Genetic Epidemiology 33: 752–835 (2009)

ABSTRACTS FROM THE

EIGHTEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

1

Enriching the Gold Dust: Extreme-Value Based Genome-Wide Association in the Post-GWAS Era

Dalin Li (1), David V. Conti (2)
(1) University of Southern California
(2) david.conti@keck.usc.edu

Based on the "common disease-common variant" hypothesis, great progress has been made in recent GWAS. However current findings are far from fully explaining the heredity of the traits studied, suggesting rare variants may contribute significantly to the genetic predisposition of human traits. With traditional GWAS design very large sample size is required to detect rare variants and the cost would be forbidding, particularly when we might rely on sequencing to explore the rare variants across the genome. Here we propose the extreme-value based GWAS in which only individuals with extreme phenotypes are used for whole genome genotyping or sequencing. Our simulations show that this extreme-value design is highly efficient in detecting rare variants with a sample size 10 to 20 times smaller than the comparable full study design. Furthermore the impact of bias effects does not increase correspondingly and the disturbance from unknown confounding and measurement bias can be significantly reduced. Moreover with the small sample size in the extreme-value design, it would be practically feasible to combine the information from the DNA sequence, DNA or histone methylation as well as RNA expression in the study. We propose an analysis framework for the extreme-value design based on maximum likelihood theory. A corresponding power calculation approach and a guideline for optimizing the extreme-value design conditional on the phenotyping/genotyping cost ratio are further proposed.

2

Fine Mapping of Common and Rare Variants Associated With Low-Density Lipoprotein Cholesterol (LDL-C) Via Sequencing Candidate Loci Following Genome-Wide Scans

Bingshan Li (1), Yun Li (1), David Schlessinger (2), Samer Najjar (2), Angelo Scuteri (2), Ed Lakkata (2), Serena Sanna (2), Mike Boehnke (1), Goncalo Abecasis (1), Manuela Uda (2)
(1) Center for Statistical Genetics, Dept. of Biostatistics University of Michigan
(2) Istituto di Neurogenetica e Neurofarmacologia (INN), Consiglio Nazionale delle Ricerche, Italy

Coronary artery disease is one of leading causes of morbidity and mortality in developed countries and strong associations have been established between lipoprotein levels and coronary heart disease. Our previous studies of >8,000 individuals through genome-wide association scans identified a number of loci associated with LDL-C levels, including previously reported and also newly implicated loci. To further understand the genetic contributions of both common and rare variants to the LDL-C level, we sequenced exons of 9 genes in associated loci in 256 unrelated Sardinian individuals with either extremely low or high LDL-C levels, along with 120 HapMap samples. Among all variants identified, 71% (81/121) nonsynonymous and 56.3% (40/71) synonymous mutations have frequency below 1%. In addition, two frame shift (in *APOB*) and two truncation mutations (in *PCSK9*) were identified. Comparisons between high LDL-C and low LDL-C groups showed that rare coding variants are enriched in one of the two groups for a set of genes (*APOB*, *LDLR*, *PCSK9*, *SORT1*). To increase power of detecting associations of variants in coding regions with LDL-C levels, we are using imputation to extend our findings to additional genotyped individuals in our 6148 sample Sardinian cohort. Equipped with this larger amount of data after imputation, fine mapping and evaluation of potential functional variants should be achieved with greater power.

3

An Integration of Genome-Wide Association Study and Gene Expression Profiling to Prioritize the Discovery of Novel Susceptibility Loci for Osteoporosis Related Traits

Yi-Hsiang Hsu (1), M. Carola Zillikens (2), Scott G. Wilson (3), Charles R. Farber (4), Serkalem Demissie (5), Estelle N. Bianchi (6), Liming Liang (7), J. Brent Richards (8), Karol Estrada (2), Yanhua Zhou (5), Nicole Soranzo (9), Atila van Nas (10), Miriam F. Moffatt (11), Guangju Zhai (12), Albert Hofman (13), Joyce B. van Meurs (2), Roger I. Price (3), L. Adrienne Cupples (5), Aldons J. Lusis (14), Eric E. Schadt (15), Serge Ferrari (6), André G. Uitterlinden (2), Fernando Rivadeneira (2), Tim D. Spector (12), David Karasik (1), Douglas P. Kiel (1)
(1) Hebrew SeniorLife Institute for Aging Research and Harvard Medical School, Boston, 02131 MA
(2) Department of Internal Medicine, Erasmus MC, Rotterdam, 3015GE, The Netherlands
(3) Departments of Endocrinology & Diabetes and Medical Technology & Physics, Sir Charles Gairdner Hospital, Western Australia
(4) Dep. Medicine, Cardiovascular Medicine and Center for Public Health Genomics, University of Virginia, Virginia
(5) Department of Biostatistics, School of Public Health, Boston University, Boston, MA, 02118
(6) Service of Bone Diseases, Department of Rehabilitation and Geriatrics, University Geneva Hospital, Switzerