is negligible. When the sign of the mean is reversed between studies, the probability of an allele ?ip increases directly with the power of the studies. We derive expressions for the mean of the odds ratio test statistic under common models that illustrate clearly how the behaviour of key model parameters impacts the probability of a genuine allele flip Using HapMap data, and r rather than r2 to highlight previously

54

A Fast-search Algorithm to Identify Communities of Interacting SNPs in GWAS Studies

Sharlee Climer (1), Lisa de las Fuentes (2), Victor G. Dávila-Román (2), C. Charles Gu (1,3)
(1) Division of Biostatistics, (2) Cardiovascular Imaging and Clinical Research Core Laboratory, Cardiovascular Division, Department of Medicine, (3) Departments of Genetics; Washington University in St. Louis, MO, USA

Most genome-wide association (GWAS) studies have analyzed the thousands SNPs individually without considering SNP–SNP interactions, probably due to the computational burden. However, although the total number of all possible interactions is prohibitively large, in real data we have almost always encountered large but sparse networks containing a number of relatively small, dense communities of interacting SNPs. In the present study, we present a novel method for finding communities in large and sparse networks. We propose a fast algorithm to identify clustering of pairwise SNP–SNP interactions that may reflect underlying biological functions. Our method deals with a critical issue of the problem: it requires a method that can scale to more than 1,000,000 vertices and be fast enough to enable the computation of hundreds of bootstrapping trials. We use a "divide-and-conquer" technique that efficiently reduces the size of the problem into small pieces that can be encoded into integral values and used to look up significant interactions in a pre-computed table. The evaluation for potential interaction between a given pair of SNPs can be terminated when the significance level of the test falls below a given threshold. The high speed of this method has allowed us to tackle a real dataset from an interim GWAS study of left ventricle hypertrophy, and identify unique and denser interacting SNP clusters than existing methods.

55

Hierarchical stochastic search with prior information

N. Cremer (1), L. Beckmann (1), J. Chang-Claude (1), D.V. Conti (2)
(1) Dept. of Cancer Epidemiology, DKFZ, Heidelberg, Germany (2) Dept. of Prev. Med., USC, Los Angeles, CA

Hierarchical modeling allows for incorporation of various types of information by modeling estimates from a first level regression on prior information. A recently proposed approach expands upon this by allowing prior information to guide a stochastic model search. To investigate performance, a simulation study was conducted consisting of ten

independent exposures; seven not conferring risk and the remaining three being associated with disease status. Of primary interest is the impact of the prior structure on inference. Thus, we investigated three different priors comprising of (1) treating all exposures equal, (2) using perfect information, and (3) random knowledge about association status with disease. Results were based on 1,000 replicates. We present both posterior estimates of test statistics and probabilities of variable inclusion. For empirical power we use the number of Bayes Factors (BFs) ≥ 3 across all replicates. Clear distinction in BF values between risk and non-risk exposures was seen in all scenarios with perfect information providing the strongest evidence. Comparison of empirical power indicates a clear separation for equal and perfect information. There were slightly elevated false positive rates for random knowledge, with empirical power only weakly impacted. While correct information clearly provides the most advantage, incorporating even poor information still separates exposures reasonably well. Comparison to alternative approaches and increased number of both risk and non-risk exposures requires further investigation.

56

Meta Genetic Association of Colorectal Cancer and SNPs at 8q24, 9p24, and SMAD7

K. Curtin (1), W.-Y. Lin (2), R. George (2), M. Katory (2), J. Shorto (2), D.T. Bishop (3), A. Cox (2), N.J. Camp (1)
(1) Univ. of Utah School of Medicine, USA, (2) Sheffield Medical School, UK, (3) Leeds Inst. of Molecular Medicine, UK

Genomewide association studies of colorectal cancer (CRC) have identified SNPs that have subsequently been studied as candidate loci. Associations of 7 SNPs in the 8q24 candidate region, 2 SNPs in the 9p24 region, and 3 SNPs in SMAD7 and CRC were investigated in two UK case-control cohorts (Sheffield and Leeds, Colorectal Cancer Study Group) and family-based cases in high-risk Utah pedigrees and matched controls (1132 cases and 1092 controls in total). Meta-statistics and Monte Carlo significance testing using Genie software provided valid analyses of family-based and independent resources. Similar to other reports, the 8q24 rs6983267 high-risk allele was associated with increased risk of CRC overall (GG vs. TT, OR 1.4, 95%CI 1.1-1.8) and in cases with age at onset <60 years (OR 1.6, 95%CI 1.03-2.4). In cases without a first-degree family history, associations with rs6983267 and rs1050477 were observed (p-trends .02 and .03, respectively). Only rs10090154, located in a different designated risk locus, was associated in cases with a family history (TT vs. CT/CC, OR 3.2, 95%CI 1.04-9.7). In three SMAD7 SNPs, associations were seen in distal colon tumors, but not in proximal or rectal cancers (rs4939827 CC/CT vs. TT, OR 0.7, 95%CI 0.5-0.9; rs12953717 TT vs. CC, OR 1.6, 95%CI 1.1-2.3; rs4484148 CC/CT vs. TT, OR 1.3, 95%CI 1.1-2.3). The association remained significant in a case-case comparison of distal vs. proximal/rectal cancers for rs4939827 and borderline significant for rs12953717. We were unable to confirm any association with 9p24 SNPs rs719725 or rs7857826 in CRC. Our investigation confirms that polymorphisms in 8q24 are associated with CRC and that these associations differ between LD sub-regions, and offers support for SMAD7 markers and an association with distal colon cancer.

57

Communicating Results from International Collaborations in the Information Age: Introducing the Genomic Applications for Humanity (*Genapha*) Website

D. Daley (1), D. Zamar (1), B. Tripp (1), M. Lemire (2), T.J. Hudson (2), and P.D. Paré (1)
(1) James Hogg iCAPTURE Center, University of British Columbia, Vancouver, BC; (2) Ontario Institute for Cancer Research, Toronto, Ontario; Canada

In order for research to have impact it must first be successfully communicated. In an era of high throughput genome wide association and candidate gene studies how do we translate the results from international genomic consortiums effectively? Only a small number of results are published, generally only statistically significant observations, leaving the vast majority of results undisclosed. To facilitate meta-analyses and the identification of genes with smaller effect sizes (OR <1.4) we are releasing association results for 98 candidate genes from our study that includes 5,565 individuals recruited into four studies from Canada and Australia. One way to communicate findings is to utilize the World Wide Web and we present *Genapha* (www.genpaha.ca) as a model for knowledge translation from high throughput genomic platforms. *Genapha* is an online database of results, providing rapid, full, and open disclosure of our research findings to the scientific community. The database has been interfaced with dbSNPs, OMIM, PubMed, and the semantic web. We have developed search engines to allow users to rapidly filter information and identify genes, SNPs, and association results most pertinent to their own research, not just from our study, but the entire body of knowledge that is available on the web. Detailed descriptions of the study designs, populations, and allele frequencies along with tutorials are hosted on the website.

58

Genetic contributions to Longevity

E.W. Daw (1), C. Kammerer (2), J.H. Lee (3), M. Feitosa (1), J. Zmuda (2), I. Borecki (1), S. Barral (3), E. Hadley (4), K. Christensen (5), R. Mayeux (3), M. Province (1)
(1) Washington Univ., St. Louis MO, (2) Univ. of Pittsburgh, PA, (3) Columbia Univ, New York NY, (4) NIA, Bethesda MD, (5) Univ. of S Denmark

Longevity is a complex phenotype with contributions from many genetic and environmental factors. The Long Life Family Study (LLFS), an international study, has recruited families based on familial clustering of longevity, with the goal to identify factors protecting against aging changes. Probands are ≥ 89 for men, ≥ 92 for women. Demographic data is collected on the whole family and clinical measures are made on the proband, siblings, spouses and children. To assess the genetic contribution to Longevity (age at death or age at exam as a censoring value), as well as clinical traits from three domains (cognitive, cardiovascular, and metabolic) associated with longevity, we estimated the heritability (h^2)

and carried out Oligogenic Segregation (OS) analyses for longevity. h^2 estimates relative contribution of genetic and environmental factors, while OS estimates number of genes contributing to trait variation, relative importance of those genes, and mode of inheritance of each gene. The h^2's ranged from ~0.2 to ~0.7. Number of genes found by OS did not correlate completely with h^2, but the number of trait genes with the highest posterior probability ranged from 7, for longevity itself, to 2-3 for measures in the cognitive domain. All traits examined were found to have some genetic component, although environment may be more important for some. These results suggest LLFS data will help identify genes contributing to longevity and each of the associated domains.

59

A Family-based Association Test for Quantitative Traits to Detect Gene-Gene Interactions

L. De Lobel (1), H. De Meyer (1), L. Thijs (2), T. Kouznetsova (2), J. Staessen (2), K. Van Steen (3)
(1) Department of Applied Mathematics and Computer Science, Ghent University, Belgium
(2) Division of Hypertension and Cardiovascular Revalidation, Department of Cardio-vascular Diseases, KULeuven, Belgium
(3) Department of electrical engineering and computer science, University of Liege, Belgium

For many complex diseases, quantitative traits contain more information than dichotomous traits. One of the approaches used to analyse these traits in family-based association studies is the Quantitative Transmission Disequilibrium Test (QTDT). The QTDT is a regression-based approach that contains a test for both linkage and assocation. It splits up the association effects in a between-family and a within-family component to adjust and test for population stratification. Furthermore, a variance components method is included in the model to be able to test for linkage. We extend this approach to test for gene-gene interactions in the family-based setting by adjusting the definition for the between-family component, the within-family component and the variance components included in the model. Different epistatic models with and without main effects are simulated to investigate the power of the approach. Simulated data under the null are used to investigate type I error rates. The proposed method is also applied to a real-life dataset on hypertension.

References:

[1] G.R. Abecasis, L.R. Cardon and W.O.C. Cookson, 2000. A General Test of Association for Quantitative Traits in Nuclear Families, American Journal of Human Genetics, no. 66: 279–292.

60

Can Novel Apo A-I Polymorphisms be responsible for Low HDL in South Asian Immigrants?

S. Dodani (MD, PhD, FCPS, MSc, FAHA), Y.B. Dong (PhD), H. Zhu (PhD)
Medical College of Georgia, Augusta, GA

Abstract: Coronary artery disease (CAD) is the leading cause of death in the world. Even though its rates have decreased worldwide over the past 30 years, event rates are still high in South Asians. South Asians are known to have low high density lipoprotein (HDL) levels. The objective of this study was to identify Apolipoprotein A-I (Apo A-I) polymorphisms, the main protein component of HDL and explore its association with low HDL levels in South Asians. A pilot study on 30 South Asians was conducted and 12 hour fasting samples for C-reactive protein, total cholesterol, HDL, low-density lipoprotein (LDL), triglycerides, Lipoprotein(a), Insulin, glucose levels, and DNA extraction and sequencing of Apo A-I gene were done. Results Apo A-I polymorphisms revealed six novel single nucleotide polymorphisms (SNPs) one of which (C938T) was significantly associated with low (<40 mg/dl) HDL levels (p=0.004). The association was also seen with total cholesterol (p=0.026) and LDL levels (p=0.032). This pilot work has highlighted some of the gene-environment associations that could be responsible for low HDL and may be excess CAD in South Asians. Further larger studies are required to explore and uncover these associations that could be responsible for excess CAD risk in South Asians.

61

Association of Variations in Inflammation-Related Genes with Susceptibility to Major Depression and Antidepressant Effect of Desipramine and Fluoxetine

C.-H. Dong, M.-L. Wong, J. Licinio
Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL33136, USA

Clinical parallels between the nature and course of depressive symptoms and inflammation suggest the activation of the immune system in major depressive disorder (MDD). To examine the association between polymorphisms in inflammation-related genes and the susceptibility to major depressive disorder (MDD) and antidepressant response, we investigated a panel of single nucleotide polymorphisms (SNPs) focused on the steroid pathway and proteasome subunit in 278 MDD patients of Mexican descent and 281 ethnicity-matched healthy controls. Significant association was found between MDD and SNPs in two genes that are critical for T cell function: rs2296840 (OR: 1.7, 95% CI: 1.3-2.1, Benjamini-Hochberg FDR<0.01) in proteasome beta 4 subunit gene (PSMB4) and rs17244587 (OR: 2.0, 95% CI: 1.4-2.8, Benjamini-Hochberg FDR<0.01) in T-box 21 gene (TBX2). The combined effect analysis of PSMB4 and TBX21 genetic variants on MDD yielded a Rothman synergic index of 1.26. Drug response analyses revealed that five SNPs in the steroidal pathway and proteasome genes were associated with clinical response (at least 50% reduction of 21-item Hamilton Depression Rating Scale) within the entire depressed group treated with desipramine or fluoxetine. Our findings support the implication of inflammation-related genes in the susceptibility to major depression and a role in antidepressant response. Further studies in independent samples are clearly warranted for replication and validation.

62

SNP Linkage Scan of Glaucoma Related Traits in the Beaver Dam Eye Study

Priya Duggal (1), Alison Klein (2), Ching-Yu Cheng (1), Kris Lee (3), Ronald Klein (3), Barbara Klein (3), Joan E. Bailey-Wilson (1)
(1) NHGRI/NIH, (2) Johns Hopkins University, (3) University of Wisconsin

Primary open-angle glaucoma (POAG) is a leading cause of blindness in the world. However, because there is not a uniform set of guidelines for POAG, it is often defined differently among epidemiological studies. The variability in definition of POAG decreases the ability to identify a causal gene. Two early better defined quantitative markers, intraocular pressure and cup-disc ratio, are strong risk factors for disease, and may help to better elucidate the causal genes. Previous familial correlations, segregation analysis and STRP linkage suggest a genetic component to these traits. We performed a genome wide scan of these POAG related traits using 486 pedigrees from the Beaver Dam Eye Study. Genotyping of 6008 SNPS on the Illumina linkage panel was completed at the Center for Inherited Disease Research. Linkage analysis was performed using the modified Haseman-Elston regression models in SIBPAL (SAGEV5.0), after removing linkage disequilibrium. We will present regions of linkage for these traits.

63

As genetic epidemiology looks beyond mapping single disease susceptibility loci, interest in detecting epistatic interactions between genes has grown. The dimensionality and comparisons required to search the epistatic space and the inference for a significant result pose challenges for testing epistatic disease models

The Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) was developed to test for multilocus models in pedigree data. In the present study we rigorously tested MDR-PDT with new cross-validation (CV) and omnibus model selection algorithms by simulating a range of heritabilities, odds ratios, minor allele frequencies, and numbers of interacting loci. Additionally, given that the permutation-based hypothesis test of the MDR-PDT does not evaluate effect modification across genotypes and that this property might inflate the type I error rate for the null hypothesis of no interaction, we chose to implement a regression-based permutation test.
We found that MDR-PDT performs similarly with 5 and 10 fold CV. Also, the sensitivity of MDR-PDT to select best models using the omnibus approach was not extremely less than the n-locus approach. We also demonstrate that fitting a regression model on the same data as analyzed by MDR-PDT is a biased procedure and is not a valid test of interaction. The regression-based permutation test implemented here conducts a valid test of interaction after a search for multilocus models, and can be used with any method that conducts a search to find a multilocus model representing an interaction.

64

Genome-wide Case Control Association Study for Familial Melanoma identifies 2 significant associations

J. Farnham (1), S. Leachman (2), N.J. Camp (1), L.A. Cannon-Albright (1)
(1) Departments of Biomedical Informatics and (2) Dermatology, Salt Lake City, Utah

We performed a genome-wide case control association study in 87 melanoma cases from 21 informative, high-risk Utah pedigrees. All relationships within pedigrees were more distant than first degree. Genotyping was performed with the Illumina 550k SNP set, after exclusion of 58,000 markers failing quality control and 14,000 markers on the X chromosome (not analyzed). For controls, we selected caucasians from the Illumina Icontrol data set (n=1,579); all were genotyped for the 550k SNPs. Our initial screen for association included naive Fisher's Exact Test, ignoring the familial relationships between cases, under three models: dominant, recessive, and an allele test. 49 distinct markers were selected for secondary screening that surpassed a significance threshold of $p < 1e\text{-}5$. Secondary screening was performed using Genie software, which accounts for the known relationships between cases.
In the secondary screen 3 markers in 2 regions reached the genomewide significance threshold of $p < 3.4e\text{-}7$. The most significant marker was on chromosome arm 6q22 (p=5e-8). The 2 other markers were both on chromosome arm 16p13, located in A2BP1/FOX1 (p=2e-7 for each). A2BP1, ataxin-2 binding protein-1 (A2BP1) gene, also called FOX1. Other regions with markers reaching $p < 5e\text{-}6$ included: chromosome arms 2q12, 2q24, 2q32, 8p23, 12p13, 17p11, 18q22, and 20q13 (20Mb from the recently reported association at 20q11). This study represents the first complete genome wide association study reported for familial melanoma.
This study shows the power and utility of association analysis in related cases from high-risk pedigrees. Two significant regions of association were identified, neither of which has been previously reported.

65

Association of promoter LIPC variants with fat distribution and HDL level: The NHLBI Family Heart Study (FHS)

M.F. Feitosa (1), P. An (1), S. Ketkar (1), R.H. Myers (2), J.S. Pankow (3), M.A. Province (1), I.B. Borecki (1)
(1) Washington Univ, Saint Louis, MO, (2) Boston Univ, Boston, MA, (3) Univ. Minnesota, Minneapolis, MN, USA

Obesity is associated with a lower level of high density lipoprotein cholesterol (HDL), which is a risk factor for coronary artery disease. Whether body fat distribution, specifically centralized obesity, affects HDL level remains unclear.Hepatic lipase plays a major role in HDL metabolism, and promoter variants in the gene (LIPC) have been associated with HDL level. Previously, we genotyped 19 tag-SNPs in FHS (2238 subjects, 591 families) and found association between rs261342 (an intronic marker in LD with the LIPC promoter region) and HDL (p=0.00067) and waist-to-hip ratio (WHR, p=0.01991). Recently, a subset of ~1,000 largely unrelated subjects was typed

with the Illumina 550K chip, and we investigated 110 tag-SNPs across 264 kb of LIPC, including the promoter. We found strong association of four SNPs in the promoter region with WHR (rs7172930, p=0.000034; rs1711050, p=0.000507; rs421705, p=0.000354; rs426684, p=0.000908), which are in moderate LD (0.42?LD?0.55). The level of HDL was associated with a promoter LIPC variant (rs1077834, p=0.000051), not in LD with the cluster of four ($r2 < 0.13$). The minor alleles of rs7172930, rs1711050 and rs421705 are associated with reduced WHR levels, while those of rs426684 and rs1077834 are associated with increased WHR and HDL, respectively. This is the first report of LIPC variants influencing both centralized fat patterning and HDL level, suggesting that the correlation between them may be in part mediated by LIPC.

66

A Two-level Modeling Strategy to Investigate Bias and Heterogeneity in Association Studies of Candidate Gene Pathways

J. Figueiredo (1), J.P. Lewinger (1), J. Poynter (1), D. Lee (1), K. Siegmund (1), D. Duggan (2), J. Potter (3), J. Baron (4), R. Haile (1), D. Conti (1)
(1) Dept. Prev Medicine, University of Southern California, LA, (2) Translational Genomics Research Institute, AZ, (3) Fred Hutchinson Cancer Research Center, WA, (4) Dartmouth Medical School, NH

Several large consortia are conducting association studies investigating hundreds of SNPs in candidate gene pathways and cancer risk. The variety of different designs and ascertainment strategies sometimes used within a single study can introduce heterogeneity. Study conclusions can be affected if this heterogeneity is not taken into account.
We propose a simple two-model strategy that takes advantage of the large number of SNPs investigated to formally test for heterogeneity in effect estimates across subgroups of subjects. In a first stage, we regress the outcome on each SNP to obtain effect estimates and their standard errors within a subgroup of subjects believed to be homogeneous. In a second stage, we assess heterogeneity across the subgroups by comparing the estimates derived in the first-stage using random effects models. We apply our two-stage approach to a case-control study of colorectal cancer and genes involved in folate metabolism to test for heterogeneity resulting from inclusion of cases recruited within varying lengths of time after diagnosis. We found no evidence that this introduces a systematic heterogeneity in estimated RR's across all SNPs. However, we found evidence that the underlying risk for selected SNPs may be heterogeneous.

67

Genotype-Environment Interaction Influencing Blood Pressure: The Strong Heart Family Study

N. Franceschini (1), K.E. North (1), S. Rutherford (2), S.A. Cole (2), J.W. MacCluer (2), L.G. Best (3), E.T. Lee (4), B.V. Howard (5), K.A. Rose (1)
(1) UNC, NC, (2) SFBR, TX, (3) MBTL, SD, (4) U Oklahoma, OK, (5) MedStar, DC

In an era of genome wide association studies, little progress has been achieved in identifying common blood pressure [BP] susceptibility alleles, partly due to naive study designs that ignore the role of social, behavioral, and environmental factors in BP regulation. We studied the interaction of additive genetic effects and behavioral (physical activity [PA], smoking, alcohol use) and socioeconomic (education and income) factors on systolic BP (SBP) in American Indians (AI) in the Strong Heart Family Study. We studied 1,894 participants, with previously identified linkage to SBP, recruited from Arizona, North/South Dakota, and Oklahoma. Resting BP was measured three times and the last two measures averaged. The mean SBP was 123 ± 17 mmHg. Fifty eight percent of AI were ever smokers and 58% reported current alcohol intake. The median PA level, quantified by a pedometer, was 4,813 steps/day. Thirty-five percent of individuals had less than a high school education and 69% had annual household incomes $<$\$25,000. Using variance component models (SOLAR), we detected evidence for distinct genetic effects on SBP among ever compared to never smokers. The additive genetic variance of SBP was 0.37 and 0.63 for ever and never smokers, respectively [P=0.001]. For alcohol intake, we detected evidence for distinct genetic effects on SBP among current drinkers compared to former or never drinkers [rhoG=0.44, P=0.004]. Lastly, we detected evidence for distinct genetic effects on SBP among individuals with incomes $<$\$25,000 versus those with income of \$25,000.00 or more [P=0.01]. Our findings suggest that behavioral and socioeconomic factors can modify the genetic effects on BP variability. Therefore, accounting for these interactions may help us better dissects the complexities of the gene effects on BP.

68

Combining linkage results for cohorts genotyped on different SNP panels

B.L. Fridley (1), W. Bamlet (1), D. Serie (1), J.D. Potter (2), and E.L. Goode (3)
(1) Div. of Biostatistics, Mayo Clinic, Rochester USA
(2) Div. of Public Health Services, Fred Hutchinson Cancer Research Center, Seattle USA
(3) Div. of Epidemiology, Mayo Clinic, Rochester USA

With continually-advancing genotyping platforms and the need for increasing sample size, analysts seek to combine linkage data from multiple studies genotyped on differing SNP panels. For example, in a study of colorectal cancer families collected within the Colon Cancer Family Registry (Colon CFR), approximately 300 families have been genotyped on the Affymetrix Mapping 10K 2.0 Array, and 200 families will be genotyped using the Illumina Infinium HumanLinkage-12 BeadChip. One method to combine the linkage results is via a meta-analysis. An alternative, potentially more-powerful approach, is a pooled analysis. Pooled parametric linkage can be done by mapping SNP sets to a common genetic map, running parametric linkage of each family set separately using a common interval, and simply summing LOD scores across the studies. However, for nonparametric linkage analysis, in which the Kong and Cox LOD score is computed, this approach is no longer appropriate, and a single nonparametric linkage analysis is

needed. Using Affymetrix chromosome 2 data on 300 Colon CFR families, we evaluated methods for pooled analysis by splitting families randomly into two groups with complete missing data on 50% of SNPs. We accounted for LD in both parametric and non-parametric analyses. We will present results from both parametric and nonparametric linkage methods which show the utility of the pooled approach in combined linkage analysis.

69

A Copy Number-Based Robust Method for SNP Genotype Calling

W.J. Fu (1), L. Wan (2), K. Sun (1), Q. Ding (3), Y. Cui (4), R.C. Elston (5) and M. Qian (2)
(1) Dept. of Epid, Michigan State Univ, USA, (2) Dept. of Math, Peking Univ, China, (3) Dept. of Biochem, Michigan State Univ, USA, (4) Dept. of Stat, Michigan State Univ, (5) Dept. of Epid and Biostat, Case Western Reserve Univ, USA

Single nucleotide polymorphism (SNP) genotype calling is of great importance for genome-wide association studies. Although a number of genotype calling methods have been studied, many of them rely on multi-array training and depend on probe intensity distributions and so may not be robust for cross-laboratory studies. In this paper, we describe a novel approach to genotype calling with Affymetrix high density SNP arrays. The method incorporates array probe sequence structure, which is invariant within the Affymetrix microarray technology, and utilizes efficiently both perfect match and mismatch probes. The binding free energy and binding affinity between probe and target sequences is modeled, and a probe intensity composite representation (PICR) model is built for the intensities of both perfect match and mismatch probes with copy numbers and binding affinity. This PICR model provides a regression model for allelic copy number estimation at each SNP, yielding a novel method for SNP genotype calling through a statistical decision based on allelic copy numbers. Using the HapMap samples, we demonstrate that this method 1) is robust for cross-laboratory studies and for studies using different array prototypes, 2) has zero no-calls and thus generates no missing data for SNP genotyping, and 3) consistently achieves high accuracy on high density Affymetrix 100K and 500K SNP arrays.

70

Consistency replicating locus linked to plasma Factor XII activity on 5q33 and identification of novel locus on 8q24

F. Gagnon (1), G. Antoni (1), A. Tuite (1), Y. Luo (1), D. Bulman (2), P. Wells (2)
(1) Univ. of Toronto, (2) OHRI, Canada

Purpose: Venous thromboembolism (VTE) is one of the most tractable common complex disease for prevention purpose i.e, well-characterized VTE-associated phenotype (APCR) and genetic variant (Factor V Leiden - F5L), along with known environmental determinants. Despite that, VTE risk prediction remains limited. Factor XII activity (FXII), which exhibits high heritability and positive correlation with VTE, is a potential therapeutic target with reduced bleeding risk. Identifying FXII Quantitative Trait (QT) loci will provide clues of the underlying mechanisms, as well as better risk profiling opportunities. The main study objective is to identify the QT loci underlying the variation in FXII.
Methods: In the framework of our "Family Study on F5L Thrombophilia", we measured ~30 hemostatic/lipid QT, along with covariates, in members (n=262) of 5 French-Canadian families ascertained through a proband with both VTE and F5L. We did a genome-wide oligogenic linkage analysis (1079 microsatellites) of FXII, based on Bayesian MCMC methods and covariate adjustments.
Results: We provide evidence for 2 strong QT loci located on chromosome (ch) 5 and 8. The ch5 locus corresponds to the F12 gene, which has been implicated in variation of FXII in the GAIT (Genetic Analysis of Idiopathic Thrombophilia) study. The ch8 locus (8q24) is novel. Adding F5L in the model did not modify the signal on ch5 but reduced it on ch8, suggesting that the later interacts with F5L. We are also investigating pleiotropic effects with other QT/VTE, as well as epistasis.
Conclusion: Using innovative ascertainment and analytic strategies, we localized strong linkage for FXII, demonstrating that with a carefully thought out design, it is possible to detect linkage with a relatively small sample. Attempts to replicate the loci will be done in 2 independent French samples with complementary designs.

71

Heritability of alcohol and nicotine dependence in Mongolian adults

J. Ganchimeg, S.I. Cho, H.J. Kim, J.I. Kim, H.S. Park, H.L. Kim (1), J.H. Sung, J.S. Seo, and The Gendiscan Study Group
Seoul National University, Republic of Korea
(1) Ewha Women¡¯s University, Republic of Korea

Background: Alcohol and tobacco consumption among the Mongolians have increased with social transformations since 1990s. We assessed the heritability of alcohol and nicotine dependence in Mogolians whose environment has been recently changing.
Methods and materials: This study is based on the Gendiscan Study, a community-based family study in Mongolians. The sample of adult Mongolians (older than 16) consisted of 95 large families (individual size=1,642), with mean age 35.93¡¾15.21. All participants were native Mongolians born and grown up in the same rural area. Alcohol and nicotine dependence were measured by standard questionnaires. Heritability was calculated by S.A.G.E 5.3 and adjusting for sex and age.
Results: Among men, 23% smoked and 14% regularly drank. In women, 2% smoked and 1% regularly drank. Alcohol dependence (n=69) in 46 family (n=997) and nicotine dependence (n=61) in 41 family (n=918) were found. Heritability of dependence was higher for alcohol (25.17%¡¾6.62%) than for nicotine (22.90%¡¾1.97%). After adjustment, heritability decreased to 19.04%¡¾5.15% for alcohol and 9.17%¡¾4.65% for nicotine dependence, and both remained statistically significant despite their low values compared to previous studies.
Conclusion: There is significant genetic influence in alcohol dependence and nicotine dependence. Changing environ-

ment and measurement errors may contribute to underestimation of genetic influence measured by heritability.

72

A generalized sequential Šidák procedure for multiple hypothesis testing

Guimin Gao and Guolian Kang
The University of Alabama at Birmingham, AL 35294, USA

Multiple hypothesis testing is a common problem in genome research, such as genome-wide studies and gene expression data analysis. Rubin et al. and Wasserman and Roeder proposed a weighted Bonferroni procedure. This method can control family-wise error rate (FWER) and can have much higher power than the widely used Bonferroni procedure when given the means of the test statistics or some prior information. To further increase the power of the weighted Bonferroni procedure, in this paper we propose a weighted Šidák procedure that is an extension of and has slightly higher power than the weighted Bonferroni procedure. Furthermore, we develop a generalized sequential Šidák procedure which uses weighted p-values and is an extension of the generalized sequential Bonferroni procedure of Holm. Under the assumption that the means of the test statistics in the multiple testing are known, we incorporate the optimal weights calculated by the weighted Šidák procedure (the weighted Bonferroni procedure) into the generalized sequential Šidák procedure (the generalized sequential Bonferroni procedure). Simulation studies show that the generalized sequential Šidák procedure has slightly higher power than the generalized sequential Bonferroni procedure, and has much higher power than the weighted Šidák procedure and the weighted Bonferroni procedure. All proposed procedures can control FWER. The proposed methods are useful when prior some information is available to estimate the means of the test statistics.

73

Predicting a binary outcome using SNPs: a case study on prediction of longevity

W. Ghidey (1), T. Stijnen (1), J. Houwing-Duistermaat (1), B. Heijmans (1), M. Beekman (1), R. Westendorp (2), E. Slagboom (1), H.C. van Houwelingen (1)
(1) Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, the Netherlands, (2) Department of Gerontology and Geriatrics, Leiden University Medical Center, the Netherlands

In genetic association studies, a large number of single-nucleotide polymorphisms (SNPs) are typed in a sample of cases and control for the purpose of identifying genes associated with a specific phenotype. With complex phenotypes, the SNPs usually act in combination (interaction effect) that an individual SNP may not be important by itself. In such cases, the identification of the best predictive model of the outcome in terms of a specific combination of a large number of SNPs is a statistical challenge. In this paper, we aim at finding a best predictive model of a complex trait (longevity) in terms of specific combinations of a number of SNPs. We explored different better alternatives to the standard logistic regression model. We evaluated the predictive power of each method and at a later stage further improvement of the predictive power was explored through combining the individual model predictions into a single model. The model is constructed as a weighted combination of individual prediction models, where the weights are determined such that a cross-validated error is minimized. We illustrate the application of the different alternative prediction models as well as that of the combination model with longevity data of the Leiden 85-plus study.

74

A Quantile-based Test Of Allelic Association For Analyzing Population-based Quantitative Trait Data

S. Ghosh, A. Ghosh
Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Although statistical methods based on case-control designs are the most popular and extensively used approach for genetic association mapping of binary traits, development of such methods for quantitative traits is currently an active area of interest. While some novel family-based methods of association have been developed for quantitative traits, population-based quantitative trait data have usually been analyzed using classical analysis of variance (ANOVA) methods. However, ANOVA is valid in a strict statistical sense only under the assumption of equality of variances in each underlying group. On the other hand, the assumption of equality of variances of the quantitative traits for the different genotypes at a QTL is genetically unrealistic, particularly if the trait is correlated with some disease outcome. Thus, it is of interest to explore for model-free alternatives which would circumvent this problem. We propose a quantile-based method to test for allelic association. The basic paradigm of the method is that a marker allele in linkage disequibrium with an allele at the QTL would have either a strictly increasing or a strictly decreasing frequency distribution across the range of quantitative trait values. We perform Monte-Carlo simulations for different genetic parameters and find that our proposed method is more powerful than ANOVA. Since case-control analyses for binary traits are known to be susceptible to population stratification, we also assess the relative effects of different degrees of population stratification on the proposed method and ANOVA.

75

Establishing Equivalence with Hardy Weinberg Equilibrium

K.A.B. Goddard (1), A. Ziegler (2), and S. Wellek (3)
(1) Center for Health Research, Kaiser Permanente Northwest, Portland, OR; (2) Institute of Medical Biometry and Statistics, University at Lübeck, Germany; (3) Division of Biostatistics, CIMH Mannheim/University of Heidelberg, Germany

Tests of Hardy-Weinberg Equilibrium (HWE) are conducted for many reasons, including as a quality control measure in genetic association studies using unrelated individuals. The interpretation of the test result is difficult because departure from HWE can occur for other reasons including proximity

to a disease locus in samples selected on phenotype (i.e., case-control studies), and multiple testing. Also, the standard goodness-of-fit (GOF) test to assess HWE is fundamentally flawed with respect to its logical basis, because it is only tailored to establish departure from HWE whereas the aim is to establish (approximate) compatibility between an observed genotype distribution and HWE. We present a logically unflawed solution to this problem using equivalence testing, and show that the new method provides exact control of the type I error risk and the power is $>80\%$ for a wide range of allele frequencies when the sample size exceeds 200. We illustrate the method using genotype distributions from 43 candidate gene studies and 2 genome-wide association studies. The conclusions of the two tests are the same for 70% of the samples. The discrepancies are explained by either a small sample size, or a large sample size and statistically significant, but unimportant, deviations from HWE using the GOF test. The new approach provides more satisfactory assessment of compatibility with HWE, especially when used in genome-wide association studies.

76

Increased risk of monoclonal gammopathy of undetermined significance (MGUS) and lymphoid tumors among first-degree relatives of MGUS cases

L.R. Goldin (1), S.Y. Kristinsson (2), N.E. Caporaso (1), M. Bjorkholm (2), I. Turesson (3), O. Landgren (1)
(1) National Cancer Inst, Bethesda,MD (2) Karolinska Inst,Stockholm, SW (3) Malmo U. Hospital, Malmo, SW

MGUS is a generally asymptomatic plasma-cell disorder with an elevated monoclonal immunoglobulin of less than 3 g/dl in the absence of a lymphoproliferative (LP) malignancy. MGUS is a precursor to multiple myeloma (MM) and other lymphoid tumors, transforming at the rate of approximately 1% per yr. Genetic factors have been shown to be important for MM and other LP tumors but the genetic relationship to the precursor trait is not known. We identified 4488 MGUS cases diagnosed in major hematology outpatient units in Sweden (1967-2005) with linkable relatives. Using the population-based Multigenerational Registry, we obtained 17,628 controls and first-degree relatives of cases (n=14689) and controls (n=58698). Relatives were linked with hospital outpatient registries and the Cancer Registry to define occurrence of MGUS and other LP tumors. We applied a marginal survival model with a sandwich covariance estimator to take into account familial dependencies. Relatives of MGUS cases were at significantly increased risk for MGUS (HR=2.84,1.45-5.57), MM (HR=2.87,1.92-4.27), Waldenstrom macroglobulinemia (HR=4.94,1.32-18.46), and chronic lymphocytic leukemia (CLL) (HR=2.05,1.22-3.43) but not for other lymphomas. Thus, shared genes likely contribute to risk of MGUS and related LP tumors, MGUS being an early genetic lesion in the pathway to malignancy. Our candidate gene studies show possible associations of LP tumors with apoptosis genes.

77

Family-based association testing of colorectal cancer risk in the 8q24 and 9p24 candidate regions

E.L. Goode (1), L. Le Marchand (2), B.L. Fridley (1), W. Bamlet (1), D. Serie (1), J.D .Potter (3)
(1) Mayo Clinic College of Medicine, Rochester, MN, USA; (2) University of Hawaii, Honolulu, HI USA; (3) Fred Hutchinson Cancer Research Center, Seattle, WA USA

Genome-wide association studies indicate that chromosome 8q24 contains risk alleles for colorectal, prostate, and breast cancer and 9p24 for colorectal cancer. We assessed whether analysis of a sparse SNP linkage panel may elucidate associations in these regions using family-based analysis methods. In 321 White families of the Colon CFR with ≥ 2 affected individuals and no mismatch repair mutations, we analyzed 135 SNPs from the Affymetrix 10K 2.0 Array (chr 8, 118-142 Mb; chr 9, 0.2-9 Mb) using single-SNP and 3-SNP haplotype sliding window using family-based association testing (FBAT). No suggestive findings were apparent on 8q24. Although the ARTIC 9p24 SNP rs719725 (Zanke et al., Nat Genet 2007) was also null, the next closest 9p24 SNP on the SNP array, rs1821892, showed association (p=0.002), as did the two subsequent SNPs (p=0.05, p=0.004) and certain haplotypes (3 p's <0.01). These associations were mainly limited to 184 population-based families (rs1821892 p=0.01; 132 clinic-based families p=0.76). Exclusion of 500 individuals (24%) who were analyzed in a population-based discordant sibship analysis showing association to rs719725 (Poynter et al., Cancer Res 2007) removed observed associations. Results emphasize the utility of population-based families for detecting modest associations, as predicted by the common disease-common variant hypothesis and the expected enrichment of clinic-based families for high-risk all

78

The effects of linkage disequilibrium in large scale SNP datasets for MDR

B.J. Grady, E.S. Torstenson, M.D. Ritchie
Vanderbilt University, Center for Human Genetics Research

In the analysis of genome-wide association studies (GWAS), an important consideration is the power to identify predictive models of disease. A confounding factor up to this point has been the presence of linkage disequilibrium (LD) in GWAS datasets. In order to look at the effect of LD on association analysis, in particular with respect to detecting gene-gene interactions, genomeSIMLA was used to simulate SNPs in LD and data was then analyzed for 2-locus interactions using Multifactor Dimensionality Reduction (MDR). MDR is a non-parametric statistical method for detecting gene-gene and gene-environment interactions. Using genomeSIMLA, we simulated datasets with varying proportions of SNPs in LD in which 15%-95% of the SNPs were in LD with at least one other SNP in the dataset, where LD is defined by an r2 of at least 0.8. In addition, we simulated 3 different scenarios for the disease susceptibility loci: a model in which both SNPs are in separate blocks of LD; a model with 1 SNP in an LD block and the other not in LD with any other SNPs; a model of 2 SNPs in the same block of LD. Multiple penetrance functions were then used to create varying effect sizes. Results from these analyses indicate that higher levels of LD begin to challenge the MDR algorithm such that the ability to detect the "functional" locus is decreased; however, there is

ample power to detect the SNPs in the blocks of LD with the functional locus. This simulation study indicates that the use of MDR in large scale SNP data with varying amounts of LD can be fruitful as long as one pays attention to the LD patterns within the dataset.

79

Search for a modifier locus of the skeletal muscle involvement in the Emery-Dreifuss muscular dystrophy

B. Granger (1), L. Gueneau (2), R. Ben Yaou (2), V. Drouin-Garraud (3), G. Bonne (2), S. Tezenas du Montcel (1)
(1) UPMC, EA3974, AP-HP, GH PS, Paris, France, (2) INSERM U582, Paris, France, (3) Medical Genetics Dpt, Charles Nicolle Hospital, Rouen, France

Mutations in the LMNA gene cause autosomal dominant Emery-Dreifuss muscular dystrophy (AD-EDMD), characterized by skeletal and cardiac involvements. We observed intrafamilial variability in a family carrying LMNA mutation characterized by a wide range of age of onset of myopathic symptoms. Thus heterogeneity suggests the contribution of modifier locus. We performed a systematic genome scan of 59 individual from a French family of 100 individuals, with 280 microsatellite markers. In a second step, we added 11 markers centred on the area underscored by the first step.
Among the 30 carrier of the LMNA mutation, 14 had skeletal muscle and heart disease, 13 only heart disease, 2 were asymptomatic and 1 had an unknown phenotype. The main criterion was the age of onset of skeletal muscle symptoms. To test linkage and segregation, we used a Monte Carlo Markov chain method as implemented in Loki (Heath 1997). This approach estimates the posterior probability for any given chromosome region of at least 1 trait locus being in that region.
Analysis of the L-score, ratio of posterior to prior probability of linkage, showed a signal on chromosome Z1 (max L-score: 12.9), confirmed after refinement of the map. These results, obtained through a method taking into account both the information of liaison and segregation in a big family, strongly suggest the existence of at least one locus involved in the age of onset of disease.

80

Accounting for heterogeneity in genome-wide homozygosity mapping

A.V. Grant, D.K. Nolan, S. Boisson-Dupuis, L. de Beaucoudrey, O. Filipe-Santos, J. Feinberg, J. Bustamante, J.L. Casanova, L. Abel
Génétique Humaine des Maladies Infectieuses, INSERM U550, Paris, France

Genome-wide (GW) model-based linkage, in particular homozygosity mapping (HM), is a powerful method to locate rare Mendelian mutations when a single locus is implicated. However, statistical power can decrease dramatically in the presence of genetic heterogeneity, which becomes increasingly likely with increasing sample size. Here, we propose a novel approach to test for linkage accounting for genetic heterogeneity in the context of GWHM. After having computed multipoint LOD scores over the whole genome, the principle of the approach is, at each map position i: a) to rank individual family multipoint LOD scores, b) to compute the sum of the highest LOD scores for the first to the kith family, where ki is the last family at the ith position with a positive LOD score. In order to evaluate the significance of these sums in the context of GW results from the entire sample, empirical p-values are calculated using permutations. The method was applied to a sample of 17 consanguineous families including one affected child presenting the syndrome denoted as Mendelian susceptibility to mycobacterial diseases (MSMD: MIM 209950). A genome-wide multipoint linkage scan was conducted using MERLIN on genotypes from a 250K SNP array under a recessive model, accounting for linkage disequilibrium. No positive LOD scores were obtained when considering the entire sample. Interestingly, when accounting for heterogeneity by our proposed approach, preliminary results point to at least one linkage region of interest involving 3 families.

81

Genetic Association of genes in the IGF signaling pathway and Metabolic Biomarkers

C. Gray-McGuire, Q. Lu, Y. Li, E.K. Larkin, S. Patel, S. Redline
Case Western Reserve Univ, USA

Obesity and its comorbidities have become a growing concern over the past 30 years, particularly among minorities. The increase in obesity is attributable to environmental factors as well as genetic predisposition. While the mechanism of homeostasis is not fully understood, insulin resistance, the primary metabolic abnormality underlying obesity, is of great interest. The analysis of our collection of over 1400 individuals from the multi-ethnic Cleveland Family Study (CFS) sample support this with significant familial correlations for a series of metabolic biomarkers that were not seen in a sample of half-siblings. This, coupled with significant evidence of linkage in European Americans (EA) of fasting insulin to the IGF1 region (12q22; LOD=3.2) and in African Americans (AA), of fasting glucose to IGFBP1/IGFBP3 (7p14; LOD=8.0), led to our current assessment of association between genes within the Insulin Growth Factor (IGF) signaling pathway and a series of metabolic biomarkers. Using a subset of the CFS, we performed an association analysis utilizing multiple regression of both nuclear and extended pedigrees, accounting for the correlation between family members and adjusting for body mass index, age and sex. We found significant association between SNPs in this pathway and multiple biomarkers, the most significant of which was adiponectin: $p=2 \times 10\text{-}22$ and $p=7 \times 10\text{-}13$ in AA and EA, respectively. Evidence of association with circulating levels of IGF was also found ($p=3 \times 10\text{-}3$). Finally, results of multivariate modeling of these biomarkers help explain correlated associations.

82

Attributing Hardy-Weinberg Disequilibrium (HWD) to population stratification and genetic association in case-control studies

V.K. Grover (1), D.E.C. Cole (2), D.C. Hamilton (1)

(1) Department of Mathematics & Statistics, Dalhousie Univ, (2) Department of Pathobiology & Laboratory Medicine, Univ. of Toronto

Loci exhibiting HWD are often excluded from association studies, because the HWD may indicate genotyping error, population stratification or selection bias. For case control studies, Wittke-Thompson (WT) (Am J Hum Genet, 2005; 76:967-986) showed HWD can result from a genetic effect at the locus. We extend the WT model to accommodate both stratification and genetic effects. Theoretical genotype frequencies and HWD coefficients are derived under a general disease model for a population with two strata assuming no HWD in either stratum. Maximum likelihood is used to estimate model parameters and a test for lack of fit identifies the models most consistent with the data. Simulations have been carried out to assess the method.
The technique was applied to a dataset consisting of a group of ethnically and clinically heterogeneous kidney stone formers and controls with a presumptive mixture of self-reporting Caucasian and Asian Canadians, both exhibiting HWD for the R990G SNP of the CASR gene. The results indicate that both stratification and genetic association together can account for the HWD. The best fitting model suggests an admixture of about 10% Asians and a recessive genetic component with an increased risk of 2.6 for those with two copies of the variant allele. The ability of this WT-based method to apportion HWD to stratification and genetic effects represents a significant advance in our ability to deal with heterogeneity in case-control genetic association studies.

83

Combining two genome wide association scans for seven smoking related phenotypes replicates published associations in the CHRNA3/5 region

F. Gu (1), A. Bergen (2), N. Chatterjee (3), J. Sheng-Shih (5), K. Yu (3), M. Yeager (4), D.J. Hunter (1), G. Thomas (4), K. Jacobs (4), M.T. Landi (3), S. Chanock (4), J. Chen (1), R. Ziegler (3), N. Caporaso (3), P. Kraft (1)

(1) Program in Molecular and Genetic Epidemiology, Harvard School of Public Health; (2) Molecular Genetics Program, SRI International; (3) Division of Cancer Epidemiology and Genetics and (4) Core Genotyping Facility, National Cancer Institute; (5) Johns Hopkins University

Smoking is a risk factor for more than two dozen diseases and conditions and a leading contributor to mortality worldwide. We performed genome-wide association scans for seven smoking behavior phenotypes using data on 1,144 breast cancer cases and 1,138 controls from the Nurses' Health Study and 1,065 prostate cancer cases and 995 controls from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. We tested the association between these phenotypes and over 518,350 SNPs on the Illumina HumanHap 550k platform that passed quality control and had minor allele frequency $>1\%$ in both studies. We combined study-specific results using weighted Z scores. Although no SNPs achieved genome-wide significance ($p<10\text{-}7$) for any phenotype, we replicated published associations between SNPs in the CHRNA3/CHRNA5 region and cigarettes per day (CPD). We also found evidence that variation in the candidate genes MAOA, TRPV1, and FOSB was associated with CPD (gene-level $p<0.01$). Our study provides further evidence that SNPs in the CHRNA3/CHRNA5 region are associated with smoking behavior, and suggests several other regions for further study.

84

Robust Multifactor Dimensionality Reduction Method for Detecting Gene-Gene Interaction in Bladder Cancer

J. Gui (1), J.H. Moore (1,2,3), A.S. Andrew (1)
(1) Dept. of Community and Family Medicine and (2) Dept. of Genetics (3) Norris-Cotton Cancer Center, Dartmouth Medical School, USA

The central goal of human genetics is to identify and characterize susceptible genes for common complex human diseases. The multifactor dimensionality reduction (MDR) [1] method successfully reduced the high dimensionality caused by combining multi-locus genotypes and provided a key step to facilitate detection of important gene-gene and gene-environment interactions. However, MDR classifies the combination of multi-locus genotypes into high-risk and low-risk groups based on a simple comparison of the ratios of the number of cases and controls to that in the entire data. This may cause false-positive findings when the two ratios are very close to each other. To tackle this problem, we propose a Robust Multifactor Dimensionality Reduction (RMDR) method that performs Fisher's Exact Test on the case-control ratio in all possible genotype combinations. We classify the combination in the high or low risk group when the test is significant; otherwise we classify it to the unknown group. In this way, only genotypes with significant case-control ratios are considered. We expect that this approach will increase the power when heritability is low. In the simulation study, with 400 samples and a heritability of 0.025, we show that RMDR has 70% power to detect the true interaction as compared to 50% from MDR. We then apply the RMDR method to detect interactions in genotype data from a population-based study of bladder cancer.

Reference:
[1] M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, 2001. Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147.

85

Expectation-Maximization Algorithm Based Test Of Informative Missingness (Em-Tim) In Genetic Studies Using Case-Parent Triads

Chao-Yu Guo (1,2)
(1) Clinical Research Program and Program in Genomics, Children¡¦s Hospital Boston
(2) Department of Pediatrics, Harvard Medical School

In genetic studies, the Transmission/Disequilibrium Test (TDT) [Spielman, et al. 1993] using case-parent triads is a popular study design attributable to its robustness to

population admixture. Recently, [Guo 2007] indicated that when offspring genotypes were missing informatively, an occurrence that can be considered as ascertainment bias, inflated type-I error and/or reduced power may occur using the TDT when incomplete triads are excluded. In an effort to avoid an erroneous conclusion from such a genetic study, [Guo, et al. 2008] proposed a method, Testing Informative Missingness (TIM), which detects informative missingness in genetic studies using case-parent triads. The TIM compares the parental genotype distribution in complete triads to that of dyads conditional on the genotypes of affected offspring. The TIM is intuitive, can be easily implemented, and is robust to population admixture. Although the TIM has a decent power, its performance is weaker when informative missingness is caused by a rarer genotype. In this article, we propose a new strategy named expectation-maximization algorithm based test of informative missingness (EM-TIM) that detects informative missingness by discordant linkage and association information between complete triads and incomplete data. According to computer simulations, the EM-TIM is free from admixture and more powerful than the TIM especially when the frequency of genotypes that causes the informative missingness is low.

86

Generalized Linear Modeling with Regularization for Detecting Common Disease Rare Haplotype Association

W. Guo (1), S. Lin (1,2)
(1) Department of Statistics
(2) Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA.

Whole genome association studies (WGAS) have surged in popularity in recent years as technological advances have made large-scale genotyping more feasible and as new exciting results offer tremendous hope and optimism. The logic of WGAS rests upon the common disease/common variant (CD/CV) hypothesis. Detection of association under the common disease/rare variant (CD/RV) scenario is much harder, and the current practices of WGAS may be underpower without large enough sample size. In this paper, we propose a generalized linear model with regularization (rGLM) approach for detecting disease-haplotype association using unphased SNP data that is applicable to both CD/CV and CD/RV scenarios. We borrow a dimension-reduction method from the data mining and statistical learning literature, but use it for the purpose of weeding out haplotypes that are not associated with the disease so that the associated ones, especially those that are rare, can stand out and be accounted for more precisely. By using high-dimensional data analysis techniques, which are frequently employed in microarray analyses, interacting effects among haplotypes in different blocks can be investigated without much concern about the sample size being overwhelmed by the number of haplotype combinations. Our simulation study demonstrates the gain in power for detecting associations for moderate sample sizes.

87

Feature Selection in the right pathway - A gene interaction study

Genet. Epidemiol.

Simone Gupta (1), Hui-Qi Low (2), Katherine Kasiman (2), Hui-Meng Chang (3), Meng-Cheong Wong (4), Christopher P.L.-H. Chen (2), Jian-Jun Liu (1), Anbupalam Thalamuthu (1) *
(1) Genome Institute of Singapore, Singapore, (2) National University of Singapore, Singapore, (3) National Neuroscience Institute, Singapore, (4) National Cancer Centre, Singapore

In the current study, a mutianalytical approach was applied to evaluate the gene interactions in one case-control candidate gene study in the Homocysteine pathway for the etiology of stroke. The primary objective in this study lies in feature (SNP) selection, which can be determined using the measure of interaction entropy. Interaction gain (IG) is described as a measure of the strength of an interaction between attributes and the case-control status (1), represented as an interaction graph. Each feature is an attribute combination that represents a path on the interaction graph. All possible combinations of the attributes up to the order of the length of the path in the interaction graph are constructed. Each feature combination is an independent model and is further assessed for disease risk. Odds ratios (OR) for each genotype combinations is computed and its confidence interval is constructed based on bootstrap samples. SNPs interactions were also assessed under a logistic regression model. The sensitivity and specificity of the genotype combinations to predict the disease status were computed using a Naïve Bayesian classifier. We identified a snp combination as high risk factor for stroke along with it its measure of predictability of the disease.

Reference:
[1] Jakulin A, Bratko I. Lecture Notes Artificial Intelligence 2780, 229-238 (2003).

88

A note on the asymptotic null distribution of likelihood ratio tests for genetic linkage in multivariate variance components models

S.S. Han (1), J.T. Chang (1)
(1) Dept. of Stat, Yale Univ., USA

This study concerns the asymptotic null distribution of certain likelihood ratio tests for detecting genetic linkage in multivariate variance component models. In previous papers and software, the asymptotic null distribution has been stated to be a mixture of several chi-square distributions, with binomial mixing probabilities. Here we show, by simulation and by theoretical arguments based on the geometry of the parameter space, that all aspects of the previously assumed asymptotic null distribution are incorrect—both the binomial mixing probabilities and the chi-square components. The true mixing probabilities give the highest probability to the case where all variance parameters are estimated non-zero, and mixing proportions and critical values depend on unknown true parameters. Correcting the null distribution gives more conservative critical values than previously stated, yielding P values up to 10 times larger. We conclude that significance assessments should be done by empirical methods based on given data. In an example application to

a simulated data set, we illustrate three well known methods for obtaining empirical P-values and compare their results on our data set.

89

Potential interactions among NOS genes in Parkinson disease

D.B. Hancock (1), E.R. Martin (2), W.K. Scott (2)
(1) National Institute for Environmental Health Sciences, Research Triangle Park, NC, (2) Institute for Human Genomics, University of Miami, Miami, FL

Nitric oxide synthase (NOS) genes (NOS1, NOS2A, and NOS3) have been implicated in Parkinson disease (PD), and regulatory interactions have been proposed. We examined interactions in 337 families with sporadic PD (337 cases, 389 controls). Coding and tagging SNPs were tested for allelic and genotypic associations with PD using the Pedigree Disequilibrium Test (PDT) and geno-PDT. Multi-locus associations of SNP genotypes were tested using the Multifactor Dimensionality Reduction-PDT (MDR-PDT) and modeled using generalized estimating equations (GEE). There were significant marginal effects for NOS1 (rs3741475 & rs2682826) and NOS2A (rs12944039, rs2297516, & rs2255929) SNP alleles (range of PDT p=0.0007-0.05) and genotypes (range of geno-PDT p=0.0009-0.05). Results were strongest in 40 families with individuals affected with PD before age 40. In the same families, MDR-PDT found significant associations of two-SNP (rs2255929 in NOS2A & rs1808593 in NOS3, p=0.046) and three-SNP (rs2297516 & rs2297515 in NOS2A & rs1549758 in NOS3, p=0.0070) models. However, GEE model building did not detect a statistically significant interaction for the two-SNP model and could not validate the three-SNP model due to convergence problems with sparse data. The MDR-PDT results suggest that NOS2A and NOS3 might jointly influence risk of PD, but since no interaction was detected in GEE models, the MDR-PDT results may simply reflect the strong single-locus effect of NOS2A. Further exploration in a larger sample of early-onset PD is needed.

90

A new approach to detect gene-environment interaction using haplotype sharing

R. Hein*, V. Rothe* L. Beckmann, J. Chang-Claude
*these authors contributed equally
Department of Cancer Epidemiology, German Cancer Research Center DKFZ, Heidelberg, Germany

We developed a new straightforward and easily implemented approach to detect gene-environment interactions (GxE) for binary exposures, which does not require any assumptions on an underlying disease model.
First, the individuals' haplotype pairs are clustered based on their sharing measured in markers IBS. Subsequently, the proportion of exposed cases is compared to the proportion of unexposed cases in each cluster i, yielding cluster values (CVi) between 0 and 1. CVi=0 indicates the absence of GxE. By contrast, a value 1 may either indicate the presence of environmental main effects or GxE. By clustering with respect to haplotype similarity, data was stratified according to genetic main effects. Hence, CVi greater than 0 does not indicate genetic main effects.
To eliminate the possibility that a value greater than 0 is due to environmental main effects, we test for similarity between cluster values. Dissimilar CVis indicate the presence of GxE. The approach was applied to simulated case-control data. Individuals were generated by drawing haplotype pairs and exposure status according to a given distribution. Disease status was assigned to individuals by a logit model containing the environmental exposure and one chosen risk variant. Compared to four alternative methods, preliminary results show that our new approach has moderate to high power, although the type I error depends on the test for similarity between cluster values. Further investigations will concentrate on that test.

91

An Integrated Autism Gene Knowledge Base

C. Hicks (1), A. Tchourbanov (1), J. Del Greco (2), R. Asfour (2), D. Brazdziunas (3)
(1) Department of Preventive Medicine and Epidemiology & Surgery, Loyola University Medical Center, Loyola University Chicago, USA
(2) Department of Mathematics and Statistics, Loyola University Chicago, USA
(3) Children's Memorial Hospital & Northwestern University, Chicago, USA

The past decade has witnessed hundreds of reports declaring or refuting genetic linkage and association with putative autism susceptibility genes. This wealth of information has become increasingly difficult to follow, much less to interpret. We have created a publicly available, continuously updated database that comprehensively catalogs all candidate genes and gene variants from linkage and association analysis in the field of autism spectrum disorder (the Autism Gene Database). We have performed genomic analysis. The database currently contains 127 autism candidate genes and mouse orthologs, 654 environment genes, genomic and protein sequences. Using our database, we have shown that autism genes interact with environment genes and are functionally related. We are currently expanding the knowledge base to include single nucleotide polymorphisms associated with risk for autism, and gene expression data. Our database provides a powerful tool for deciphering the genetics of autism spectrum disorder, modeling gene and epigenetic regulatory networks in autism, and it serves as a potential model for tracking the most viable candidate genes in other common complex diseases.

92

Estimation of the heritability and the contribution of known genetic factors from twin data applied to Rheumatoid Arthritis

J.J. Houwing-Duistermaat (1), D. van der Woude (2), R.E.M. Toes (2), T.W.J. Huizinga (2), A.H.M. van der Helm-van Mil (2), R.R.P. de Vries (3)

(1) Dept. of Medical Statistics, LUMC, Netherlands, (2) Dept. of Rheumatology, LUMC, Netherlands, (3) Dept. of Immunohematology and Blood Transfusion, LUMC, Netherlands

Purpose: Binary traits are often modeled using logistic regression yielding odds ratios. We added a Gaussian random effect to the model in order to estimate the heritability and the contribution of known genetic factors from twin data(1). We analyzed 148 twin pairs in which at least one twin had Rheumatoid Arthritis (RA) and for which HLA shared epitope (SE) and HLA-DR3 genotypes were typed. HLA SE predisposes specifically to anti-citrullinated protein antibody positive (ACPA+) RA and HLA-DR3 predisposes specifically to ACPA- RA. The prevalence of RA, ACPA+ RA and ACPA- RA for the various SE and DR3 genotypes were known which allowed us to estimate population based parameters from the ascertained twin pairs.
Results: The overall heritability of RA was 66% with a 95% confidence interval (CI) of 44-75%. For ACPA+ RA, heritability was 68% (CI: 55-79%) and the contribution of HLA SE to the genetic variance was 19%. The heritability of ACPA- RA was 66% (CI: 38-83%). The HLA-DR3 alleles contributed 12% to the genetic variance.
Conclusion: The heritability of ACPA+ RA is comparable to the heritability of ACPA- RA. The two genetic risk factors explain only a part of the genetic variance. Our analysis demonstrates that many individual genetic risk factors remain to be identified.

Reference:
[1] J.J. Houwing-Duistermaat, et al. 2000. Biometrics 56:808–14

93

Evaluation of CNV calling algorithms in identifying T-ALL related cancer genes

P. Hu (1), I. Matei (1), E. Parkhomenko (1), C. Guidos (1,2), J. Danska (1,2), J. Beyene (1,2)
(1) Hospital for Sick Children Research Institute, Toronto, Canada
(2) University of Toronto, Toronto, Canada

Characterization of copy number variation (CNV) patterns in human oncogenome provides a major challenge in understanding the biological mechanisms responsible for cancers. Although the importance of CNVs in cancer studies is becoming widely accepted, the optimal statistical methods for identifying these variants are still not well-evaluated. Taking T-acute lymphoblastic leukemia (T-ALL) as an example, we evaluated three CNV calling algorithms on mice BAC array based aCGH data (not published) and human T-ALL samples genotyped with Affymetrix 500K SNP array (Mullighan et al. Nature 446:758:764, 2007). We measured the performance of different calling algorithm based on different criteria: (1) sensitivity to the selected normalization methods; (2) recurrence of CNV regions using STAC method (Diskin et al. Genome Res. 16:1149-1158, 2006).; (3) false discovery rate in simulated data (Willenbrock et al. Bioinformatics 21: 4084-4091, 2005). Although some methods tend to identify large CNV regions while other methods are easier to detect small CNV regions, we found some of genes in the CNV regions consistently detected by these methods are enriched in signaling-related pathways, such as PI3K/AKT Signaling and PTEN Signaling, which have been implicated in the aggressiveness of a number of different cancers, especially T-acute lymphoblastic leukemia
Taking together, we provided a genome-wide list of common copy number alternation regions in T-ALL and further identified novel frequently amplification and deletion regions. Many of the genes associated with these regions represent likely novel oncogenes or tumor suppressors. Our data also demonstrates that findings of study on chromosomally unstable mice tumors have implications in discovery of the biological driver evens in the human oncogenome.

94

A General Framework for Studying Haplotype Effects and Haplotype-Environment Interactions, With Applications to Untyped SNPs

Y. Hu, D. Y. Lin, D. Zeng
Department of Biostatistics, University of North Carolina, USA

Assessing associations between haplotypes (or untyped SNPs) and disease phenotypes is an important and challenging task in genetic epidemiology. The common practice of using probabilistically inferred individual haplotypes in association analysis is generally biased and inefficient. We present a unified likelihood-based approach to this problem. We consider all study designs, including cross-sectional, case-control, cohort, nested case-control, and case-cohort designs. The phenotypes can be disease indicators, quantitative traits, or potentially censored ages at onset of disease. The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate a variety of genetic mechanisms and gene-environment interactions. We allow genetic and environmental factors to be correlated. We construct appropriate likelihood functions and show that the maximum likelihood estimators are consistent, asymptotically normal, and statistically efficient. We develop fast and stable numerical algorithms to implement the corresponding estimation and testing procedures. We extend our approach to the related problem of untyped SNPs by combining the likelihood of an appropriate reference panel (e.g., HapMap) with the likelihood of the study data. Simulation studies demonstrate that the new methods perform well in practical settings. Applications to several genetic epidemiological studies are provided. A computer program is freely available.

95

Modeling the RET gene in Hirschsprung disease

A.S. Jannot (1,2), J. Amiel (2), F. Clerget-Darpoux (1), S. Lyonnet (2) and the Hirschsprung Disease consortium
(1) INSERM, U535, Villejuif F-94817, France
(2) INSERM, U781, Paris F-75743, France

The genetic background of Hirschsprung disease (HSCR) involves one major gene, the RET protooncogene. Many studies demonstrated that several polymorphisms at the RET locus both in coding and non coding sequences are associated with the disease risk, but up to now, the RET gene has never

been modeled with a unified approach. The International HSCR Consortium , which gathers French, American, Italian, Dutch, Spanish and Chinese teams, has collected 780 families for whom mutations on coding sequence have been searched for and 16 SNPs have been genotyped. We found a geographic genetic heterogeneity, but no sex and phenotypic heterogeneity for RET involvement in HSCR, in contradiction with the results of many studies. Using the "combination test" (1), we have then identified the variants that are the most associated to HSCR and shown that several common variants are necessary to explain RET involvement in the disease. Finally, using the genotypes composed by the most associated variants, we were able to test several models with the MASC method (2), which uses both linkage and association information, in order to estimate genotypic risks and their confidence interval, a necessary step towards a better understanding of the disease and genetic counseling. HSCR being regarded as a model for complex diseases, this study promotes a two-step strategy (combination test + MASC) to model candidate genes.

Reference:
[1] A.S. Jannot, et al., 2003. Genet Epidemiol 25:158–67.
[2] Clerget-Darpoux, et al., 1988. Ann Hum Genet 52:247–58.

96

A wavelet based method in association

R.F. Jiang, J.P. Dong, Y.L. Dai
Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931

An important problem in association studies is to find a method which is adaptive to LD structures. If a single SNP has enough LD information, then methods based on a single SNP have more power than multilocus methods. Otherwise, multilocus methods are more powerful. The effectiveness of multilocus methods often depends on LD structure. Consequently, we face the challenge of choosing the right method for given LD structure. It is ideal to have a method which can automatically adapt to the LD structure, and always be able to provide the most powerful test. This is accomplished by a new wavelet thresholding method that we propose in this work. We compare the proposed score test based on wavelet transform with the following commonly used tests: the score test based on Fourier transform; a test obtained by fitting a regression function with a single SNP, then use a Bonferroni correction to find the global p-value; and a likelihood-ratio test based on logistic regression. Simulation studies show that our new test has a significantly higher power over other commonly used tests. It also has the correct type I error rates. It can be applied to both qualitative and quantitative traits.

97

Enhanced Detection of Genetic Association of Left Ventricular Diastolic Dysfunction by Analyzing Novel Latent Phenotypes

Jyh Ming Juang (2), C. Charles Gu (1), Lisa de las Fuentes (2), Alan D. Waggoner (2), Victor G. Dávila-Román (2)
(1) Div. of Biostatistics, Washington Univ., USA
(2) Dept. of Cardiovascular Imaging and Clinical Research Core Lab., Washington Univ., USA

Diastolic heart failure (DHF) accounts for 40-50% of heart failure in clinical practice.It manifests early as left ventricular diastolic dysfunction (LVDD).Most DHF treatment guidelines do not provide definite recommendations because of insufficient molecular and/or genetic information and lack of universal echocardiographic diagnostic criteria.The complexity of LVDD led us to examine intermediate phenotypes,e.g.echocardiographically derived measures,to gain clues to the genetic underpinnings.We applied independent component analysis (ICA) to extract latent LVDD traits from panels of multi-dimensional echocardiographic measures.Based on the latent values,we classified 403 Caucasians into different risk groups of LVDD.We tested genetic associations of the latent LVDD traits with 79 single nucleotide polymorphisms (SNPs) ($r2 > 0.8$) in three PPAR genes involved in the transcriptional regulation of fatty acid metabolism. 64 SNPs with minor allele frequency $\square 5\%$, genotype call rate $\square 90\%$, missing rate $\square 10\%$, and Hardy-Weinberg equilibrium p-value $\square 0.01$ were retained for analysis.The association between dependent variables and individual SNPs were tested by multivariable regression.We found that 14 SNPs were associated with the latent LVDD trait compared with, at the most, 5 SNPs associated with the echocardiographic measures.In conclusion, using the latent trait can improve the detection of genetic underpinning of LVDD.

98

Allelic based Gene-Gene Interaction in Case-Control Study

J. Jung
Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis

Abstract: In the case-control study of complex diseases, it has shown that the interaction caused by multiple single nucleotide polymorphisms (SNPs) within a gene as well as by SNPs at unlinked genes plays an important role to influence risk of the diseases. I propose a new statistical approach that can detect gene-gene interactions at the allelic level contributing to a disease trait. The proposed method assigns a score to each unrelated subject according to their allelic combination inferred from the observed genotypes at two or more unlinked SNPs, and then tests for association of the allelic score at the all possible allelic combinations with a disease trait. By testing for the association of allelic combinations at multiple unlinked loci with a disease trait, the interaction can be assessed both in cases where the SNP allelic association can and cannot be detected as a main effect. Based on the non-centrality parameter approximation and simulation study, I investigate the analytical properties of the proposed methods in terms of type I error rate and power, and demonstrate that the proposed method is an extension of Armitage test for multiple trend in proportions and is identical to the score test derived by the logistic regression method.

Keyword: Armitage trend test, score test, Interaction effect

99

An Efficient Multilocus Monte Carlo Approach for Gene-Centric Genome-wide Association Studies

Guolian Kang
Section on Statistical Genetics, Department of Biostatistics, The University of Alabama at Birmingham, AL 35294

With the development of advanced molecular biotechnologies, genome-wide association (GWA) studies on high-throughput genomic data have been the current state-of-the-art approach in detecting genes underlying complex human diseases. The GWA studies have been focused on single-locus association tests or haplotype-based analyses. However, both types of association studies have their own potential pitfalls in the context of replication and interpretation of association findings. Recently, we proposed a multi-locus analysis approach, an entropy-based gene-centric GWA studies, which is to test disease gene association by treating all single nucleotide polymorphisms (SNPs) in a gene as a functional unit in a genome-wide scale. However, this method can not test the single variant association with disease, and will lose power when there is no or weak linkage disequilibrium (LD) among SNPs within a gene. In this article, we develop an efficient two-stage strategy based on a multiple-locus Monte Carlo approach for gene-centric genome-wide association studies. A proportion of genes are selected by the multilocus Monte Carlo procedure in the first stage (the "gene-screening stage"); the associated SNPs within all associated genes selected in the gene-screening stage are identified by a step-down Monte Carlo procedure in the second stage (the "SNP-testing stage"). The new multiple-locus Monte Carlo approach considers SNPs within one gene simultaneously and is developed based on approximating the joint distribution of all multiple SNPs within one gene. We examine the cases of strong, weak and no LD among SNPs with varying amount of association between SNPs with disease. Using computer simulations, we evaluate the overall family-wise error rate and compare the power when the Bonferroni procedure or Monte Carlo procedure or entropy-based procedure is used to determine significance. Our results show that this proposed approach, which properly accounts for the correlated nature of SNPs data, not only provides accurate and robust control of the overall family-wise error rate, is also substantially more powerful than the standard Bonferroni procedure, especially when the SNPs within one gene are in strong LD or there is more than one disease variant within one gene. Furthermore, it is more robust and mostly more powerful than the entropy-based procedure. So, the proposed two-stage multilocus Monte Carlo procedure, which is computationally efficient and functionally robust, can be applied for a large number of candidate gene association analyses and gene-centric GWAS when the genome-wide genic SNPs become available.

100

Parsing of Bipolar Disorder: a Latent Class Approach

B. Kerner (1), L. Huynh (1), B.O. Muthen (2)
Department of (1) Psychiatry and 2) Education and Information Sciences, UCLA

Background: The heterogeneous phenotype in bipolar disorder (BPD) is one of the major obstacles in identifying genetic and environmental risk factors. Latent class analysis is a suitable tool for addressing phenotypic heterogeneity.
Design: Our study included 5427 individuals from 802 families of the National Institute of Mental Health Bipolar Genetics Initiative (NIMH-BPGI). Latent class cluster analysis on co-morbid conditions was performed on the entire sample, as well as on the bipolar probands only using the computer software program Mplus.
Results: The analysis in the family sample identified five classes: 1. BPD with co-morbid panic disorder, 2. BPD with co-morbid alcohol dependence, 3. BPD with psychotic symptoms without any co-morbid conditions, 4. major depressive disorder, and 5. Individuals without mental illness. Allowing for more classes split the last groups even further, but left the bipolar groups largely unchanged. The analysis of the probands only revealed two subclasses, those with co-morbid substance abuse and those without. Panic disorder or the presence of psychotic symptoms did not characterize a specific subgroup of BPD patients overall, but appeared to aggregate in some families as indicated in the latent class analysis of all family members combined.
Conclusions: Bipolar disorder appears to be a heterogeneous group of disorders characterized by co-morbid conditions. These conditions often precede the onset of the bipolar phenotype and therefore, they might be useful as indicators for more homogeneous subtypes in genetic linkage and association analysis.

101

Genetic effects of physical activity (PA) on different intensity in Korean twin family study

H.J. Kim (1), J. Sung (1), J.Y. Min (1), Y.M. Song (2), K. Lee (3), S.I. Cho (1)

(1) Department of Epidemiology, School of Public Health, Seoul National University, South Korea, (2) Department of Family Medicine, Samsung Medical Center, and Center for Clinical Research, Samsung Biomedical Research Institute, Sungkyunkwan University School of Medicine, Seoul, South Korea, (3) Department of Family Medicine, Busan Paik Hospital, School of Medicine, Inje University, Busan, South Korea

We assessed the genetic effects of physical activity in different intensity such as walking, moderate, and vigorous activity. The study populations are based on the first phase of the¡®Healthy Twin¡¯from 2005 to 2007. participations were 547 families (individual size=1,941) including 469 MZ and 139 DZ twins. We estimated adjusted genetic effects of intensity-specific as well as total physical activity score (MET-minutes/week) using SOLAR. We found that genetic variation in the intensity of physical activity was statistically significant (except for moderate activity). Genetic effects was the highest for vigorous activity (29.7%¡¾39.9%), and lowest for moderate activity (5.8%¡¾34.4%). Total PA was explained by genetic (a2) (14.9%¡¾34.3%), common environment (c2) (9.0%¡¾35.1%), and non-shared environment (e2) (76.1%¡¾35.9%). Physical activities, especially for vigorous or strenuous activities had more significant genetic components than walking or moderate activities.

Acknowledgements: This study was supported by the Center for Genome Science, Korea, National Institute of Health research contract, budgets 2005-347-2400-2440-215, 2006-347-2400-2440-215, 2007-090-091-4854-300.

102

Linkage analysis of gene expressions related to insulin resistance

K.Z. Kim, J.Y. Min, J.H. Sung, S.I. Cho
Seoul National Univ. SPH.

Background: Insulin resistance is an important risk factor of CVD, acting through metabolic disorders. However, the mechanism of their inter-relationships remains poorly understood. This study aims to explore the loci for a common regulator of the gene expressions related to insulin resistance.
Methods: 30 genes known to associated with insulin resistance were matched with the expression QTs (eQTs) in the GAW15 Problem 1 data set. Principal component and factor analyses were performed to identify the clustering of eQTs. Heritability estimation and VC linkage analyses were performed on the each of eQTs, PCs, and factors.
Results: Heritability of the 3 best grouped factors was 24.9∼37.3%, and that of the first 3 PCs was 2.4∼37.5%. The eQTs of UGP2, IGF1, and APOBEC3B showed the greatest heritability in each factor, with 28.6∼56.9%. In the linkage analyses, the highest peak was on chromosome 23 for factor 1 & 2 (LODs 9.2, 3.9), and on chromosome 11 for factor 3 (LODs 2.0). PCs were mapped on similar areas with lower LOD scores. Individual eQTs generally showed strong signals at their coding regions. Factors tended to show higher peaks than the eQTs within the factor. UGP2 and IGF1 were mapped with highest peak on chromosome 23, similar to factors 1 & 2. However, APOBEC3B was mapped on chromosome 2 (LODs 2.5), with only a weak signal on chromosome 11 (LODs 0.8).
Conclusion: Gene expressions related to insulin resistance appear to form subsets sharing common regulators that are mapped on chromosomes 23, 11, and 2. Factor analysis useful approach to identify the relationships among genes involved in a regulatory network.

103

Detecting and Estimating Genetic Association in Extended Pedigrees with a Regression-based Method

S. Kim, N.J. Morris, R.C. Elston
Department of Epidemiology & Biostatistics, Case Western Reserve University, USA

There has been an explosion of statistical developments for genetic association studies using independent samples. However, testing association using pedigree data, but ignoring their dependencies, can lead to misleading conclusions. We first explain the regression-based method for quantitative traits measured on members of an extended pedigree. For a model with a relatively simple familial correlation structure, the statistical mixed model can be used; the likelihood can be efficiently computed by a peeling algorithm. Here we propose a model with more complex familial correlations, allowing a cluster of equally correlated individuals to be members of several nuclear families, and describe how its likelihood can be computed. The regression-based method can be extended to analyze binary traits by using an appropriate link function to form a Generalized Linear Mixed Model (GLMM). Estimation using GLMM is not in general straightforward, but a probit link function allows the likelihood to be calculated as a function of standard normal integrals. We compare this model with the model that uses a logit link function, which allows simpler interpretation in terms of odds ratios. Over most of the range of possible probabilities, these two different link functions yield almost identical results.

104

Linkage analysis of gene expressions related to bone density

T.H. Kim, J.Y. Min, J. Sung, S.I. Cho
School of Public Health, Seoul National University, Republic of Korea

Loss of bone density results in osteoporosis, a common disease with strong genetic and environmental influences. This study was undertaken in order to identify the loci that determines the gene expressions related to bone density.
We used data from GAW15 Problem 1 that includes 3,500 expression QTs (eQTs) in 14 three-generation CEPH Utah families. Based on literature review, 18 genes related to bone density were identified in the data. Factor analysis was used to assess the latent structure in the clustering of the expression patterns. Linkage analysis was performed on each gene expressions as well as common factors.
Some eQTs were mapped on their own coding gene, whereas others peaked on different loci. Suggestive loci were found on chromosome 4, 6, 7 regions (LOD score > 2.0) for OPG gene expression, on chromosomes 17 (LOD score=2.3) for PLOD, and on chromosome 2, 3, 9 (LOD score > 2.0) for TCIRG1. Factor analysis suggested three major clustering of the selected genes. With linkage analysis, factor 1 was mapped on chromosomes 1, 2, 4, factor 2 on chromosome 14, and factor 3 on chromosome 2. Loci for factors did not generally overlap with the individual genes in the factor.
The expressions of bone density-related genes are influenced by many loci that were not reported from previous studies. Linkage analysis combined with factor analysis may provide new insights into complex regulatory network for bone density.

105

Genetic heterogeneity and the power of population- and intra-familial based tests of association for quantitative traits

Y. Kim (1), A.J.M. Sorant (1), A.F. Wilson (1)
(1) Genometrics Section, Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD, USA

In this study computer simulation was used to determine the effect of genetic heterogeneity on tests of association between a quantitative trait and a causal SNP in both population and intra-familial study designs (unrelated individuals, and nuclear and extended families). G.A.S.P. was used to simulate

a single quantitative trait in two sub-populations. In the "association" sub-population, the trait was based on a single SNP marker; in the "non-association" sub-population, the trait was caused by random effect. In each simulation experiment, the sub-populations were combined in different proportions (100%, 50%, 30%, and 10% from the association sub-population) to produce samples with various degrees of genetic heterogeneity. The heritability of the trait in the combined sample was fixed to be 0.05 by altering the locus-specific heritability of the marker in the association sub-population. ANOVA [SAS] was used to test for association in population-based tests of unrelated individuals. ASSOC [S.A.G.E.], FBAT and ROMP were used to perform intra-familial tests on trios and nuclear families. The proportion of samples with a significant result was used to estimate the power at the causal locus and the type I error rate at a second unassociated locus. Although the power of the test decreased as the proportion of the association sub-population decreased for all methods tested, the least reduction occurred with the intra-familial based ASSOC analysis, which also had the greatest power in each scenario.

106

Genome-wide association of serum cystatin C and creatinine in type I diabetics

E. Kistner (1), A. Paterson (2), A. Pluzhnikov (3), A. Tikhomirova (4), J. Below (5), A. Konkashbaev (4), C. Roe (4), D. Nicolae (4,5), N. Cox (4)
(1) Departments of Health Studies
(2) Genetics and Genome Biology, Hospital of Sick Children, Toronto
(3) Section of Genetic Medicine,Internal Medicine
(4) Human Genetics
(5) Statistics, University of Chicago

Both serum cystatin C and creatinine are measured approximations of glomerular filtration rate in clinical practice. Cystatin is preferred as estimates only vary slightly with age, race, gender, and weight. Recently, cystatin has been shown to be both a marker of kidney function and a predictor of cardiovascular mortality. Teasing apart whether cystatin is a sensitive measure of kidney function and thus associated with cardiovascular health, or whether cystatin may be an indicator of additional mortality risk is of interest. The Genetics of Kidneys in Diabetes (GoKinD) cohort of 1825 probands with type I diabetes ascertained as diabetic nephropathy cases and controls, was genotyped using the Affymetrix 500K SNP platform. Genome-wide association of serum cystatin was tested in hopes of uncovering regions implicated in either kidney function or cardiovascular disease. Serum creatinine levels were also tested for association. Significant results were replicated in the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) cohort not ascertained for kidney failure. A region on chromosome 20 where the CST3, CST4, and CST9 genes are located shows significant association with cystatin levels (p=6.09E-07) and is replicated in EDIC. Further work is necessary to determine the causal variation in the associated regions.

107

Detection of Complex Genetic Patterns Using MDR

S. Knight, N.J. Camp
Depart. of Biomedical Informatics, Univ. of Utah, USA

Multifactor dimensionality reduction (MDR) has been suggested as a powerful method for finding susceptibility genes and gene-gene interactions. We investigated MDR's ability to detect haplotype and interactions effects and to correctly identify the direction of the effects. We simulated 5 genetic models of increasing complexity. The most complex model involved 6 components including multiple independent variants (major and moderate), haplotype effects and interactions across genes. The simulations were generated from simulated sequence variation for 5 genes and a set of tagging-SNPs were established for each gene. We performed analyses using these tSNPs, and then repeated the analyses with the disease SNPs (dSNPs). For MDR, an exhaustive search to 7 levels was performed and interaction dendograms were examined to indicate the direction of effects. These results were compared to stepwise logistic regression (LR) for up to three-way interactions. All results were similar for the dSNPs and tSNPs analyses. For the simplest model (an interaction without main effect), both MDR and LR were able to correctly identify the interactions and the directional effect. For a multiple insult haplotype, simulated such that risk was only evident when all 3 variants were present, both MDR and LR identified the effect based on at most 2 SNPs on the haplotype. With increasing complexity, MDR continued to identify most of the simulated variants, however, the directional effects identified were incorrect –falsely concluding redundancy in place of synergy (60% of time for MDR vs. 10% for LR). For the most complex model, MDR was unable to capture a dSNP on a gene which had 2 independent dSNPs and incorrectly classify the interaction compared to LR which captured and correctly classify all effects. The fact that MDR and the interaction dendograms were unable to identify the correct pattern for the complex genetic model is a concern. We suggest exercising caution interpreting MDR's results.

108

Random Coefficients Models, Factor Analysis, and Linear Mixed Model association tests

A.T. Kraja (1), I.B. Borecki (1), J.M. Ordovas (2), J.M. Pankow (3), J.E. Hixson (4), R.J. Straka (5), S.J. Lin (1), D.K. Arnett (6), M.A. Province (1)
(1) Div. of Stat Genom, Washington U, MO; (2) Nut & Genom Lab, Tufts U, MA; (3) Dep of Epi, U of Minnesota, MN; (4) Hum Gen Ctr, U of Texas, TX; Dep of Exp & Clin Pharm, U of Minnesota, MN; (6) Dep of Epi, U of Alabama at Birmingham, AL

The Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study is a clinical familial study examining fenofibrate effects on the lipids profiles. Five factor analyses (FA), corresponding to 5 different visits were performed on 10 metabolic syndrome (MetS) risk variables. The MetS lipid-insulin domain showed a factor structural change before and after fenofibrate treatment. We used a random coeffi-

cients model (RCM) to produce random intercepts and slopes for each subject on the repeated factor scores. Factor scores for each visit, and random slopes from these scores, measuring the subjects' response to treatment, were used as phenotypes in the association tests with 109 SNPs of 28 candidate genes in mixed linear models to account for familial dependencies. A number of SNPs on 11q23, in APOA5 (rs3135506, rs662799), APOC3 (rs4520, rs2854117), APOA4 (rs5104); and PPARA gene on 22q13.31 (rs11703495, rs8138102) were significantly associated with factor scores of MetS lipid-insulin domain. The association tests on RCM slopes identified polymorphisms on 7q36.1 NOS3 (rs1799983, rs1800783, rs743507) gene, as probable candidates associated with the effects of fenofibrates. Hence, combined FA and RCM for repeated measures identified pleiotropic effects of polymorphisms in the MetS lipid-insulin domain change.

109

Application of Bayesian Classification with Singular Value Decomposition Method to Polycystic Ovary Syndrome

S. Kwon, J. Cui, K.D. Taylor, R. Azziz, M.O. Goodarzi, X. Guo
Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA

Genome-wide association studies (GWAs) are aimed at finding genetic variations over the entire genome responsible for a disease. A large number (m) of single nucleotide polymorphisms (SNPs) are engaged in GWAs. The high cost of patient recruitment often limits sample size (n). Conventional statistical approaches are not applicable when $m < n$. We developed the Bayesian classification with singular value decomposition (BCSVD) method and the related test procedure for selecting significant genes, which utilizes permutation and generalized likelihood ratio. Using the simulated Rheumatoid Arthritis data set provided by the Genetic Analysis Workshop 15, we have shown that the BCSVD method was able to identify the disease causing mutations. We applied this newly developed method and gene selection procedure to our polycystic ovary syndrome (PCOS) study sample which includes 134 white women with PCOS and 77 white controls. Given that insulin resistance (IR) is present in most women with PCOS, 317 SNPs in 41 insulin signaling pathway genes were studied using the oligo-ligation assay on an Illumina Bead Station with the goal of identifying genes for PCOS. 12 out of 41 genes (GSK3B, RHOQ, PIK3R1, FLOT1, FYN, PPP1R3A, RAPGEF1, PTEN, SORBS1, FRAT2, GYS2, and INSR) were found to be associated with PCOS. These identified genes may provide an explanation for the observed link between IR and PCOS. The BCSVD is a good choice for analyzing large scale association data when $m > n$, and the test procedure developed here is practical.

110

Evidence for a dominant major gene predisposing to Hepatitis C Virus (HCV) infection in an endemic population

C. Laouénan (1), S. Plancoulaine (1), M.K. Mohamed (2), N. Arafa (2), I. Bakr (2), M. Abdel-Hamid (3), C. Rekacewicz (4), D. Obach (1,4), A. Fontanet (4) and L. Abel (1)
(1) INSERM U550, Paris, (2) Ain Shams Univ, Cairo, Egypt, (3) NHTMR Inst., Cairo, Egypt, (4) Inst. Pasteur, Paris

HCV infects 130 millions persons worldwide and thus is a major public health problem. In developing countries, unsafe injection and blood transfusions are thought to be the major routes of transmission. However, our previous work in an Egyptian population from rural area has shown strongly significant familial correlations suggesting the existence of a genetic predisposition to HCV infection.
To test this hypothesis, we performed a segregation analysis for HCV infection, defined as seropositive/seronegative HCV status, in the same population. We used the regressive logistic model, which allows taking into account simultaneously in addition to the genetic effect, the familial correlations (father-mother, father-offsprings, mother-offsprings and sibs-sibs) and the relevant associated risk factors. A total of 312 pedigrees (4509 subjects) were analysed. The overall HCV seroprevalence was 12.2% increasing with age. The main associated risk factors were previous treatment for schistosomiasis and blood transfusion. We found evidence for a dominant major gene predisposing to HCV infection. The frequency of the predisposing allele was 0.013, indicating that 2.6% of the subjects, in particular those younger than 20 yo, were predisposed to HCV infection. The present study provides evidence for the role of host genetic factors in susceptibility/resistance to HCV infection in an endemic population.

111

Linkage as a $p >> n$ problem - Shrinkage strategies

J.J. Lebrec, H.C. van Houwelingen
Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands

We show how data from linkage studies can be viewed as arising from a regression model with many predictors and too few observations. The predictors are the few thousand human gene effects while the observations are the identical-by-descent (IBD) signal traditionally used in non-parametric linkage analysis. Classical penalization/shrinkage techniques such as Ridge or LASSO regressions may then be applied in order to circumvent the high-dimensionality of the problem. From a Bayesian perspective, these techniques essentially consist in choosing a specific prior for the distribution of gene effects (normal/double exponential distribution for Ridge/LASSO, respectively). For oligogenic and complex traits, such unimodal priors are probably inappropriate so we propose to use 2-point mixture priors as an alternative. Such so-called 'spike and slab' models allow selective shrinkage of gene effects and can be fitted using standard Markov Chain Monte Carlo techniques (as implemented in e.g. the R package OpenBugs) assuming uninformative priors for hyperparameters. The performance of the 'spike and slab' procedure is compared to standard methods in a couple of simulations using Receiver Operating Characteristics (ROC) curves. Finally, some extensions of the method including meta-analysis of linkage studies as well as models for

integration of linkage data with biological pathway information are briefly sketched.

112

Genetic Analysis of Hereditary Endometriosis Families in Puerto Rico

E.M. Ledet (1), I. Flores (2), J.E. Bailey-Wilson (3), D.M. Mandal (1)
(1) Dept. of Genetics, LSUHSC, USA; (2) Dept. of Microbiology, Ponce School of Medicine, PR; (3) NHGRI, NIH, USA

Endometriosis is defined by the growth of endometrial tissue outside of the uterine cavity. This gynecological disease reportedly affects 10% of women in their reproductive years. In Puerto Rico, the estimated prevalence of endometriosis is 4%.
Puerto Rican families with two or more cases of endometriosis were recruited. Blood samples, clinical histories, and surgical confirmation of diagnosis were verified and marker genotypes were obtained. Assuming heterogeneity, on chromosome 10, non parametric multipoint linkage analysis was performed on 42 families and we obtained a LOD score of 0.55. Linkage analysis with 12 families on chromosome 8 produced a LOD score of 0.60, but linkage analysis on chromosomes 1, 3, and 7 did not yield any significant findings. Presently, we have included patient clinical information as covariates and used ordered subset analysis (OSA) to examine the evidence for linkage in the presence of heterogeneity.
An improvement was observed in LOD scores after performing OSA. On chromosome 10, the most significant change in LOD score was observed with a maximum LOD score of 1.12 when 'age at onset' was included as a covariate. On chromosome 8, LOD scores of 1.24 and 1.17 were obtained at the same marker loci with the addition of covariates. OSA on chromosomes 1, 3, and 7 did not yield any significant changes in LOD score with the addition of covariates. In conclusion, for chromosomes 8 and 10, OSA identified genetically more homogeneous subsets of endometriosis families and refined disease gene location.

113

Familial aggregation of measures of kidney function

K.E. Lee, R. Klein, B.E.K. Klein
Department of Ophthalmology and Visual Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI

Serum cystatin-C (cysC) and serum creatinine (creat) are routinely measured markers of renal function. High values of these quantitative measures are used to define kidney disease. The purpose of this analysis is to see if these measures aggregate within families in a general population. Both creatinine and cystatin-C were measured from frozen serum samples obtained in the Beaver Dam Eye Study. Serum creatinine was measured by reflectance spectrophotometry on the Vitros analyzer (Johnson & Johnson Clinical Diagnostics, Inc., Rochester, NY). Serum cystatin C was determined nephelometically using the Dade Behring BN100 nephelometer (Deerfield, IL). The measurements were adjusted for age, gender and body surface area. S.A.G.E (FCOR) was used to estimate correlations. The sib-sib correlations (1125 pairs) were 0.11 and 0.12 for creat and cysC respectively. Similarly the parent-child correlations (511 pairs) were 0.09 and 0.07 and the cousin-cousin correlations (1788 pairs) were 0.02 and -0.03 for creat and cysC respectively. Only the sib-sib correlations are significantly different from 0. The spousal correlations (81 pairs) were 0.43 and 0.35 for creat and cysC respectively suggesting that there are important environmental (or other personal) exposures influencing these measures in addition to some small genetic influences.

114

Discovery of rare variants via sequencing: implications for association studies

B. Li, S.M. Leal
Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030

Common diseases can be due to functional variants with a wide spectrum of allele frequencies. To understand the architecture of allele frequencies and more importantly provide a biomedical resource for common diseases, the 1000 genome project is initially sequencing ~1,000 individuals from 10 different ethnic backgrounds. Currently genome-wide association studies use tagSNPs to uncover genes for common diseases caused by common variants. In the future there will be a shift to analyzing sequence data to understand the etiology of common diseases due to rare variants. We determined for a set sample size what proportion of rare variants can be identified. For a gene with 20 rare variants each with a frequency 0.1% if 1,000 individuals are sequenced the probability is 0.055, 0.956 and 0.999 that 100%, 75% and 50% of the variants will be uncovered. If only 200 individuals are sequenced these probabilities decrease to $2.3x10^{-10}$, $1.5x10^{-4}$ and $8.6x10^{-2}$, respectively. If each of the 20 variants are functional and have a genotypic relative risk of 5.0; sequencing 200 cases will identify 100% and 50% of the variants with probability $4.6x10^{-9}$ and 0.197, respectively. Sequencing genomes of 1,000 individuals from the same population will reveal ~87% of the variants with frequency of 0.1% with a power of 0.9; the number of identified variants falls to ~18% if only the genomes of 100 individuals are sequenced. The proportion of identified variants will be lower if variant frequencies are 0.1%. For the 1,000 genome project due to ethnic specific sample sizes many rare variants including those involved in disease etiology may not be identified. It is also not advisable to sequence a sample subset in order to discover rare variants and then genotype the remaining sample, since missing variants will dramatically reduce power of association studies.

115

A Multiple Testing Adjustment for Correlated Multi-Degree of Freedom Tests

Dalin Li, Juan Pablo Lewinger, David V. Conti
Department of Preventive Medicine, Univ. of Southern California, USA

Large scale candidate gene studies and genome-wide association studies require strong multiple testing adjustments to guarantee a target global type I error. While simple adjustments such as the Bonferroni correction can be conservative due to the correlation of genetic data, exact permutation-based adjustment can be impractical due to their heavy computational burden. Recently, Conneely and Boehnke (2007) introduced an adjustment for correlated one degree of freedom tests (PACT) that can attain the accuracy of a permutation based-adjustment in much less computation time. However, in commonly encountered situations such as the analysis of haplotype associations, gene-gene, and gene-environment interactions, correction for multi-degree of freedom tests is required. Here we introduce an extension of PACT for correlated tests with any number of degrees of freedom. To calculate the adjusted p-values, the multi-df tests statistics are transformed into Z statistics and the contribution of each independent observation to the score is used to estimate their correlation matrix. As in the original PACT approach, the adjusted p-values are obtained from the tail probabilities of the corresponding multivariate normal distribution. A simulation study in which p-values of correlated multiple-df joint tests are adjusted with our approach shows accurate control of the family-wise type I error.

116

Detecting SNP-SNP interactions in trios with affected probands

Qing Li (1), Thomas A Louis (1), M Danielle Fallin (2), Ingo Ruczinski (1)
(1) Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
(2) Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

We present a novel methodology and document open source software for the detection of SNP-SNP interactions in trios with affected probands. We carried out simulation studies to generate high-order snp-snp interactions among case-parent trios using a novel and efficient haplotype and mating table based algorithm. Inference is based on an extension of the logic regression methodology (Ruczinski et al., 2003), embedded in a conditional logistic regression framework. Missing data are addressed using a haplotype-based imputation algorithm. We apply the new methodology to data from a Schizophrenia study (Fallin et al, 2005) consisting of 325 SNPs in 64 genes, typed in each of 300 trios with affected probands.

117

Family-based association method for expanded pedigree with half-sib data

Y.W. Li (1, 2), Y.J. Li (1)
(1) Ctr. for Human Genetics, Duke Univ. Med. Ctr, USA, (2) Dept. of Stat, NC State Univ, USA

Most, if not all, family-based association tests focus on families with parent-offspring triads, full siblings with or without parents, or extended pedigrees including related nuclear families and full sibships. In this study, we aim in developing a family-based association method that can incorporate half-siblings. The new method builds upon the pedigree disequilibrium test (PDT), in which we incorporated identical by decent (IBD) status between two half-siblings. In the case of missing parental genotypes, we infer possible parental genotypes using sibling genotypes. Our proposed Expanded Pedigree Disequilibrium Test (EPDT) includes both single marker and multilocus haplotype association analysis. The EM algorithm was applied to address haplotype ambiguities. The type I error and statistical power were evaluated by simulating various combination of full and half sibpairs with or without parents. Two existing methods, PDT and FBAT, were used for comparison. The simulation results demonstrated that EPDT has correct type I errors (0.043 to 0.056). Type I errors in the FBAT test, however, are inflated ($\sim$ 0.112) when the number of half sibpair sharing 1 IBD is large. For the complete full and half sibpairs with parents, respectively, the EPDT and PDT showed similar power pattern. However, EPDT had slightly greater power than PDT when the ratio of concordant half sibpairs over full sibpairs is increased (89.5% in EPDT and 83.1% in PDT for data with 150 full sibpair and 150 half sibpair families). We observed inflated type I error in FBAT for half sibpairs with parents, which leads to higher power in FBAT than in EPDT or PDT. Since PDT and FBAT will not utilize the information from discordance half sibpairs without parents, we observed the gains of power in EPDT for the datasets with this nature. Overall, EPDT can serve as a good alternative for existing family-based association tests.

118

Common Genetic Variants in Candidate Genes and Risk of Familial Lymphoma (LP)

X. Liang, D. Ng, M.L. McMaster, O. Landgren, N. Caporaso, L.R. Goldin

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

Although the etiologies of LPs are generally unknown, familial aggregation, linkage and case-control studies have supported the role of germ line genes in etiology and suggest that some genes are shared among diverse LPs. Using a custom-designed Illumina panel containing 1536 SNPs, we investigated the effect of common genetic variants in 152 candidate genes from pathways including apoptosis, DNA repair, immune response, and oxidative stress among 44, 50, 28, and 71 unrelated familial chronic lymphocytic leukemia (CLL), Hodgkin lymphoma (HL), non-Hodgkin lymphoma (NHL), and Waldenström macroglobulinemia (WM) patients from high-risk families, respectively, and 107 spouse controls. Several findings were notable. SNPs in BCL2 showed an association with all four phenotypes. SNP rs9989529/BCL2 was significant in both NHL and WM with ORs of 2.28 (95% CI: 1.19-4.37) in NHL and 1.66 (95% CI: 1.08-2.56) in WM. SNP rs4987827/BCL2 was significant in both NHL and HL. Polymorphisms in IL10 (including a variant in the promoter region previously reported to be associated with NHL), TRAIL and TRAILR1 were found associated with CLL and WM. Consistent with

prior data, IL6 polymorphisms were associated with WM and HL.
To our knowledge, this is the first large-scale dense-SNP candidate gene study in familial LPs and highlights the importance of both extrinsic and intrinsic apoptosis pathways in LP etiology. Future investigations are needed to replicate our findings and to define functional roles of genetic variants associated with the risk of developing LPs.

119

Using Observed Population Structure to Investigate Potential Confounding

J.H. Liu (1), F.D. Gilliland (1), C.A. Haiman (1), D.V. Conti (1)
(1) Dept. of Preventive Medicine, USC

Population stratification may lead to spurious associations as a result of sub-populations with different rates of disease and allele frequencies. Using genetic information, many methods exist to estimate and adjust for the underlying structure. However, there remains some ambiguity as to the level of admixture within real populations and to what degree this may lead to confounding. In order to study observed population structure, we select 233 ancestry informative markers (AIMs) that can identify four parental populations, and estimate structure within a multiethnic population containing self-identified African-American, East Asian, Native Hawaiian, Latinos, Caucasians, and Native American. To investigate the impact of this structure on inference, we overlay simulated disease rate and gene frequency differences on the empirical estimates of individuals' admixture. We examine how observed levels of population structure impact bias and type I error, and how various methods perform when controlling for the resulting confounding. In this setting, we discuss the impact of reducing the number of AIMs and the importance of using individuals from different source populations. Furthermore, we expand on current methods by identifying the particular sub-structure driving the confounding. To accomplish this, we traverse the levels of estimated genetic structure and analyze the resulting effect estimates using smoothing splines. We demonstrate this approach via simulations and apply it to multiple SNPs genotyped on the USC Children's Health Study containing self-identified White and Hispanic individuals.

120

Comparison of classification methods for detecting associations between SNPs and chick mortality

N. Long (1), X.-L. Wu (2), D. Gianola (1,2), G.J.M. Rosa (2), K.A. Weigel (2), S. Avendano (3)
(1) Department of Animal Sciences, University of Wisconsin-Madison, USA
(2) Department of Dairy Science, University of Wisconsin-Madison, USA
(3) Aviagen Ltd., Newbridge, UK

Multi-category classification methods were used to detect SNP-mortality associations in broilers. The objective was to select a subset of whole genome SNPs associated with chick mortality. This was done by categorizing mortality rates and using a filter-wrapper feature selection procedure in each of the classification methods evaluated. Different numbers of categories (2, 3, 4, 5 and 10) and three classification algorithms (naive Bayes classifiers, Bayesian networks and neural networks) were compared, using early and late chick mortality rates in low and high hygiene environments. Evaluation of SNPs selected by each classification method was done by predicted residuals sum of squares and a significance test-related metric. A naive Bayes classifier, coupled with discretization into two or three categories generated the SNP subset with greatest predictive ability. Further, an alternative categorization scheme, which used only two extreme portions of the empirical distribution of mortality rates, was considered. This scheme selected SNPs with greater predictive ability than those chosen by the methods described previously. Use of extreme samples seems to enhance the ability of feature selection procedures to select influential SNPs in genetic association studies.

121

C2 and CFB Genes in Severity of Age-related Macular Degeneration

B.A.C. Longville (1,2,3), J. Xiao (2,3), A.X.J. Tan (2,3), X. Feng (2), X. Wu (2,3), C.J. Adams (2), N.M. Warrington (3), P.A. McCaskie (3), J.P. Beilby (4), W.K. Greene (1), I.J. Constable (2), L.J. Palmer (3)
(1) Murdoch U., (2) Lions Eye Inst., (3) Gen.Epi. & Biostats., UWA, (4) PathWest, Perth, Australia

Age-Related Macular Degeneration (AMD) is the leading cause of blindness in the elderly and is classified into subphenotypes based of type and severity of symptoms. The Complement Component 2 (C2) and Complement Factor B (CFB) genes are plausible candidate genes for AMD progression and severity due to their role in the complement cascade; of which the Complement Factor H (CFH) gene, which contributes to AMD risk, is a key regulator.
We investigated the association of C2/CFB gene polymorphisms with subphenotypes of AMD using a haplotype-tagging set of 19 single nucleotide polymorphisms (SNPs) genotyped across the adjacent C2/CFB genes in a comprehensively phenotyped cross-sectional AMD case series (n=1,013) collected as part of the Western Australian Macular Degeneration Study. The multivariate associations of tagging SNPs and AMD phenotypes were tested.
Significant associations between C2/CFB genetic variants and measures of AMD severity such as neovascular AMD (P=0.008) versus early AMD, legal blindness (P=0.04), and composition of neovascular lesions (P=0.005) were observed; all were independent of interaction with the previously reported CFH Y402H variant, and remained significant after adjustment for multiple testing and plausible covariates. Our results provide supporting evidence for the role of C2/CFB dysfunction in the progression and severity of AMD.

122

A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies

X.Y. Lou (1), G.B. Chen (1,2), L. Yan (2), J.E. Mangold (1), J.Z. Ma (3), J. Zhu (2), R.C. Elston (4), M.D. Li (1)

(1) Depts. of Psychiat & NB Sci, & (3) Public Health Sci, Univ. of Virginia, USA; (2) Institute of Bioinformatics, Zhejiang Univ, China; (4) Dept. of Epi & Biostat, Case Western Reserve Univ, USA

Widespread multifactor interactions present a significant challenge in determining risk factors of complex diseases. Several combinatorial approaches have emerged as a promising tool for better detecting gene-gene (G x G) and gene-environment (G x E) interactions. We recently developed a general combinatorial approach, namely the generalized multifactor dimensionality reduction (GMDR) method that can entertain both qualitative and quantitative phenotypes and allow for both discrete and continuous covariates to detect G x G and G x E interactions in a population-based design. We report here the development of a novel algorithm that can be used to study G x G and G x E interactions for family-based designs. Compared to the MDR-PDT (pedigree disequilibrium test) method, our family-based GMDR method has three major improvements: (1) allowing for covariate adjustment, (2) providing a unified framework for analyzing both continuous and dichotomous phenotypes, and (3) coherently handling different family types and patterns of missing marker genotypes. Our simulations provide further evidence that the family-based GMDR method is superior in performance to identify epistatic loci compared to the MDR-PDT. Finally, we applied our novel approach to a genetic dataset on nicotine dependence (ND) and found a significant interaction between two taste receptor genes, TAS2R16 and TAS2R38, in affecting ND.

123

Detecting Epistasis among Candidate Genes using Quantitative Traits and Family Trios

J.P. Lozano and H. Bickeböller
Department of Genetic Epidemiology, University of Göttingen, Germany

The *Transmission Disequilibrium Test* (TDT) is a widely-used method to elucidate the role of genetic factors in complex diseases in families. Variations of the TDT have been proposed to enhance its applicability in different settings. The *Quantitative Transmission Disequilibrium Test with Mating Type Indicator* ($QTDT_M$) (Gauderman, Genet Epi 2003, 25:327-338) is a method for quantitative traits based on linear regression incorporating parental mating types as fixed effects and can be extended to test epistasis. This method has been shown to be more efficient than other methods for quantitative traits in the analysis of candidate genes.
In this study, we simulated family trios with genotypes of two unlinked biallelic loci. We simulated datasets with and without epistasis where the quantitative trait takes on a normal, multimodal or skewed distribution. The $QTDT_M$ was applied to determine epistatic, main gene and overall genetic effects. Different assumptions about the genetic model were considered. Different effects can be well detected by the $QTDT_M$ when the trait is indeed normally distributed. However, $QTDT_M$ may show problems of low power as well as false detection of epistasis under non-normal traits depending on the analyzing genetic model.

124

Performance comparison of variance components and Bayesian MCMC approaches to linkage analysis

J. Ma (1), E.W. Daw (2), C.I. Amos (1)
(1) Dept. of Epidemiology, UT M.D. Anderson Cancer Center, Houston, TX, USA, (2) Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA

Variance components (VC) and Bayesian Markov-chain Monte Carlo (MCMC) approaches are the two commonly used methods for linkage analysis of complex quantitative traits using extended pedigrees. Using simulated data, we compared the performance of the VC and MCMC approaches by calculating their power in detecting genes at a fixed significance level (5%). For the VC approach, we used the multic program implemented in the ACT package and utilized the likelihood ration test (LRT) as the test score for linkage, while for the Bayesian MCMC approach, we used the Loki package. The empirical statistics for linkage test using Loki include the following scores: the L-score implemented in Loki, the log of placement posterior probability ration (LOP), and the posterior probability of linkage (PPL). For data simulated using a simple model with a major gene and a polygenic effect (in terms of variance, major gene: polygene: residual=1:1:9), we found that both VC and MCMC approaches work almost equally well with the test scores (power $>95\%$). However, for the trait height in the GAW 13 data, which was simulated under a genetic model with multiple genes, the Bayesian MCMC approach has a larger power using any one of the three scores (86% for L-score, 83% for LOP, and 87% for PPL) than the VC approach (24% for LRT) in detecting the largest major gene. For the second largest major gene, both approaches have poor powers ($\sim$ 30% for MCMC scores, $\sim$10% for VC LRT).

125

Development of Predictive model with Logic Regression in R

S Mahasirimongkol (1)
(1) National Institute of Health, Department of Medical Sciences, Ministry of Public Health, Thailand

Genome wide association provided invaluable information for the genetic risks of complex diseases, because difference allelic architecture of susceptible alleles are specific for each disease. Based on common disease common variants hypothesis, interaction of identified genetic risks should be the basis of accumulation of risks beyond disease particular disease threshold. It is of interest to develop the interaction model based on novel data mining tool, one of the method that were developed to find the lists of plausible interaction models is Logic regression. While the method is restricted to used only the binary predictor, it provide the computation efficient approach for model development. With tools for evaluate the predictive model in ROCR module, the validate of the the model can be conveniently visualized and interpreted.
The Logic regression module and ROCR module had been applied to a data set of 7 binary clinical and 10 genotypic variables derived from a GWAS of pharmacogenomics study of genetic risk of rash development and the result of the analysis were presented. The training set included 160 cases

of patients, of which 80 of them developed rash after drug intake. The validating data sets are separate groups of 160 patients with the similar phenotypes. Applying logic regression to genotypic and clinical binary predictors resulted in an easily interpreted model and the validation of these models in the validated data set supported several model with AUC of ROC curve in the model developed larger than 0.7, indicated that the models is worth validating in new data set.

126

Gene-gene and gene-time effects in cohorts: Simulation study of a nonparametric longitudinal approach and results on real data

D. Malzahn (1), A. Schillert (1), M. Müller (2), I.M. Heid (2), H.E. Wichmann (2), H. Bickeböller (1),
(1) Dept. of Gen. Epi., Univ. of Goettingen, Germany, (2) Inst. for Epi., HelmholtzZentrum Munich/Neuherberg and Univ. of Munich, Germany

A variety of phenotypes for common disease change longitudinally. Current longitudinal analysis is typically restricted to multivariate normal phenotypes. We propose a nonparametric longitudinal method. A much simpler version has been used as linkage-test for families at a single locus (Kulle et al., BMC Genet 4 Suppl I: S85, 2003) without testing for interactions. We modify and generalize the test for association analysis of quantitative longitudinal phenotypes in cohorts with respect to genetic variants at mutually independent loci including additional covariates and including tests for gene-gene and gene-time interactions. Genetic variants and covariates are modelled as factors. Phenotypes of different individuals are assumed to be independent. Longitudinal observations of the same individual can be arbitrarily dependent. No assumption about the distribution of the longitudinal phenotype is made. The nonparametric null-hypothesis of no interactions is tested by an ANOVA-like rank sum statistic.
We show results from simulation studies and an application to a population-based cohort. Under normality, our method has comparable power in comparison to its parametric ANOVA counterpart but retains its validity and power under non-normality. We apply the method to analyze association of 10-year body mass index with two established susceptibility SNP's in the genes INSIG2 and MC4R in data from the South German cohort KORA.

127

Identification of prostate cancer susceptible locus in high-risk African-American families

D.M. Mandal (1), E.M. Ledet (1), J.E. Bailey-Wilson (2)
(1) Department of Genetics, LSU Health Sciences Center, New Orleans, LA, USA, (2) NHGRI, NIH, Baltimore, MD, USA

Family history is a primary risk factor for prostate cancer irrespective of race. Researchers within the Department of Genetics at LSU Health Sciences Center in New Orleans initiated a genetic linkage study of prostate cancer. A total of 41 high-risk hereditary prostate cancer families (15 African-American and 26 Caucasian) have been recruited (¡Ý3 affected cases/family). Our goal is to identify any region harboring susceptibility genes for prostate cancer in our population. In the preliminary study, three African-American families were genotyped. Fluorescence-labeled PCR primers for 406 unique microsatellite markers with an average spacing of ~10 cM were used. The genotyped data were checked for misspecified family relationships using PEDCHECK. Mendelian inconsistencies were checked with the SIBPAIR program. Single point and multipoint parametric and nonparametric linkage analysis was performed using the Genehunter and GENEHUNTER-PLUS programs. Preliminary analysis on chromosome 22 produced a LOD-score of 0.4, which is close to a region identified by other prostate cancer studies. Recruitment is ongoing to increase the power of the study to detect linkage and more families will be genotyped in the near future. Family history in combination with racial information may aid in effective screening strategies and lead to better understanding of the disease in different ethnic groups.

128

European Collaborative Study of Early-Onset Bipolar Disorder: Evidence for genetic heterogeneity according to age at onset in 2q14

M.H. Dizier §(1,2), F. Mathieu § (3) and the European Collaborative Study of Early-Onset Bipolar Disorder § equal contributors
1. INSERM U535, France
2. IFR69, France
3. INSERM U841, France

As part of the European Collaborative Study of Early Onset Bipolar Disorder, we recently performed a genome-wide linkage analysis in 70 families ascertained through an early onset bipolar type I proband. In this study, we identified 8 regions that were genetically linked to bipolar disorder (BPAD): 2p21, 2q14, 3p14, 5q33, 7q36, 10q23 16p23 and 20p12 (Etain et al., 2006). The aim of the present study is to perform a fine mapping of these regions, using additional markers and an extended family sample (N=120). Moreover, we test the genetic heterogeneity of bipolar disease according to the phenotypic heterogeneity of siblings (BPAD I, BPAD II, early or later forms). Two heterogeneity tests are used, the Predivided Sample Test and the Maximum Likelihood Binomial approach, which both use allele parental sharing distribution among affected sibships. For all regions but two (2p21 and 10q23), the genetic linkage was confirmed, with a maximum non parametric lod-score observed on chromosomes 2q14 and 16p23. Genetic heterogeneity is detected for the 2q14 region (p=10-4) between early and later BPADI, using the two methods. For BPADI sib-pairs concordant for early age at onset, allele sharing proportion is higher (0.58) than for discordant ones (0.28). All these results show for the first time the underlying genetic heterogeneity in bipolar affective disorder and validate the age at onset as a relevant factor in its genetic vulnerability.

129

Heritability of Endophenotypes of Pulmonary & Physical Function in Long Life Family Study

A.M. Matteini (1), M.D. Fallin (1), C.M. Kammerer (2), N. Schupf (3), A.I. Yashin (4), R. Mayeux (3), R.G. Barr (3), K.G. Arbeev (4), K. Christensen (5), E.C. Hadley (6), A.B. Newman (2), J.D. Walston (1)
(1) Johns Hopkins University, (2) University of Pittsburgh, (3) Columbia University, (4) Duke University, (5) University of Southern Denmark, (6) National Institute on Aging

Heritability of physiologic measures associated with exceptional longevity varies. To overcome this heterogeneity, endophenotypes (linear combinations of correlated measures) may improve gene detection. Using current data (1567 persons of 266 families) from the LLFS, we 1) derived endophenotypes representing linear combinations of original measures, 2) estimated heritability of these for future gene finding efforts. We hypothesize underlying latent constructs of longevity may be better detected in genetic analyses of these derived traits than single measures. We randomly selected 1 person per generation per family (n=689 older, 878 younger generation) & used 36 measures from cardiovascular, cognitive, physical function, pulmonary & metabolic domains to estimate the correlation matrix in each generation. We repeated random sampling 1000 times, averaged correlation matrices & performed Principal Components Analyses. In both groups, principal component 1 (PC1) represented 15-18% of total variance & included pulmonary & physical function measures. Second & third PCs represented 10 & 8% of variance, respectively, but content varied widely by generation & sex. Heritabilities of these PCA-generated endophenotypes were estimated. Development & genetic analyses of these endophenotypes should help identify underlying genetic mechanisms driving functional longevity.

130

Association testing with principal-components-based correction for stratification: When and how does it work?

M.S. McPeek (1,2), J. Zhang (1), M. Abney (2)
(1) Dept. of Statistics, (2) Dept. of Hum. Genet., Univ. of Chicago, USA

We consider principal components (PC) methods to correct for population stratification in genome-wide case-control association testing. Published simulations (Price et al. 2006 Nature Genetics 38:904-909) have shown that PC methods work well, at least in certain circumstances. We move beyond simulations to provide a direct theoretical justification. In particular, we address the question: under what conditions should such a procedure give correct type 1 error? We give particular models, assumptions, and choices of test statistic under which the asymptotic type 1 error of the PC-corrected test is correct and others under which it is not and discuss the implications for use of PC in genome-wide association testing.

131

Interaction between Smoking and STAB2 Gene on the Severity of Rheumatoid Arthritis

Jin-Young Min (1), kyoung-bok Min (2), Joohon Sung (3), Sung-Il Cho (3)
(1) Institute of Health & Environment, Seoul National University
(2) Ajou University School of Medicine, Department of Preventive Medicine
(3) Department of Epidemiology, School of Public Health, Seoul National University

Rheumatoid arthritis (RA) is a chronic autoimmune disorder characterized by inflammation of the synovial tissue and deterioration of the joint and bone. A recent study reported a potential gene-environmental interaction between HLA-DR and smoking. The present study was to investigate whether a specific gene was related to the association between smoking and the severity of RA (rheumatoid factor levels >20 IU/ml). We used the resources the NARAC family collection of GAW 15 database, and 1209 subjects were included in the current analysis. The linkage panel contained 5858 SNP markers, and 5744 SNPs passed quality control criteria. Linear regression analyses using the PLINK software and generalized estimating equation regression models were used to test for association between the SNPs and the severity of RA according to smoking groups. The severity of RA in smokers was associated with rs703618 ($p=1\times10^{-5}$), which lies in the intronic region of the stabilin 2 (STAB2) gene on chromosome 12. There were significant differences in the levels of RF between ever smokers and never smokers according to the rs703618 genotype. We investigated whether a specific gene acts as a mediator between smoking and the severity of RA, and found that the STAB2 gene could affect this relationship. Our finding indicates that smoking may mediate RA severity by affecting the expression level of a specific gene.

132

A simple method for co-segregation analysis to evaluate the pathogenicity of DNA variants of unknown significance in BRCA1 and BRCA2

L. Mohammadi (1,4), M.P.G. Vreeswijk (2), J. Wijnen (1), P. Devilee (2,3), C. J. van Asperen (4) and J. C. van Houwelingen (4)
Depts. of (1) Clinical Genetics, (2) Human Genetics, (3) Pathology, (4) Medical Statistics. Leiden University Medical Center

Purpose: Uncertainty about the association of rare DNA variants with disease makes genetic counseling difficult. To classify these variants as disease causing or not, we want to derive likelihood ratios (LR). The aim is to determine whether or not variants are likely to be deleterious mutations.
Method: The analysis of patterns of co-segregation of the variant with disease in families is a powerful tool to obtain likelihood ratios. There are limitations to the procedures proposed in the literature, e.g. genetic linkage software is usually needed for calculations. In this study, we describe a simple method for the analysis of co-segregation of rare variants with disease. We present an algorithm to calculate the likelihood ratios without the need for genetic linkage software.
Results: We applied our algorithm to obtain likelihood ratios in favor of causality of BRCA1 and BRCA2 variants. Our data contained pedigrees with at least one carrier of a BRCA variant. The magnitude of the likelihood ratio depends on the numbers of people with the mutation and with breast or ovarian cancer.

Conclusion: This is a simple and powerful method in analyzing co-segregation. We present a plain algorithm which does not need linkage packages. It can be easily run in the counseling setting as it requires only two affected genotyped persons, gender and the age of onset for breast and/or ovarian cancer.

133

Multifactor Dimensionality Reduction 2.0

J.H. Moore (1), C.S. Greene (1), P.C. Andrews (1)
(1) Computational Genetics Laboratory, Department of Genetics, Dartmouth College, USA

Multifactor dimensionality reduction (MDR) was designed as a nonparametric and genetic model-free approach to identifying, characterizing and interpreting gene-gene interactions in genetic and epidemiologic 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 2.0 of the open-source MDR software package that has been made freely available to the genetic epidemiology community since February of 2005 from www.epistasis.org. This new version includes an estimation of distribution algorithm (EDA) for carrying out a stochastic search for the optimal combination of interacting polymorphisms. The new EDA algorithm provides an alternative to exhaustive search that may not be computationally feasible when the number of polymorphisms is large as in a genome-wide association study. The key feature of this new algorithm is the ability to use expert knowledge in the form of prior statistical evidence (e.g. LOD scores, ReliefF) or biological evidence (e.g. chromosomal location, KEGG pathway, Gene Ontology) to probabilistically select polymorphisms for consideration in an MDR model. Previous studies have shown that detecting interactions in the absence of large marginal effects in genome-wide association studies is not computationally feasible without expert knowledge.

134

Likelihood Ratio Test for Linkage in the Multivariate Variance Component Models

N.J. Morris, C.M. Stein and R.C. Elston
Department of Epidemiology & Biostatistics, Case Western Reserve Univ., USA

The asymptotic distribution of the likelihood ratio test (LRT) for linkage using a multivariate variance component model is not well understood. The general consensus about the asymptotic distribution is that "this issue warrants further detailed attention" (Marlow et al. 2003, Am J Hum Genet 72(3): 561-570), and, "there is an urgent need to characterize the asymptotic distribution associated with these multivariate tests" (Evans et al. 2004, Eur J Hum Genet 12: 835-842). In this work we investigate the asymptotic distribution of this LRT statistic when a number of different possible constraints are put on the alternative hypothesis. We point out that the literature already contains solutions to some of these problems under restrictive assumptions. For more general situations, a simple and computationally efficient approach to calculating asymptotic significance levels is suggested. This approach involves decomposing the parameter space into direction and length components, thus reducing the dimension of the parameter space. Comparisons are made to previously suggested distributions. Some simulation results show the type I error rates for different sample sizes.

135

QTL-ALL: software for QTL linkage analysis

Nandita Mukhopadhyay (1), Samsiddhi Bhattacharjee (1), Chia-Ling Kuo (1), Daniel E. Weeks (1,2), Eleanor Feingold (1,2)
(1) Dept. of Human Genetics, Univ. of Pittsburgh, (2) Dept. of Biostatistics, Univ. of Pittsburgh

Many new statistics for linkage mapping of quantitative trait loci (QTL) in humans have been developed in the last few years, but few have been implemented in end-user software. In particular, score statistics based on the usual variance components likelihood have shown great promise for use with selected samples and/or non-normally distributed data. QTL-ALL is a user-friendly program that provides a wide variety of statistics for performing QTL linkage analysis. Score statistics are emphasized, as are statistics appropriate for ascertained (non-population) samples such as concordant and discordant sibling pairs. QTL-ALL is portable across many platforms such as SOLARIS, Linux, Dec-Alpha, and Macintosh OS X. QTL-ALL is highly automated: starting from input data files in a slightly modified linkage format, it guides the user through a set of simple menus to select marker and trait loci for analysis, does error checking, lets the user select from a list of statistics appropriate for the pedigree structures in the data, and then computes the selected statistics on the data, providing readable, formatted text output as well as graphical plots of p-values.

136

Gene-Environment and Gene-Gene Interactions in GWAS

C.E. Murcray, J.P. Lewinger, W.J. Gauderman
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

It is a commonly held belief that most complex diseases are affected in part by interactions between genes and environmental factors. However, investigators conducting genome-wide association studies (GWAS) typically only test for the marginal effects of each genetic marker on disease. We propose an efficient and easily implemented two-step analysis of GWAS data aimed at identifying genes involved in a gene-environment interaction. Our method expands on the traditional 1-step test for gene-environment interaction in a case-control study by incorporating a preliminary screening step constructed to efficiently use all available information in the data. Specifically, we first screen markers using a case-only analysis applied to all study subjects, and for only those that pass the screen, test for GxE interaction using the

traditional case-control analysis. We demonstrate analytically and by simulation that the two-step method preserves Type I error, even when there is strong population level association between gene and environment. Furthermore, we show that our two-step method is consistently more powerful than the traditional test for gene-environment interaction under many alternative models. For example, when the interaction odds ratio was simulated to be 3.0, power was 33.2 percent using a standard one-step approach, compared to 57.9 percent using our two-step method. We also extend this method to identify markers involved in gene-gene interactions.

137

Comparative Study of Type I Error Rate and Power of Two Methods for Multiple-Testing Adjustment in Case-Control Genetic Association Studies

R. Nickolov, R. Smoak
Department of Mathematics and Computer Science, Fayetteville State University, USA

An important aspect in the modern genetic association studies is the problem of multiple testing. Since hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested for association to a trait of interest many of the tests may be correlated with each other due to linkage disequilibrium (LD) between nearby SNPs. On the one hand classical approaches for multiple-testing adjustment such as Bonferroni correction and Sidak procedure tend to be very conservative. On the other hand, permutation testing provides valid adjustment, but is not computationally efficient. Here we study two recently proposed methods for multiple-testing adjustment in case control association studies. Both of these methods provide accurate calculation of P values adjusted for multiple testing in much less computation time.
The first method [1] adjusts P values for correlated tests directly by comparing the observed tests statistics with their asymptotic distributions through numerical integration. The second method [2] accurately calculates P values by employing importance sampling to considerably decrease the number of sampled permutations, and uses the LD decay property for SNPs to improve running time. We compare the type I error rate and power of the above two methods through simulation of datasets under different realistic patterns of LD, sample sizes, and disease models.

Reference:
[1] Conneely and Boehnke, 2007. Am J Hum Gen 86:1158–116.
[2] Shamir and Kimmel, 2006. Am J Hum Gen 79:481–492.

138

Joint First-Pass Single SNP Analysis and Feature Selection/Dimensionality Reduction for Whole Genome Association Studies (WGAS) using Filter Approaches

K.K. Nicodemus (1,2), Y.Y. Shugart (3)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, UK
(2) Department of Clinical Pharmacology, University of Oxoford, UK
(3) Johns Hopkins School of Public Health, Baltimore, MD, USA

One challenge posed by high-throughput genomics is the development of computationally efficient yet statistically powerful methodologies. Adapting strategies from computer science, we tested whether correlation-based filters could be used as a first-pass analysis for WGAS using a simulation-based study of ~300K SNPs. We simulated whole genome data with two different disease models (6 disease loci: 2 dominant, 2 recessive, 2 multiplicative; odds ratios (ORs) ranged from 1.25-6.25) one considered independent risk loci whereas the 2nd also included 2-SNP interactions. Analyses were conducted using a fast correlation-based filter (FCBF) and minimum redundancy maximum relevance feature selection (mRMR). Based on preliminary results (50 replicates) and focusing on a recessive disease locus with OR=1.66, FCBF selected SNPs within an average 238 SNP window of the disease locus whereas mRMR was less able to localize this modest signal (average SNP window=2076). However, each method reduced the number of SNPs for follow-up to less than 100. Analyses considering all 311915 SNPs were computationally tractable, taking ~20 minutes for mRMR and <1 hour for FCBF. Filter approaches for WGAS may be an efficient way to reduce dimensionality while retaining information. Future work will compare performance of filters with single SNP exact tests for the selection of interesting features.

139

Optimizing Measured Genotype Genome-wide Association in Large Pedigrees

Jeffrey R. O'Connell
University of Maryland School of Medicine, Baltimore, USA

The measured genotype (MG) is a mixed model for quantitative trait association analysis in pedigrees that incorporates SNP genotypes as fixed effects and adjusts for residual familial correlation through a polygenic component. MG analysis provides a flexible regression framework to incorporate environment covariates, multiple genetic models and gene-by-gene interactions. The limiting computational factor in maximizing the MG likelihood is inverting the variance-covariance matrix, whose complexity scales as $O(n^3)$, where n is the dimension of the kinship matrix. Thus, the time for a single SNP analyses quickly increases from seconds to minutes as the dimension increases from tens to hundreds, making genome-wide analysis with SNP chips infeasible. A standard solution is to break the pedigrees into smaller units at the cost of increased Type I error.
We present a solution that allows analysis using the full pedigree based on diagonalizing the variance-covariance matrix to reduce the complexity the likelihood calculation to $O(pn^2)$, where p is the number of covariates. Since the kinship matrix is independent of the SNP, diagonalization is required once for the entire genome-wide scan, providing significant performance gains. For example, analysis of 350K SNPs in 860 Old Order Amish subjects connected into a single 8100 pedigrees required 70 minutes, compared to an estimated 583 days using SOLAR, representing a 12,000-fold

speed up. Our algorithm has been implemented into a user-friendly and optimized software program with a variety of options to facilitate genome-wide MG analysis in large pedigrees.

140

Non-redundant association can prioritize gene regions and provide genomewide significance

Nathan Pankratz, PhD
Indiana University

After performing a genomewide association study (GWAS), it is often difficult to know which regions to follow-up, especially without genomewide significance. Regions with multiple markers showing evidence of association might be prioritized. However, these markers are often in high linkage disequilibrium (LD) with one another ($r^2 > 0.80$), which indicates that these additional markers are providing redundant information. I propose a non-redundant summary (NRS) statistic that down-weights the contribution of additional markers in proportion to their pairwise LD. The NRS statistic is computed for all index SNPs ($p < 0.0001$) and incorporates information from all SNPs within a given window around the index SNP (i.e. 150kb) that exceed inclusion threshold ($p < 0.01$). The p-values are sorted and the negative log of each p-value is multiplied by a weight (one minus the maximum r^2 value between that SNP and any SNP that is more significant than it). These weighted values are then summed and divided by the square root of the total number of SNPs in the window. Empirical p-values are computed by permuting the phenotypes. I will demonstrate this method using brand new GWAS data for Parkinson disease (PD). No individual SNP or haplotype exceeded a Bonferroni correction of 1.5×10^{-7} or had a genomewide $p < 0.45$. The NRS statistic, however, was able to identify two regions at genomewide significance (p=0.01) that contain genes associated with PD. When SNPs were analyzed individually, these regions were ranked 9th and 13th. Novel regions were also nominated. This method could prove to be a powerful tool to help identify susceptibility alleles for complex diseases.

141

Multilocus Analysis of Genome-wide Association (GWA) Studies by Applying Random Forests and Logistic Regression

R. Parisi, D.T. Bishop, M.M. Iles, J.H. Barrett
University of Leeds

In GWA studies of common diseases, testing loci singly has proven successful so far, but may miss loci that in combination have a larger effect than their individual main effects. Using realistically-simulated data we evaluate a tree-based method, Random Forests (RF), for multilocus analysis under a range of effect sizes, modes of interaction and disease allele frequencies and compare its performance with an established method, logistic regression (LR). We assume a two-stage study, with the multilocus methods applied to the most promising loci selected after a singlelocus analysis. Initially RF and LR are applied to 1000 simulated datasets of 1500 (unlinked) SNPs including 16 disease-related loci (10 with a main effect and 3 pairs interacting). Assuming additive, dominant and recessive genetic models and varying the main effect of the interacting loci, the interaction term and disease-allele frequencies, we examine the results using either the importance measure (in RF) or significance level (in LR). A pair of interacting loci is said to be detected if one or both is sufficiently highly ranked using RF and if a sufficiently stringent significance level for the interaction term, measuring departure from a multiplicative model, is reached for LR. Further simulations examine the effect on power of varying the number of disease-related and non-disease-related SNPs. Disease-allele frequency and size of interaction are strong determinants of performance. Increasing the number of non-disease-related SNPs reduces the power of RF; this suggests care when priorotizing SNPs for multilocus analysis.

142

Studying genomic impact of copy number variation on gene expression profiles using Sparse Canonical Correlation Analysis

E. Parkhomenko (1), D. Tritchler (2,3), P. Hu (1), C. Guidos (1,3), J. Danska (1,3), J. Beyene (1,3)
(1) Hospital for Sick Children Research Institute, Canada
(2) Ontario Cancer Inst., Canada
(3) University of Toronto, Canada

Gene expression profiles are widely used in genetic epidemiology to gain insight into complex phenotypes. There has been growing interest to understand genetic causes of variation in gene expression. Studies showed that SNPs can explain underlying patterns of variation. Another form of structural genetic variation affecting gene expressions is copy number variation (CNVs). Recent studies of contribution of CNVs to variation in gene expression are based on association or one-gene-at-a-time correlation. Expression profile may have multiple genetic regulators while one regulatory region may be associated with several expression profiles. Therefore, investigation of connection between sets of genes is preferred. We use Sparse Canonical Correlation Analysis (SCCA) to study the relationships between gene expressions and regions of copy number variation. SCCA performs data integration and simultaneous analysis of the entire sets of variables of different types to identify associated subsets. It is applicable in large-scale studies with limited sample size and a large proportion of measured variables representing noise. SCCA provides sparse solution facilitating biological interpretability and hypothesis generation. We present the performance of our method using simulations and illustrate the application of SCCA to studying contribution of CNVs to variation in gene expression profiles using the study of leukemia in mice.

143

Genome-wide association study of time to long-term diabetic complications

A.D. Paterson (1,4), A.P. Boright (3), D. Waggott (2), Y. Zuo (2), E. Shen (2), L. Mirea (4,2), L. Zhu (2), O. Huang (4), Y. Yoo (2), M. Hosseini (1), P.A. Cleary (5), J.M. Lachin (5) L. Sun (4,1), S.B. Bull (2,4), DCCT/EDIC Research Group

(1) Sickkids, (2) Lunenfeld, (3) UHN, (4) Pub Health Sci, UofT, Toronto, Canada, (5) Biostats, George Washington University

We aimed to identify common alleles from across the genome that are associated with time-to-event for retinal and renal complications of type 1 diabetes. We used data from the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study probands who have repeated measures of complications and risk factors collected for 14 to 20 years. 867,874 SNPs from the Illumina 1M assay with minor allele frequency >1% in 1304 individuals were tested for association with time-to-event diabetic complications using simple and extended Cox Proportional Hazards models. The latter included DCCT treatment group, repeated measures of glycemic exposure, and other baseline measures. Renal outcomes were (1) two consecutive AER > 30 mg/day (PM), (2) AER > 300 mg/day or end stage renal disease. Retinal outcomes were (1) severe non-proliferative diabetic retinopathy (DR) or scatter laser treatment and (2) clinically significant macular edema or focal laser treatment. The most significant results were for DR where rs10491776 had p=6x10-7 and 3x10-6 in extended and simple models, respectively. Another SNP (rs10803641) was associated with PM, p=6x10-7 and 3x10-6 in extended and simple models respectively. We identified a number of other loci associated with time-to-complications in addition to these. Replication is in progress.

144

A Principal Components Approach to Adjust for Population Structure using Family Data

G.M. Peloso (1), J. Dupuis (1), J.N. Hirschhorn (2), K.L. Lunetta (1)
(1) Dept. of Biostat, Boston University, (2) Children's Hospital Boston, The Broad Institute, Harvard Medical School

Even in populations that are considered homogeneous, population structure can result in spurious associations. The Framingham Heart Study is a longitudinal study of over 13,000 individuals begun in 1948. The study consists of 3 generations of individuals in families. A panel of ~550,000 genome wide SNPs were genotyped on >8000 individuals from this study. We applied a principal components (PC) approach using EIGENSTRAT to infer orthogonal axes of continuous genotype variation. We selected a subset of unrelated subjects to infer the weights of the eigenvectors and then applied the weights to the remaining individuals in the sample to determine the eigenvectors for all subjects. The first two PCs of the genotype data show a gradient similar to what has been previously reported in European samples. We found that the first two PCs were associated with adult height in the Framingham sample. The LCT gene is known to contain SNPs that have varying frequency across the European continent. Using linear mixed effects models to test for population association, we found that SNPs near the LCT gene were highly associated with height prior to adjustment for the PCs, but that family based association tests provided no indication of association. The LCT SNP associations with height were non-significant after adjusting for the PCs. Using simulation, we compare the power and type I error of the population-based association approach with PC adjustment to that of FBAT when PCs are and are not associated with the phenotype.

145

Forward-time simulations of admixed populations with complex human diseases

B. Peng, D. Redden, C.I. Amos

Many simulation methods have been proposed to simulate datasets with high-density markers suitable for genome-wide association studies. They are designed for particular applications and vary in their abilities to mimic allele frequency distribution and linkage disequilibrium patterns of real human populations, to reflect the impact of past demographic (such as population expansion and admixture) and genetic (such as natural selection) features, and to produce samples with realistic disease or quantitative trait model. We introduce a forward-time simulation method that can be used to simulate samples with high-density markers with realistic allele frequency distribution and linkage disequilibrium patterns. Using a carefully controlled evolutionary process, this method can simulate isolated or admixed human populations with complex demographic and genetic history. Because the result of such simulations are large populations, different ascertainment methods can be applied to the same simulated population. This allows researchers to study a disease at the population level, and compare head-to-head gene mapping methods based on different ascertainment schemes. One of the key features of this method is its ability to control sample disease allele frequencies, which is critical for the evaluation of the performance of gene mapping methods using a large number of replicates.

146

Regional Genetic Variation and Linkage Disequilibrium in Quebec

M.-H. Roy-Gagnon (1), C. Moreau (1), D. Sinnett (2), C. Laprise (3), H. Vézina (4), D. Labuda (2)
(1) CHU Ste-Justine Research Center, QC, Canada; (2) Dept. of Pediatrics, Université de Montréal QC, Canada; (3) Dept. of Fundamental Sciences, and (4) GRIG, Université du Québec à Chicoutimi, QC, Canada

Founder/isolated populations may be advantageous for gene mapping studies of complex diseases. The population of Quebec (Canada) is a young founder population with a large number of founders that divided into several regional founding events. Genome-wide genetic diversity and linkage disequilibrium (LD) extent for Quebec and its regional populations as well as population structure need to be investigated for appropriate design and analysis of genetic epidemiological studies.
We studied 50 individuals from 5 regional populations of Quebec with the HumanHap650 Illumina panel, which covers the genome with 650,000 Single Nucleotide Polymorphisms (SNPs). We compared allele frequencies, heterozygosity, and the distribution and extent of LD between our

Quebec sample and 50 founders from the HapMap CEU sample (an outbred sample of European descent), and we examined population structure using principal components analysis.
Considering 510,842 common SNPs (frequency?5%), we found that allele frequencies were similar in Quebec and HapMap CEU (less than 0.5% significant differences). The proportion of SNPs 20-50 kb apart in strong LD ($r^2 > 0.8$) was ~8% in Quebec compared to ~7% in the CEU. Four distinct populations were identified by principal components analysis.
Overall, we observed similar genetic variation and slightly higher LD in Quebec compared to HapMap CEU. However, these results varied across regional populations.

147

In search of causal variants: refining disease association signals using cross-population contrasts

N.L. Saccone (1), S.F. Saccone (2), A.M. Goate (2), R.A. Grucza (2), A.L. Hinrichs (2), J.P. Rice (1,2), L.J. Bierut (2)
(1) Dept. of Genetics, Washington University, U.S.A.; (2) Dept. of Psychiatry, Washington University, U.S.A

Genome-wide association (GWA) using single nucleotide polymorphisms (SNPs) is now a state-of-the-art approach to mapping human disease genes. A challenge to interpreting results is that a disease-associated SNP usually represents association with a set of several highly correlated SNPs as measured by r^2. The goal is to distinguish among these correlated loci to highlight potential functional variants. We implemented a systematic method for filtering correlated variants by testing for heterogeneity of genetic effects across diverse population samples having differing linkage disequilibrium (LD) patterns, using logistic regression. The hypothesis is that important biological mechanisms are shared across populations, though allele frequencies may vary. We applied this method to correlated SNPs in the cholinergic nicotinic receptor subunit gene cluster *CHRNA5-CHRNA3-CHRNB4*, in cocaine dependent cases and controls (504 European-Americans and 583 African-Americans). Through simulations we evaluated the power to filter out SNPs in this dataset. Of the 10 SNPs genotyped in the $r^2 \geq 0.8$ bin for *rs16969968*, 3 demonstrated significant cross-population heterogeneity and are filtered from priority follow-up. The results focus attention on a smaller set of SNPs that includes the non-synonymous *CHRNA5* SNP rs16969968. Our approach is an effective tool to enrich for variants more likely to be important and causative, and can help interpret results from GWA studies.

148

On the detection of pleiotropic QTLs in non-random and large pedigrees: empirical evaluation of different multi-trait linkage tests

A. Saint-Pierre (1), M. Cohen-Solal (2), K. Toye (3), A. Ostertag (2), M.C. de Vernejoul (2), J.M. Kaufman (3), M. Martinez (1)
(1) INSERM U563, France; (2) INSERM U606, France; (3) Gent University, Belgium

Several model-free linkage methods have been proposed for detecting QTLs with pleiotropic effects: bivariate tests, with/out constraint on the QTL correlation parameter; Combined Test [Mangin et al., Biometrics, 1998] using composite phenotypes from principal component analysis. The statistical properties of these tests have been evaluated in unselected sib-pair data. CT has been shown to be liberal and less powerful than the bivariate or the univariate test (PC1) on the 1st principal component. But, the theoretical distributions of the tests may not be appropriate [Amos et al., Human Heredity, 2001] [Gorlova et al., Ann Hum Genet, 2002]. Here, we extend this investigation in the context of data of non-random large pedigrees. We used real traits and family data (NEMO: 103 extended pedigrees; size ranges from 8.0 ± 7.3; ascertained through a male with low Bone Mineral Density values). We generated genotypes of a single marker, unlinked to neither BMD trait, using SIMULATE, and analyzed the data with SOLAR. We also found that the use of asymptotical critical values leads to liberal bivariate and CT tests. CT is supposed to be, asymptotically, equivalent to the constrained bivariate (rho QTL=0) [Mangin et al., 1998]. Yet both CT and unconstrained bivariate LODs were found positively and highly correlated in our simulated data. We have also investigated the effect of ascertainment, and marker map on the empirical distributions of the tests.

149

Nonparametric Kernel Score Statistics for Associations of Genotype Similarity with Trait Similarity

Daniel J. Schaid, Jason P. Sinnwell
Division of Biostatistics, Mayo Clinic, Rochester, MN, USA

A number of authors have recently developed methods to evaluate the association of a large number of genetic markers with a trait based on generalized linear mixed models, where adjusting covariates are treated as fixed effects and genetic markers are modeled as random effects; the potential advantage of random effects over fixed effects is a reduction in the number of parameters. Although the published methods tackle the problem somewhat differently, they tend to converge to a score statistic for a variance component, using either parametric functions (usually linear model for both fixed and random effects), semi-parametric functions (linear model for fixed effects and kernel methods for random effects), or nonlinear functions (mainly modeling the random effects with nonparametric functions, based on kernels). A key aspect is how genetic markers are modeled in terms of a genomic sharing matrix, yet this has not been carefully evaluated, such as which kernels are best for different genetic mechanisms. We develop a novel measure of genetic "distance" between pairs of subjects, allowing for missing genetic markers, and illustrate it by application to single nucleotide polymorphisms (SNPs) on chromosome 8 from the Cancer Genetic Markers of Susceptibility study of prostate cancer cases and controls. In addition, limited simulations suggest that a carefully chosen kernel-distance measure has reasonable power to detect SNP-SNP interactions, yet without the need to explicitly model all main effects and interactions.

150

Genome-Wide Association Studies: Implications for Family Disease Risks

Daniel J. Schaid
Division of Biostatistics, Mayo Clinic, Rochester, MN, USA

Genome-wide association studies (GWAS) based on single-nucleotide polymorphisms (SNPs) have rapidly provided genetic clues for common diseases. Almost 100 loci for approximately 40 common diseases have been robustly identified and replicated over the past two years. Most GWAS have been powered to detect common alleles with genotype relative risks 1.2-1.5, so it is not surprising that this range of allelic effect size is commonly reported. Perhaps greater surprises are that few risk alleles involve previously suspected genes and many risk alleles are in regions without known genes. Furthermore, the number of replicable independent risk alleles for some diseases supports a polygenic basis of common diseases. Using results from recent prostate cancer GWAS as a prototype, the polygenic basis of common disease will be explored in terms of the distribution of risk alleles in the general population and in families, and in terms of sensitivity-specificity ROC curves for disease screening. Prostate cancer is a good model, because of the large number of replicated GWAS. Further issues surrounding the polygenic nature of disease will be discussed, in terms of the familial aggregation and genetic epidemiology of prostate cancer.

151

Entropy based marker selection for Mantel Statistics Using Haplotype Sharing on a genomewide scale

A. Schulz (1), C. Fischer (2), M. Guedj (3), J. Chang-Claude (1), L. Beckmann (1)
(1) Dept. of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany
(2) Inst. of Human Genetics, Univ. of Heidelberg, Heidelberg, Germany
(3) Ligue Nationale contre le Cancer, programme Carte d'Identité Tumeurs, Paris, France

Haplotype analysis has been approved of for testing association in extended candidate genes or in regions initially identified in genomewide scans. On the genomewide level, however, haplotype analysis is impaired by computationally unattainable haplotype estimation and arbitrary marker selection.
Recently, we combined a marker selection procedure with Mantel Statistics Using Haplotype Sharing, a test for marker-phenotype association. In the first stage, markers are selected, not necessarily consecutively, by an entropy-based criterion of multilocus linkage disequilibrium, and used for local haplotype estimation. In the second stage, haplotype sharing analysis is applied to the resulting haplotype distribution. Preliminary results show that this algorithm can improve the identification of causal loci in simulated candidate gene scenarios.
In the study presented here, we apply, as proof of principle, the new algorithm for entropy-based marker selection and subsequent haplotype-sharing analysis, intended for the genomewide scale, to simulated data sets.
We discuss the new algorithm with respect to computational feasibility, start and stop criteria, and maximum number of selected markers allowed. The statistical power is compared to the performance of haplotype sharing analysis using overlapping, sliding windows of a fixed number of markers.

152

Racial differences in lung cancer risk associated with SNPs on 15q25 and variation in LD patterns

Ann G. Schwartz, PhD (1), Michele L. Cote, PhD (1), Angela S. Wenzlaff, MPH (1), Susan Land, PhD (1), Christopher I. Amos, PhD (2)
(1) Karmanos Cancer Institute, Wayne State University, 2) University of Texas M.D. Anderson Cancer Center

Three recent genome wide association studies, in Europeans and European Americans, identified a region on chromosome 15q25 associated with lung cancer and nicotine addiction. This region includes nicotinic acetylcholine receptor subunit genes CHRNA3 and CHRNA5. African Americans have not previously been studied. Lung cancer risk associated with 3 SNPs in the region, rs1051730, rs931794 and rs8034191, was evaluated in 1058 NSCLC cases from 4 population-based studies in Metropolitan Detroit and 1314 population-based controls matched within study by age, race and sex. 39% of the cases were African American. African Americans were less likely than whites to carry minor alleles at these 3 SNPs. Under a dominant model, risk associated with the minor alleles of rs1051730 (OR=1.7; 95% CI 1.1-2.6) and rs931794 (OR=1.7; 95% CI 1.2-2.4) was increased in ever smoking African Americans after adjusting for pack-years of exposure. Among whites, pack-years varied by genotype suggesting that lung cancer risk associated with these SNPs in whites is mediated through pack-years of exposure. There is strong LD between all SNPs in European Americans with r2 values >0.88. In our population of African Americans, r2 values were 0.20 or less. When considering all SNPs in a stepwise fashion in African Americans, only rs1051730 was predictive of risk. These findings suggest that the region including CHRNA5/CHRNA3, and not the genes in the region of LD for whites, are involved in determining lung cancer risk.

153

A Statistical Test of Homogeneity with Applications to The International HapMap Samples

N.M. Scott (1), W.C.L. Stewart (2), J.C. Long (1)
Depts. of (1) Human Genet and (2) Biostat, Univ. of MI

We describe a test of the null hypothesis that: two individuals are unrelated members of the same randomly mating population. Our test is novel in that it does not require allele frequency or population membership information. Instead, it only uses simple estimates of homozygosity to compare the multilocus genotypes within and across a pair of individuals. Two estimates are obtained from the average number of homozygous sites within each individual, while a third is obtained from the multilocus genotypes of the pair. Specifically, if b.k denotes the site-specific probability that a randomly chosen allele from the genotype of one individual is identical in state to a randomly chosen allele

from the genotype of the other, the third estimator is (b.1 + … + b.L)/L, where L denotes the total number of sites. Under the null, all three estimates are consistent for the average number of homozygous sites between two copies of the genome, and orthogonal contrasts are used to construct a chi squared test statistic with two degrees of freedom.
From the analysis of simulated data, we show that our test has the correct Type I error, and we examine its power under various alternatives. We apply our test to the genotype data of The International HapMap samples, and we demonstrate its ability to identify pairs who are inconsistent with the null. As a result, our test has the potential to (1) improve inference in genome-wide association studies, (2) inform breeding programs for endangered species; and, (3) infer kinship in situations where only molecular data are available.

154
Inferring Gene-by-Environment Interaction: Do Transmission Rates Reflect Genotype Relative Risks of Disease?
J.-H. Shin, B. McNeney, J. Graham
Department of Statistics & Actuarial Science, Simon Fraser University, Canada

Complex diseases are thought to result from an interplay between genes (G) and environmental or non-genetic attributes (E). The association between the disease and genes is often measured by genotype relative risk (GRR). Statistical interaction between G and E occurs when GRRs vary with the value of E. Transmission rates of a risk allele from heterozygous parents to affected offspring are often compared to their Mendelian expectations under no association or no linkage. Since GRRs that vary with E lead to transmission rates that do too, transmission rates have been used to make inference about GxE interaction. In this project, we investigate the validity of this practice by deriving theoretical transmission rates under different penetrance models and levels of dependence between G and E in the population. We take E to be a continuously varying attribute and G to be the genotype of a single nucleotide polymorphism. Our results illustrate how variation in transmission rates can give a misleading picture about variation in GRRs and hence about GxE interaction under various scenarios. We conclude that if statistical interaction is of interest, direct inference from GRRs is preferred over inference from transmission rates.

155
SNP Selection Strategies from Genome-Wide Association Studies
J.P. Sinnwell, D.J. Schaid
Division of Biostatistics, Mayo Clinic College of Medicine, Rochester, MN (USA)

Selection of subsets of single nucleotide polymorphisms (SNPs) that are associated with at trait is challenging, because of the large number of SNPs from modern genome-wide association studies (GWAS). Typical selection strategies focus on forward-selection regression methods, using, for example, logistic regression for case-control studies. However, forward-selection methods have well-known problems, particularly when the regression covariates (e.g., SNPs) are highly correlated. Although ridge-regression has been proposed to account for collinear SNPs, and thereby shrink the regression parameter estimates, ridge-regression merely shrinks estimates, but does not select the most likely important subsets of SNPs. Alternative shrinkage and selection methods that might be useful for modeling a large number of SNPs include Lasso, adaptive Lasso, and elastic net; tuning parameters for these methods are determined by cross-validation. These novel methods are applied to a large number of SNPs on chromosome 8 from the Cancer Genetic Markers of Susceptibility (CGEMS) study of prostate cancer cases and controls. Contrasting these methods with forward-selection illustrates the potential benefits of simultaneously shrinking parameter estimates while selecting subsets of SNPs. Our results suggest potential guidance on choice shrinkage/selection method, although extensive simulations will be needed to determine the statistical properties of the competing methods.

156
The Genetic Architecture of Leukoaraiosis in Hypertensive Sibships
J.A. Smith (1), S.T. Turner (2), Y.V. Sun (1), M. Fornage (3), T.H. Mosley (4), E. Boerwinkle (3), M. de Andrade (5), S.L.R. Kardia (1)
(1) Dept. of Epid, Univ. of MI, (2) Div. of Nephrol/Hyt, Mayo, MN, (3) Human Genet Ctr, Univ. TX-Houston, (4) Dept. of Med, Univ. of MS, (5) Div. of Biostat, Mayo, MN

White matter hyperintensity on magnetic resonance imaging (MRI) of the brain, referred to as leukoaraiosis, is associated with increased risk of stroke and dementia. Hypertension may contribute to leukoaraiosis by accelerating the process of arteriosclerosis in the brain. Leukoaraiosis volume is highly heritable, but shows significant interindividual variation that is not well predicted by any clinical covariates or single SNPs. As part of the Genetics of Microangiopathic Brain Injury (GMBI) Study, 777 individuals (74% hypertensive) underwent brain MRI and were genotyped for 1956 SNPs from 256 genes known or hypothesized to be involved in arterioscleriosis and related pathways. We examined SNP main effects, epistatic (gene-gene) interactions, and context-dependent (gene-environment) interactions between these SNPs and biological covariates for association with leukoaraiosis volume. Three methods were used to verify significant associations: 1) a ten-iteration four-fold cross-validation scheme, 2) false discovery rate adjustment (FDR), and 3) an internal replication design. A multiple variable model that included the four most highly predictive SNP-SNP and SNP-covariate interactions was able to predict the most variation in leukoaraiosis volume in an independent test sample (11.34%). These results strongly implicate both gene-gene and gene-environment interactions as playing a key role in leukoaraiosis pathology.

157
Use of Haplotype Analysis to Locate Prostate Cancer Susceptibility Loci in a Genome-Wide Association Study (GWAS)

P.L. Smith (1), A.A.A. Olama (1), J. Morrison (1), R.A. Eeles (2), G.G. Giles (3), D.R. English (4), J.L. Hopper (4), D.E. Neal (5), D.F. Easton (1)
(1) CRUK Genetic Epidemiology Unit, University of Cambridge, UK, (2) The Institute of Cancer Research, Sutton, UK, (3) Cancer Epidemiology Unit, The Cancer Council Victoria, Australia, (4) Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Australia, (5) Department of Oncology and Surgery, University of Cambridge, UK. Following a recent GWAS on prostate cancer(1) we applied a localized haplotype clustering method, BEAGLE(2), to compare haplotype frequencies in 1854 prostate cancer cases and 1893 population-screened controls with a low prostate-specific antigen (PSA) concentration (<0.5ng/ml) across 497,699 autosomal SNPs.
We identified six loci which show associations to prostate cancer. Five of these regions had previously been identified in the GWAS(1). The sixth locus is in a novel region.
This program may identify loci not found by conventional GWAS analysis.

Reference:
[1] R.A. Eeles, et al., 2008. Nature Genetics 40(3):316–321.
[2] S.R. Browning, B.L. Browning, 2007. Am J Hum Genet 81:1084–1097.

158

LinkPower: Automated Linkage Power Analysis for Large Complex Pedigrees Using MCMC

Yeunjoo Song (1), Sungho Won (1), Shili Lin (2), Yuqun Luo (1)
(1) Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA
(2) Department of Statistics, Ohio State University, Columbus, OH, USA

Power computation for a linkage study has been traditionally carried out using SIMLINK. With the decreased cost of marker genotyping and the increased variety of linkage study designs in human, such power computation has become less of a concern. However, large complex pedigrees with multiple affected individuals is a powerful design for gleaning linkage information, and is almost the exclusive design in many species other than human, where marker genotyping costs are still substantial. Furthermore, power analysis can provide information on which part of the pedigree we should pursue further in the case of insufficient power. SIMLINK, developed over ten years ago, is not able to accommodate large complex pedigrees with many polymorphic markers. We have developed a user friendly software, LinkPower, that provide power analysis of proposed linkage studies using large complex pedigrees via Markov chain Monte Carlo (MCMC). The correct thresholds for claiming genome-wide significance when there are only a few large pedigrees have been investigated. We demonstrate the use of LinkPower with a highly inbred, 130-member, Poodle pedigree that segregates idiopathic Epilepsy.

159

The Association of Copy Number Variation with the Gene Expression Levels of Matrix Metellopeptidases in Transformed B-Lymphocytes

Y.V. Sun, S.L.R. Kardia
Dept. of Epid., Univ. of Michigan, USA

Copy number variation (CNV) is structural variation of genomic DNA and can potentially affect a number of common human diseases including atherosclerosis. Previous studies indicate that Matrix Metallopeptidases (MMPs) are involved in the atherosclerotic lesion. To explore the potential CNV effects on atherosclerosis, we studied the association of genome-wide CNVs with gene expression levels of MMP1, MMP2, MMP7 and MMP9 using independent samples of 60 Caucasians, 60 Africans, 90 Asians (45 Han Chinese and 45 Japanese). For each sample, the genome-wide variations were measured using Affymetrix 6.0 chip including measurements of 906,000 single nucleotide polymorphisms and 946,000 CNV probes. The gene expression profiles of the transformed B-lymphocyte were measured for the same samples. Among the 210 samples, we identified 2,568 CNV regions on 22 somatic chromosomes. By testing the association between CNVs and the gene expression levels of four MMPs in each racial group, we identified a 638bp CNV region located on chromosome 5 significantly associated with MMP2 expression, and a 127.6 kbp CNV region located on chromosome 14 significantly associated with MMP9 expression, in all three groups (alpha=0.05). The MMP9 associated CNV overlaps with 3 previously reported CNV regions, while the MMP2 associated CNV has not been reported. Although the biological functions of the two loci relating to MMP2 and MMP9 expression are unclear, the significant associations suggest that the CNVs may contribute to the risks of developing atherosclerosis through affecting the mRNA expression of candidate gene.

160

Lack of agreement among intra-familial tests of association for quantitative traits with low heritabilities

H. Sung (1), J.E. Herrera-Galeano (1,2), A.J.M. Sorant (1), R.A. Mathias (1), A.F. Wilson (1)
(1) Genometrics Section, Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD
(2) Dept. of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD

Several different methods are now available for testing for associations between quantitative traits and SNPs in family data. These methods use different kinds of information and have different strengths and weaknesses with respect to their statistical properties. In a study of platelet aggregation, Herrera-Galeano et al. [ASHG, 2007] used several different association methods and found little correlation between results. Computer simulation was used to investigate the lack of agreement among methods. G.A.S.P. [v3.3] was used to generate 10,000 samples, each with 200 nuclear families with sibship size three. A quantitative trait was simulated based on a single biallelic locus with equally frequent alleles. The underlying genetic model was additive and heritabilities considered included 0, 0.001, 0.005, 0.01, 0.05 and 0.1. The data availability was modeled as complete or 50% missing. Five tests of association were performed: ASSOC (SAGE), FBAT using empirically corrected statistic, GEE (SAS GENMOD), ROMP and ROOP. Pair-wise correlations of

p-values were calculated and McNemar tests using 0.001 as cutoff value were performed to test for significant differences between the results of each pair of methods. In general, ASSOC and GEE methods had the highest correlation (greater than 0.89). McNemar tests showed little agreement among methods except when heritability was zero.

161

Performance of model selection criteria in Bayesian network analysis

Y.J. Sung (1), D.C. Rao (1)
(1) Division of Biostat, Washington University in St Louis, USA

Although the importance of gene-gene and gene-environment interactions has long been recognized, association analysis is often performed using single phenotypes and single markers or haplotypes of tightly linked multiple markers. Bayesian networks can be used for associations among multiple markers and phenotypic variables. However, the complexity of the resulting network highly depends on the choice of model selection criterion and sample size. We present the performance of three commonly-used criteria Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and Bayes factors for various levels of model complexity and sample size. We simulated two scenarios, the first with 185 possible statistical models and the second with over one million. Sample sizes ranged from 10 to 50,000. Our most important findings were that AIC outperforms BIC and Bayes factors for small sample size, whereas BIC outperforms AIC and Bayes factors for large sample size. Their percentage of inferring the true model can widely differ, one being near zero and another being near 100%. With the simple model and sample size 2500, the true model was inferred 100% for BIC and 0% for Bayes factors. However, no criterion was always the best. With the complex model and sample size 2500, the true model was inferred 0% for BIC and 64% for AIC. Our results clearly show the importance of selecting a good criterion, which depends on the model complexity and sample size. BIC selects simpler models, whereas AIC and Bayes factors select more complex models. This can offer a guidance for choosing a criterion.

162

Surviving in a non-normal world using adaptive methods as sensible alternatives to parametric and nonparametric ANOVA

S. Szymczak, B.-W. Igl, A. Ziegler
Institute of Medical Biometry and Statistics, University at Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

In order to identify genetic association between single nucleotide polymorphisms and gene expressions parametric and nonparametric analysis of variance methods are common statistical tools. However, often model assumptions are violated since conditional distributions of gene expressions given genotype are highly skewed, heavy tailed or contaminated with outliers. To overcome these problems we consider two competing adaptive methods to detect differences in location parameters. In doing so the underlying common idea is to use information about the empirical data distribution to select an appropriate test statistic. On the one hand, we analyze an adaptively trimmed version of the Welch F-test. On the other hand, we select a certain linear rank statistic in dependence of tail length and skewness.
In our work, we compare both adaptive methods with classical parametric and nonparametric analysis of variance procedures using simulated expression and genotype data. We observe substantial differences between standard and adaptive methods yielding different genes being significantly associated with certain single nucleotide polymorphisms. In the present data situations both standard statistical procedures can fail with an extremely high type I error and a dramatically low power. In these non-normal scenarios we propose adaptive procedures as reliable and sensible statistical mechanisms.

163

Survival adjusted lung function severity score in Cystic Fibrosis modifier gene studies

C. Taylor (1), R. Dorfman (1), A. Sandford (2), P.D. Paré (2), J. Zielenski (1), P. Durie (1), M. Corey (1)
(1) Sick Kids Research Institute, Toronto, Canada
(2) James Hogg Centre, UBC, Vancouver, Canada

Mutations in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene cause CF but do not predict lung disease severity which is the main cause of morbidity and mortality. The Canadian CF Modifier Study conducts candidate gene analyses to find genetic modifiers of CF lung disease in a population based sample.
Forced Expiratory Volume in 1 second (FEV1) predicts survival in CF and is accepted as the best lung phenotype. However, mortality selection in patients beyond teenage years restricts its usefulness as a disease severity indicator.
We estimated the number of deceased patients at each age using cohort survival probabilities from the Canadian CF Registry. Survival adjusted percentile values for FEV1 were computed, and converted to normal z-score, by setting values for deceased patients to 0. We used FEV1 z-score as the phenotype to re-run analysis of a candidate gene, solute carrier family 9, isoform A3 (SLC9A3), which encodes for a Na+/H+ exchanger which we have previously implicated as a potential modifier of rate of decline in FEV1 using a longitudinal mixed model analysis in CF children.
The association of the FEV1 z-score with SLC9A3 genotype was stronger than that seen in our previous longitudinal analysis and we identified a strong age group effect in an analysis of 1580 patients aged 6 to 50 years.
Analysis of other candidate genes is underway and may help explain apparent lack of consistency between gene modifier studies.
Supported by Genome Canada, Canadian CF Foundation.

164

Association of regions on chromosomes 6 and 7 with blood pressure in Nigerian families

B.O. Tayo (1), A. Luke (1), X. Zhu (2), A. Adeyemo (3), R.S. Cooper (1)
(1) Department of Preventive Medicine and Epidemiology, Loyola University Chicago Stritch School of Medicine,

Maywood, USA, (2) Department of Biostatistics and Epidemiology, Case Western Reserve University, Cleveland, USA, (3) NIH Intramural Center for Genomics and Health Disparities, Bethesda, USA

Hypertension shares a level of heritability similar to many other traits related to cardiovascular risk, however specific susceptibility loci have been difficult to localize. We conducted a multi-stage study of blood pressure as a continuous trait in a low-risk West African population where it was anticipated that environmental exposures would be reduced in complexity and intensity. In our earlier genome-wide linkage study for blood pressure in this population, strong linkage evidence was noted on chromosomes 6 and 7. We subsequently genotyped a total of 3431 single nucleotide polymorphisms (SNPs) in three regions (viz, 135.12 – 148.79 cM on chromosome 6, 3.26 – 33.62 cM and 107.89 – 121.37 cM on chromosome 7) in 713 individuals from 199 families. We conducted family-based association analysis using individual SNP while controlling for covariate effects of sex, age and body mass index on blood pressure. After controlling for multiple testing, one SNP on 6q26 and three SNPs on 7q31 retained statistical significance ($p < 0.05$) for the association with blood pressure. The haplotypes on which these SNPs resided were more strongly associated with blood pressure ($p < 0.01$). The frequency of the "at risk" haplotypes ranged from 10% to 39%. These data provide preliminary evidence that regions on chromosomes 6 and 7 may influence susceptibility to elevations in blood pressure.

165

An application of the latent p-value method to assess linkage in asthma pedigrees

C.C. Teerlink, A. Thomas
Department of Biomedical Informatics
University of Utah

The latent p-value is a recently developed, general, empirical method for assessing evidence against a null hypothesis in a stochastic system (statistical model) involving latent, unobservable variables. It is the distribution of the p-values that would be obtained for each configuration of latent variable values, weighted by the probability of each configuration conditional on the values of other related, observed variables. It is particularly applicable to genomewide genetic analysis as it allows the fair evaluation of evidence under multiple testing over genetic markers, pedigrees and phenotype models. We describe its application to a linkage analysis of asthma in 81 extended pedigrees containing __ people genotyped at 533 microsatellite markers. Our interest is in both evaluating the feasibility of the latent p-value method in such a dataset, and in evaluating linkage evidence for asthma. Since the latent p-value distribution is found empirically using Markov chain Monte Carlo methods, this is a computationally challenging problem. However, we establish that the method is feasible for genomewide linkage analysis, and that there is strong evidence for a recessive gene influencing asthma on chromosome 5q13 (median latent p-value=0.03).

166

Modifier genes of age of onset in SCA diseases

S. Tezenas du Montcel (1), B. Granger (1), G. Stevanin (2), S. Forlani (2), A. Durr (2), A. Brice (2) for the EUROSCA Group

(1) UPMC, EA3974, AP-HP, GH PS, Biostat Unit, Paris, France; (2) UPMC, INSERM U679, APHP, Paris, France

Major advances have been made in the understanding of autosomal dominant cerebellar ataxias since the 1980s. A polyQ-coding (CAG)n repeat expansion has been identified as responsible for the disease in five genes: SCA1-3, SCA6-7. The clinical symptoms of these SCA subtypes appear above a threshold number of CAG repeats with a negative correlation between the number of CAG repeats and the age at onset. However, the correlation factor ranges from 0.5 to 0.7 in most studies suggesting that other genetic factors contribute to the variability.
Regression analysis with familial dependency was used to test in 1232 patients (SCA1:314, SCA2:309, SCA3:386, SCA6:167, SCA7:55), recruited through the EUROSCA consortium, the influence of the size of the expended and the normal alleles and of 8 other polyQ genes (SCA1-3, SCA6-7, SCA17, DRPLA, HD, SBMA) on the age at onset. We evidenced an interaction between the expended and the normal alleles in the SCA1 and SCA7 subtypes and the influence of an additional polyQ gene in the SCA3 and SCA6 subtypes. However, the variance taken into account by the major gene and, when appropriate, by the additional gene remains small. A segregation analysis is therefore currently performed on the residual variability after taking in account the available information on the major and the known modifier genes on the 211 SCA families with at least two affected patients. The identification of a new major modulator gene of age at onset will lead to a genome wide study on this population.

167

A Bayesian Change-point Algorithm for the Analysis of SNP-data

F. Thomas (1), S. Pounds (2)
(1) University of Tennessee Health Science Center, Memphis, TN; (2) St. Jude Children's Research Hospital, Memphis, TN

High-resolution genomics data in the form of single nucleotide polymorphism (SNP) arrays can be used in a paired data context to compare cancer tissue to normal samples in an effort to identify regions of genomic amplification or deletion. Such regions potentially contain oncogenes or tumor suppressor genes and are therefore of particular interest. We apply here a Bayesian change-point algorithm to pre-normalized signals from SNP microarrays obtained from a set of leukemia samples in an effort to infer regions of copy number alteration. This algorithm detects multiple change-points where a change can be in the mean of the subsequent measurements, in their variance, in their autocorrelation structure, or in a combination of two or all of these aspects

168

Genome-wide Analysis of Gene-Gene Interaction in Alzheimer Disease

S.D. Turner, E.R. Martin, G.W. Beecham, J.R. Gilbert, J.L. Haines, M.A. Pericak-Vance, M.D. Ritchie

Recent genome-wide association studies (GWAS) in Alzheimer Disease (AD) have refined previous associations and identified new genes of interest. None have found strong main effects outside of APOE, suggesting other effects, such as gene-gene interactions (GxG), may be important. Using Multifactor Dimensionality Reduction, we searched for GxG in a GWAS dataset of 492 AD cases and 496 cognitively normal controls. The model with the highest prediction accuracy (72.3%) included APOE and rs2161082. Considering all 2 and 3-locus interactions of the top 1500 associated SNPs excluding the APOE locus, resulted in 29 models with a cross-validation consistency (CVC) of 10. The top model includes rs72133889-rs1116525-rs2808542 in PECAM1, WWOX and GABBR2, respectively. The average accuracy for this model is 63.5%. In addition to a significant single locus association ($\div^2$=28.7, P<10-6), rs2808542 was present in 14 of the 29 MDR models, suggesting an important marginal effect of this locus. To validate MDR models, we halved the dataset, reserving half as an independent sample for regression analysis. MDR analysis found 13 models with CVC of 9 or 10. In the regression analysis, there were two significant interactions, one between rs6473522 and rs11265191 (OR=2.93, p=.006) and another between rs12683393 and rs3791426 (OR=.41, p=0.034). Our analysis highlights several combinations of loci that show significant effects on AD risk and demonstrates the utility of deep exploration of GWAS data.

169

Testing for genetic association in an affected sibling pair – control design taking into account phenotypic information of relatives

H.-W. Uh (1), M. Beekman (1), P.E. Slagboom (1), J.J. Houwing-Duistermaat (1)
(1) Dept. Of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands

In the Leiden Longevity Study, 450 long-lived siblings (90 years or older) and the partners of their offspring were genotyped (Beekman et al., 2006). Additional phenotypic information (current age or age at death) is also available for un-genotyped (or deceased) family members. Genetic association study in an affected sibling pair – control design is considered. The outcome of interest is exceptional longevity (90 years or older). To take into account different expectancies in the different birth cohorts, we used life tables for the Dutch population.
Incorporation of phenotypic information about relatives who have missing genotype data may increase efficiency in genetic association study. For this purpose, the "more powerful" quasi-likelihood score (MQLS) test (Thornton and McPeek, 2007) can be employed. Firstly, the MQLS test accounts for the correlations among related individuals to obtain the proper type I error rates. Secondly, it uses phenotypic information of both genotyped and un-genoyped families for optimal weighting scheme. We propose to extend the allelic MQLS test to the corresponding genotypic test assuming multiplicative model (Sasieni, 1997).

170

Multi-marker methods lead to detect genetic variants at 2 loci on 21q21 associated with asthma-related phenotypes in the EGEA study

A. Ulgen (1), E. Bouzigon (1), H. Aschard (1), E. Corda (1), M.H. Dizier (2), M. Lathrop (4), C. Amos (5), F. Demenais (1)
(1) INSERM U794, (2) INSERM U535, (4) CNG, CEA France, (5) MD Anderson Cancer Center, USA

A bivariate linkage analysis using principal components (PC) approach led to detect a pleiotropic QTL in 21q21 region for two asthma-related traits: %FEV1 (forced expiratory volume in 1 second) and SPTQ (allergen polysensitization) in 295 EGEA families. Evidence for linkage came mainly from PC2 (p=2x10-5). To identify the genetic variants associated with these traits, we performed family-based association analysis (FBAT) for %FEV1, SPTQ and PCs using two microsatellites and 27 SNPs belonging to three candidate genes (IFNAR2, IL10RB, IFNGR2), located in the vicinity of the linkage peak. The strongest association signals, using single marker analysis, were obtained for PC2 with D21S1252 (p=0.003 for multiallelic test; p=0.003 for allele 2 and p=0.001 for allele 11) and IFNGR2_rs2834213 (p=0.003), these 2 loci being 3 Mb apart. Associations with %FEV1 or SPTQ were weaker (p>0.01). This association analysis was extended to multi-marker analysis considering D21S1252 alleles and rs2834213 using two methods: FBAT-LC (linear combination. of FBAT statistics) and FBAT-MM (multivariate test for markers). Both methods increased evidence for association with PC2 when considering jointly allele 11 of D21S1252 and IFNGR2_rs2834213: p=1x10-4 with FBAT-LC and p=7x10-4 with FBAT-MM. This result shows that genetic variants at two 21q21 loci, separated by a few megabases, are likely to be involved in two asthma-related phenotypes.

171

Testing genetic association in the presence of population stratification

K. Wang
Department of Biostatistics, University of Iowa, USA

Genome-wide case-control association study is gaining popularity, thanks to the rapid development of genotyping technology. In such studies, population stratification is a potential concern especially when the number of study subjects is large as it can lead to inflated false positive rate. Current methods addressing this issue include genomic control, structured population and principal components. While useful, these methods may still lead to excessive false positive rate. A simple method that corrects for population stratification is proposed. Similar to genomic control, this method applies a multiplier to Armitage trend test. However, rather than being a constant, this multiplier automatically adjusts for the variability in genotypes of the SNP of interest - the higher the variability, the larger the multiplier. Unlike genomic control, this multiplier is guaranteed to be less than

1. This method is introduced first for simple case and control design and then for more general situations that have covariates. Preliminary simulation study indicates that this new method tends to be conservative but still has satisfactory power to detect association.

172

Cancer Mortality in Familial and Sporadic Pancreatic Cancer Kindreds

L. Wang (1), K.A. Brune (2), K. Visvanathan (1,3), E. Palmisano (2), M. Raben (3), R.H. Hruban (2,3), A.P. Klein (1,2,3)
(1) Dept. of Epi, (2) Pathology, (3) Oncology, Johns Hopkins Univ., Baltimore MD

An estimated 5-10% of pancreatic cancers have a familial basis. To determine if other cancers also cluster in pancreatic cancer families we examined cancer mortality in families enrolled in the National Familial Pancreas Tumor Registry (NFPTR).
Overall, 8,492 subjects with at least one first-degree relative (FDR) with pancreatic cancer were included in the analyses after controlling for ascertainment. Of these 4,161 were from Familial Pancreatic Cancer (FPC) Kindreds (a kindred with a pair of FDRs with pancreatic cancer). Standardized Mortality Ratios (SMRs) were calculated to comparing the observed number of cancer deaths with those expected adjusting for age, sex and calendar year and we compared risk estimates stratified by family history, number of affected first-degree relatives with pancreatic cancer, proband age of diagnosis, and tertile of family history score.
In our cohort, overall cancer mortality other than pancreatic cancer was significantly increased (SMR 1.2; 95% CI, 1.11 - 1.29) and cancer kindreds members were at an increased risk of death due to cancer of the breast (SMR 1.34; 95% CI, 1.03-1.71), gall bladder (SMR 2.96; 95%CI, 1.28-5.84) as well as other sites. Overall cancer mortality was higher in more highly aggregated pancreatic cancer families (based-upon number of affected relatives or family history score). Our results demonstrate that FDRs of pancreatic cancer patients are at higher-risk of several cancers, suggesting these cancers share either environmental or genetic factors with pancreatic cancer.

173

Assessing a multilocus linear model for association with interaction terms between adjacent SNPs only

J. Wason, F. Dudbridge
MRC Biostatistics Unit, Cambridge, UK

Several authors have discussed the performance of multilocus linear models with different orders of interaction. The main-effects-only model, with a parameter for each SNP, outperforms the saturated model, with a parameter for each haplotype, in most situations. However in some situations, such as when the causal locus is rare, higher-order models are preferable. We study an intermediate model including the main effects and interactions between adjacent SNPs only. This is motivated by recent studies suggesting that the majority of linkage disequilibrium can be attributed to correlation between adjacent SNPs. We aim for consistently good power across a range of scenarios, and to improve localisation of a disease locus by modelling the LD of the region in a parsimonious way.
We used Cosi to simulate genotypes under a European like LD structure, and varied the genetic parameters and disease model in order to test how well the adjacent-interactions model performs in comparison the to main-effects model and a model with higher orders of interaction.
With respect to power, most situations favour main-effects, with adjacent interactions close behind and the higher order model performing worst. Some situations which favour a higher order model (for instance, when the causal SNP is rare and untyped) had the adjacent-interactions model performing better than the main-effects model. These results indicate that when the underlying disease model is unknown, the main-effects with adjacent interaction terms is a good compromise model.

174

Complex Segregation Analysis of an Isolated Leprosy Population from North of Brazil

R.I. Werneck (1), F.P. Lazaro (1), A. Alcaïs (2), L. Abel (2), M.T. Mira (1)
(1) Graduate Program in Health Sciences, Center For Biological and Health Sciences, Pontifical Catholic University of Paraná; (2) INSERM U550, Necker Medical School, France

Leprosy is a chronic infectious disease caused by Mycobacterium leprae that affected ~300.000 new individuals in 2006, mainly in Brazil and India. Genetic analysis has been successfully applied to the identification of host genetic factors impacting on susceptibility to leprosy. However, a consensus regarding the exact mode of inheritance is yet to be achieved, in part due to the heterogeneity of the studied populations. The objective of this study was to conduct a Complex Segregation Analysis (CSA) on leprosy using data from the Prata Colony, an isolated former leprosy colony founded in the 1920's on the outskirts of Brazilian Amazon presenting: large multiplex, multi-generation leprosy families; high disease frequency; homogenous environmental and socio-economical variables. Our enrollment strategy was complete ascertainment leading to the examination and inclusion of the whole colony, totalizing 2868 individuals (225 affected) distributed in 112 pedigrees. CSA was performed by using REGRESS, which specified a regression relationship between the probability of a person to be affected and a set of explanatory variables. CSA identified a best-fit codominant model, with the frequency of predisposing allele estimated as 0.22 (p-value=5,04 e-07). Given the unique characteristics of the studied population, we believe these results contribute significantly to the cumulative evidence of a strong genetic component in host susceptibility to leprosy.

175

Genomewide linkage and association of ocular refraction in the Framingham Eye Study

R. Wojciechowski (1), G. Ibay (1), L.D. Atwood, (2), J.E. Bailey-Wilson (1) and D. Stambolian (3)
(1) IDRB/NHGRI, USA, (2) Boston U., USA, (3) U. of Penn, USA

Purpose: Refractive disorders are the most common causes of visual impairment worldwide. Previous studies have reported linkage of myopia or ocular refraction to a number of loci, but no genomewide association results have been published. We report results of genomewide linkage and association scans in Framingham Eye Study (FES) families.
Methods: Eye exams were conducted on 2,540 FES participants in 293 families. We performed quantitative trait linkage and association analyses on ocular refraction, defined as the spherical equivalent refractive error. Genotypes were available for 1,240 individuals at ~600 autosomal STRP and ~113,000 SNP markers. Variance-components (VC) linkage and family-based association analyses were performed using MERLIN and FBAT, respectively. FBAT was performed under additive genetic models for single SNPs and haplotypes using a 3-SNP sliding window.
Results: VC linkage analyses yielded a peak LOD score of 4.16 (p=0.00001) at 124 cM on chr. 2q14.3. In the association analyses, four individual SNPs showed nominal significance levels at $p<0.0001$. The strongest association in the haplotype analyses was found in the chr. 2 linkage region (p=0.0009 at ~80.96 Mb or 103 cM) .
Conclusions: We found significant linkage of ocular refraction to a region on chr.2q. Though not genomewide significant, haplotype association results are consistent with the presence of a locus influencing ocular refraction on chr. 2. Further investigation of this region with a dense SNP map is warranted.

176
Phase uncertainty in case-control association studies
S.W. Won, R.C. Elston

There have been many SNP-based tests suggested for association analysis in a case-control design. The possible evidence for association comprises three types of information: differences between cases and controls in allele frequencies, in parameters for Hardy Weinberg disequilibrium (HWD), and in parameters for linkage disequilibrium (LD). In principle, LD between marker and disease alleles results in a difference in at least one of the three types of parameters and these three types of information correspond to existing methods of analysis. We have quantified the standardized expected differences of the parameters that use the three types of information according to disease mode of inheritance. The parameters for LD require knowledge about phase, which is usually unknown, making the LD contrast test without modification infeasible in practice. There are several methods that handle phase uncertainty. First, the most probable haplotype pair for each individual can be considered as the true phase. Second, a weighted average of haplotypes can be used. Finally, we can consider the composite LD, which does not require any information about phase. We compare these methods to handle phase uncertainty in terms of validity and efficiency. Whereas the difference in allele frequencies is usually the most informative test except in the case of a recessive disease, the LD contrast test can be more powerful if the markers are dense enough. The LD contrast test that uses a weighted average of haplotypes to calculate the LD is recommended for best statistical power with preserved type I error.

177
Incorporation of Genetic Covariates into Statistical Analysis of Family Studies
C.C. Wu, S. Shete
Department of Epidemiology, M. D. Anderson Cancer Center, Houston, Texas, USA
To determine the genetic basis of a complex trait, it is important to use methods that simultaneously take into account the joint effects of multiple genetic components underlying the trait. It is even more important when mutations at a single locus appear to explain only a small portion of familial aggregations of a disease. Even in the presence of a stable mutation segregating in a family, the age of onset within a family is often highly variable, suggesting the presence of additional genetic effects. This is true for many common complex traits that have been studied, such as the impact that BRCA1/2 mutations have upon time to onset for breast cancer. Because usual segregation analysis is efficient only for Mendelian traits, we propose to incorporate genetic covariates (mutation carrier status) in segregation analysis models that account for the genetic complexity and heterogeneity of a complex trait. We first used an independent genetic covariate (p53 mutation status) in 6 extended families of Li-Fraumeni syndrome on the basis of Cox proportional hazards regression. Because distributions of hereditary mutations are inter-dependent within families, we are using simulation-based approaches to quantitatively assess the effects of dependent genetic covariates. We illustrated this approach by analyzing cancer incidence in Li-Fraumeni syndrome families with germline p53 mutations.

178
Genetic variation in ORM1-like 3 (ORMDL3) and gasdermin-like (GSDML) and childhood asthma
H. Wu (1), I. Romieu (2), J.J. Sienra-Monge (3), H. Li (1), B.E. del Rio-Navarro (3), S.J. London (1,4)
(1) LRB, (4) EB, DIR, NIEHS, NIH, DHHS, RTP, NC, USA
(2) NIPH, Cuernavaca, Morelos, Mexico
(3) HIMFG, Mexico City, Mexico

Background: A genome wide association study has identified ORM1-like 3 (orosomucoid 1-like 3, ORMDL3) as a potential asthma candidate gene. Single nucleotide polymorphisms (SNPs) in the region including ORMDL3 on chromosome 17q21 were related to childhood asthma risk and expression levels of ORMDL3 in Europeans.
Objective: We examined whether polymorphisms in ORDML3 and the adjacent gasdermin-like (GSDML) gene that were associated with asthma in the genome wide association study are related to childhood asthma and atopy in a Mexico City population.
Methods: We genotyped rs4378650 in ORMDL3 and rs7216389 in GSDML in 615 nuclear families consisting of

asthmatic children aged 4 to 17 years of age and their parents. Atopy was determined by skin prick tests to 25 aeroallergens. **Results**: Homozygosity for the minor allele of either rs4378650 or rs7216389 was associated with decreased risk of childhood asthma (relative risk (RR)=0.57, 95% confidence interval (CI) 0.38-0.86, p=0.008 for rs4378650, and RR=0.57, 95% CI 0.38-0.87, p=0.009 for rs7216389). Linkage disequilibrium between the two SNPs was high (r2=0.92) in our population. The two SNPs were not associated with the degree of atopy.
Conclusion: Our results provide evidence to confirm the finding from a recent genome wide association study that polymorphisms in ORMDL3 and the adjacent GSDML may contribute to childhood asthma.

179
Population stratification in linkage
Chao Xing
University of Texas Southwestern Medical Center at Dallas

Population stratification as a confounding factor in genetic association studies has long been recognized and proper measures are taken; however its impact on validity of linkage methods has long been neglected. Population stratification directly affects allele frequency estimation at the population level, and subsequently has an impact on linkage analysis depending on the study design. In this study, by taking an analogy between genetic linkage and association analysis we elucidate the role of population stratification in linkage. Focusing on model-free linkage analysis of sibship data on a binary trait we propose detecting population stratification by genotyping additional members other than the affected in a proportion of families and comparing the proportion of alleles shared identical-by-descent estimated from large families with that estimated from small families, and propose controlling for it by a matched study design with paired comparison tests. Controlling for population stratification will be particularly useful for diseases of late-at-onset and studies of sib pair design in admixture populations. Caution should be taken when interpreting linkage results from studies without founders' genotypes unambiguously resolved, and former linkage peaks from such studies should be re-examined by methods controlling for population stratification.

180
A Bayesian approach for imputation of missing genoytpes
H. Xu, J. Choi
Department of Biostatistics, Medical College of Georgia, USA

Large volume of genetic information is available with the advancement of high-throughput single-nucleotide polymorphism (SNP) genotyping techniques. Consequently it is now possible to study gene-gene interaction and develop genetic models for complex diseases. However, missing genoytpes calls are common problems for typical high-throughput genotyping techniques. The missing genotype information can cause considerable problem for a large- scale genome-wide analysis. In this study, we propose a Bayesian approach for imputing these missing genotypes which improves the investigation of genetic analysis such as genome-wide association studies. The method was applied to the genome-wide association data from North American Rheumatoid Arthritis Consortium (NARAC) and compared with other competing approaches for imputation.

181
Transcription Factor 7-Like 2 (*TCF7L2*) Polymorphism is Associated with Impaired Fasting Glucose in Caucasian Participants of the Atherosclerosis Risk in Communities (ARIC) Study
Y. Yan (1), K.E. North (1), A. Kottgen (2), J.S. Pankow (3), E. Boerwinkle (4)
(1) UNC, Chapel Hill; (2) JHU, Baltimore; (3) UM, Minneapolis; (4) UT, Houston

Variants in *TCF7L2* are consistently associated with type 2 diabetes but few studies have assessed the effects of *TCF7L2* on impaired fasting glucose (IFG), especially among African Americans. We investigated the association between rs7903146 and IFG in 2,961 African American and 9,585 Caucasian adults in the ARIC cohort.
At baseline, IFG, defined as fasting glucose 100-125 mg/dl without a prior diabetes diagnosis or treatment, was present in 1,227 (41.44%) African American and 3,782 (39.46%) Caucasian ARIC participants. Generalized linear models with a log link and binomial distribution were used to estimate prevalence ratios of IFG in comparison to those with normal fasting glucose (<100mg/dl). Using an additive model, heterozygote CT and homozygote TT individuals had higher prevalence ratios (95% CI) of IFG compared to CC individuals in Caucasians [1.07 (1.03, 1.11), 1.15 (1.07, 1.24), respectively], adjusting for age, gender and study center. In contrast, no association between rs7903146 and IFG was noted in African Americans. Our study demonstrates a role of *TCF7L2* for IFG in Caucasians. The absence of association between IFG and rs7903146 in African Americans may reflect distinct patterns of linkage disequilibrium in this population, limited power to detect a genetic effect, or confounding by unmeasured covariates that is differentially distributed in African American and Caucasian participants, all of which warrant further investigation.

182
Genomic dissection of preferential amplification/hybridization based on three large-scale genome projects
H.-C. Yang (1), L.-H. Li (2), M.-C. Huang (1), W.-H. Pan (2)
(1) Institute of Statistical Science, Academia Sinica, Taiwan, (2) Institute of Biomedical Sciences, Academia Sinica, Taiwan

The coefficient of preferential amplification/hybridization (CPA) is used to improve estimation accuracy of allele frequencies. Here, we characterize genomic patterns of CPA based on samples from three genome projects. The first dataset consists of 367 and 448 Taiwanese samples genotyped using the Affymetrix Human Mapping 100K Set and 500K Set, respectively, from the Taiwan Han Chinese Cell and Genome Bank. The second dataset consists of 175 and 198 hypertension patients genotyped using the 100K Set and

500K Set, respectively, from the Taiwan Young-Onset Hypertension Project. The third dataset consists of 270 samples genotyped using both the 100K Set and 500K Set from the HapMap Project. We calculate CPA based on a proposed unbiased estimator and examine the relationship between CPA and factors including: (1) sample size; (2) time of data acquisition; (3) effects of laboratories; (4) gene chip type; (5) phenotypic status; (6) ethnicity effects; and (7) molecular features. We have constructed large public CPA databases. The results are summarized as follows: (1) CPA variance decreases significantly with increasing sample size; (2) CPA is independent to genotyping time/laboratories; (3) the 100K and 500K gene chips yield similar CPA patterns; (4) genomic distributions of CPA between phenotypic groups and between ethnic groups are positively correlated; (5) CPA distributions vary by GC content and genotypes; and (6) log-normal distributions capture the distributions of CPA well.

183

A Characterization of the Parameter Space for High-order Epistasis

W. Yang, C.C. Gu

Evidence from real studies suggests a more prominent role by interactions in complex diseases. In particular, gene-gene interactions (epistasis) have made substantive contributions to the development of diabetes, hypertension, and other human diseases. Traditional analysis paradigm focuses on the "main" (marginal) effects and treats interactions as deviance from the main effects, with less power. In a typical genome-wide association (GWAS) study, data were collected jointly, but the analyses have been carried out individually on each marker. However, interactions among genetic variants may or may not lead to detectable marginal effects. Therefore, it is important to know, for given set of marginal effects, how much variability there is in terms of interactions. One difficulty is the high-dimensional nature of the parameter space: penetrance tables for 6 bi-allelic loci require traversing the space of 729-dimension. We propose a novel model to characterize multilocus penetrance for arbitrary marginal constraints. We show that the parameter space of joint penetrance can be characterized by a set of canonical form of multilocus penetrance. Using this canonical expression of penetrance, one may effectively sample the space of high-order epistasis conformal to given marginal penetrance of any number of loci. We describe several computational algorithms to do so and demonstrate their utility by simulation studies of 6-loci disease models under multiple scenarios with varying level of marginal effects. It seems that the new method of penetrance modeling provide a viable means for systematic treatment of the vast space of parameter space of high-order gene-gene interactions, and may facilitate effective simulation analysis of gene-gene interactions in the setting of GWAS study.

184

Effects of ACE D/I Polymorphism and 11 SNPs Heterozygosity in the Same Gene on Health Risks in Framingham Study

A.I. Yashin, I.V. Culminskaya, K.G. Arbeev, S.V. Ukraintseva, I. Akushevich, A. Kulminski
Center for Population Health and Aging, Duke University, USA, aiy@duke.edu

ACE D/I polymorphism has been broadly studied for its associations with longevity and common disorders, such as hypertension. We compared the effect of D/I polymorphism with the effects of homo/heterozygosity in set of 11 SNPs in the ACE gene on risks of major diseases (cancer, CVD, diabetes) and impaired physiological states (hypercholesterolemia and obesity) in 776 adult participants of the Framingham Heart Study. Analyses showed that heterozygotic D/I females had a higher mortality from cancer and CVD after age 80, and the lower risk of becoming overall unhealthy after age 30, as compared with D/D and I/I homozygotes. Males with D/D genotype had higher risks of obesity (BMI $\geq$ 30 kg/m2) and hypercholesterolemia ($\geq$ 240 mg/dl) compared to other genotypes. In females, the effects were opposite. Analysis of frequency of genotypic classes for the set of SNPs showed that heterozygote at all sites of this set was the most common genotype in the analyzed sample. Individuals heterozygous by all 11 SNPs (both males and females) had significantly and substantially reduced risks of hypertension, obesity, cancer and diabetes (more than by 50% in some ages) as compared with the rest of the sample, especially at ages above 60. Overall, the effects of heterozygosity by 11 SNPs in ACE gene on health risks appeared to be much more pronounced than those of D/I polymorphism. Possible reasons for the overall health benefit of the heterozygosity by these 11 SNPs will be discussed.

185

Interacting Effect of ACE D/I Polymorphism and Smoking on Health Risks in Framingham Study

A.I. Yashin, S.V. Ukraintseva, K.G. Arbeev, I.V. Culminskaya, I. Akushevich, and A. Kulminski
Center for Population Health and Aging, Duke University, USA, E-mail: aiy@duke.edu

Smoking is well known for its negative health effects, including on CVD. These effects may, however, be modulated by a presence of particular genotype(s). The ACE D/I polymorphism is known for its associations with CVD, and we suggested that its effect may interact with that of smoking. We evaluated the age-specific effects of the D/I polymorphism and smoking on six health-related traits including diabetes, high cholesterol ($\geq$ 240 mg/100mL), hypertension (systolic/diastolic blood pressure $\geq$ 160/100 mm Hg), cancer, overweight or obesity (body mass index $\geq$ 25 kg/m2), and acute coronary heart disease (ACHD) in sample of 3,058 adult participants of the Framingham Heart Study. The analyses showed that the effect of smoking on health risks was genotype-specific for most studied conditions. Particularly, smoking increased risk of diabetes in individuals with ACE I/I genotype, but decreased that risk for D/I genotype (at ages above 60) in both sexes. For some disorders, the joint effect of genotype and smoking was sex-specific. Smoking decreased the risk of hypertension for D/D genotypes in females. In males the protective effect was evident for I/I carriers, while for other genotypes (D/D and D/I) smoking

increased the risk. Risk of overweight or obesity was lower in smokers with I/I genotype for both sexes. For males only smoking increased the risk of obesity in D/D carriers but decreased that risk in D/I carriers. We conclude that genetic background is significant modifier of health risks influenced by

186

Bioinformatic approach for selecting candidate genes: a case study of nicotine dependence

R.K. Yu, S.S. Shete

Integrating existing abundant biomedical information from various sources as prior knowledge for selecting genes and SNPs can improve the efficiency of genetic association studies. We programmed in Perl a literature mining system to systematically query, retrieve, update and summarize the key information from PubMed (and/or other literature databases). We exhibit utility of this approach using study of nicotine dependence. Using "nicotine dependence (addiction)" as a key word, we identified 5206 abstracts.
From these abstracts, we used 8 different key words related to nicotine dependence phenotype and obtained 78 papers, and using 9 genetic keywords we obtained 64 papers. From these papers, we were able to generate a list of genes or genomic regions of interest. Using tools such as Pathway Studio, we built up a biological network of the genes and novel candidates that we didn't encounter in the literature. We ranked these candidate genes using several approaches such as sample sizes used in the paper, statistical significance, number replications etc. Thus, we obtained 50 genes for further review and evaluation. Further probing these genes with additional databases such as NCBI's dbSNP and HapMap, we selected 28 genes and 429 SNPs. It is important to note that such approach heavily relies on the quality and reliability of the prior knowledge. For example, the SNP rs1800487 showed significant association with dependence and was considered to be in the D2 dopamine receptor (DRD2) gene. But dbSNP database does not reflect this SNP in DRD2. Later molecular biology findings proved that this SNP is in exon 8 of gene ANKK1. This is one of the major areas we consider for further improvement.

187

Bayesian Classification to Identify Epistases

M. Zhang (1), D. Zhang (1), M.T. Wells (2)
(1) Dept. of Statistics, Purdue University, USA, (2) Dept. of Biological Statistics and Computational Biology, Cornell University, USA

Simultaneous identification of genetic markers for both main and epistatic effects raises the statistical issue of variable selection with high dimensional low sample size data. We propose a two-step Bayesian classification approach to handle such data using a multiple regression model. Due to the fact that most predictors have no effect and that the nonzero effects may be asymmetric, a mixture-of-three-component prior distribution is proposed for each coefficient. More specifically, the prior distribution includes a point mass at zero, and two truncated normal distributions are specified for the positive and negative effects respectively. A Gibbs sampling algorithm is developed to stochastically search for significant variables among a large number of candidates.
We did a simulation study with 221 observations and 1596 predictors (56 main and 1540 epistatic effects), which include two (and three) nonzero coefficients for main (and epistatic) effects. Among the 100 simulated datasets, the two main effects can be correctly detected in 90 and 100 datasets, and the three epistatic effects can be detected in 90, 90, and 96 datasets respectively. The simulation study revealed that it is easier to detect larger effects as well as epistases that have main effects. The method has been successfully applied to quantitative trait loci (QTL) mapping and we are investigating its utility for human association studies with SNP markers in candidate genes.

188

Correlation Matrix Diagonal Segmentation (CMDS): A Genome-wide Approach for Identifying Recurrent DNA Copy Number Alterations across Cancer Patients

Q. Zhang (1), L. Ding (2), A. Kraja (1), I. Boreki (1), M.A. Province (1)
(1) Division of Statistical Genomics, Washington University School of Medicine, USA
(2) Genome Center, Washington University School of Medicine, USA

DNA copy number alteration (CNA) is one of the significant hallmarks of genomic abnormality in tumor cells. Identification of CNAs in cancers may provide an important insight into the molecular mechanism of oncogenesis and produce useful information for the diagnosis and treatment of cancer patients. We propose a genome-wide approach, Correlation Matrix Diagonal Segmentation (CMDS), for identifying recurrent DNA copy number alterations (RCNAs) which take place at the same chromosomal region across multiple samples. With no need of data discretization for individual samples and using a diagonal transformation strategy, CMDS significantly reduces computational burden and therefore is particularly suitable for genome-wide and large-population-based analysis. We use simulated data to investigate the statistical power of CMDS under a variety of configurations, which shows that CMDS achieves higher power compared with typical discretization-based approaches, especially when the amplitude of RCNA is small. We apply CMDS to the Affymetrix and Illumina DNA array data of over 1000 samples of matched tumor and normal tissues from lung and brain cancer patients in the Tumor Sequencing Project (TSP) and The Cancer Genome Atlas (TCGA), showing that CMDS is able to identify multiple chromosomal regions with RCNAs in lung and brain cancers.

189

Extended Homozygosity Score Tests to Detect Positive Selection for Genome-wide Scans of Human Genome

M. Zhong (1), K. Lange (2), J. C. Papp (2), and R. Fan (1)
(1) Department of Statistics, The Texas A&M University, USA, (2)
Department of Human Genetics, University of California, Los Angeles, USA

At the population level, recent positive selection can act to increase the frequency of advantageous alleles of genetic traits and can lead to high levels of linkage disequilibrium among genetic loci and nearby genetic markers. It is important to develop novel statistical methods to detect regions of the human genome where departures from the neutral model of human gene flow occur. In this article, we develop score test statistics to test the excess homozygosity. Three test cases are considered: (1) genotype-based test; (2) hidden Markov model; (3) haplotype-based test. The genotype-based test tests both linkage equilibrium and Hardy-Weinberg equilibrium. The haplotype-based test solely tests the excess homozygosity while linkage disequilibrium is allowed. The genotype-based test can be very sensitive, and most of test statistics can be significant. This will give too many false positives suggesting positive selection. On the other hand, the haplotype-based test may lead to few significant results. The hidden Markov model can provide results which stand between those of genotype-based test and haplotype-based test. The genotype-based test and haplotype-based test are two extreme cases, and the hidden Markov model stands between the two. We evaluate the robustness of the three test cases via type I error calculation and comparis

190

Single-Marker and Haplotype Analyses for Detecting Imprinting Effects in Families with Both Parents and Families with One Parent

J.Y. Zhou (1), W.K. Fung (1), S. Lin (2), Y.Q. Hu (1)
(1) Dept. of Statistics & Actuarial Science, University of Hong Kong, China, (2) Dept. of Statistics, Ohio State University, Columbus, Ohio, USA

Genomic imprinting is important in genetic trait study. Some statistical methods may be invalid or fail to detect linkage or association for imprinted genes. For case-parents trios, the parental-asymmetry test (PAT) is simple and powerful in detecting imprinting. Meanwhile, haplotype analysis is generally advantageous over single-marker analysis in complex trait study. As such, HAP-PAT, an extension of PAT, was constructed in haplotype analysis. However, it is common to collect families with both parents and those with only one parent. In this paper, when only one parent is available for each family, we develop 1-PAT to test for imprinting using single marker analysis. Combining families with both parents and those with one parent, C-PAT is proposed. We also introduce HAP-1-PAT and HAP-C-PAT to test for imprinting using haplotype analysis. A permutation procedure is devised to determine the significance of HAP-1-PAT and HAP-C-PAT. The validity of the statistics is verified by simulation. A power study shows that using the additional information from families with one parent in the analysis greatly improves the power of the tests, compared to that based on families with both parents. Also, utilizing all affected children in each family, the proposed tests have a higher power than when only one affected child from each family is selected. Furthermore, there are significant gains in power from haplotype analysis compared to single-marker analysis.

191

Maternal and embryonic genotypic interactions and risk for selected structural birth defects

H. Zhu (1), M.S. Gallaway (2), K. Waller (2), M. Canfield (3), R.H. Finnell (1)
(1) Center for Environmental and Genetic Medicine, Institute of Biosciences and Technology, TAMHSC, (2) School of Public Health, UTHSC, (3) Birth Defects Epidemiology and Surveillance Branch, Texas Department of State Health Services

Neural Tube Defects (NTDs) are a group of common, structural malformations that are associated with excess morbidity and mortality. A specific etiologic agent cannot be identified in the majority of individuals with NTDs, and in this group of patients the condition is believed to be a genetically complex trait. Both maternal diabetes and prepregnancy obesity have long been identified as significant independent risk factors for NTDs. Evidence accumulated to date suggests that the excess of NTDs among infants born to obese and to diabetic mothers may share some common underlying mechanisms. We investigated the genetic susceptibility to the induction of NTDs as a result of altered maternal glucose homeostasis and embryonic glucose transport related mechanisms in a Texas population using both case-trios and case-control study designs. Child-bearing age women who are genetically susceptible to diabetes/obesity may have altered glucose homeostasis, even if they are non-diabetic; therefore, producing a highly compromised in utero environment and exposing the developing embryos to teratogenic molecules (e.g., extra glucose flux). Genetic factors predisposing the developing embryos themselves to diabetes/obesity-related teratogenicity may modify any given infants' risk of having an NTD. Interactions between "maternal" genotypes and "embryonic" genotypes are evaluated.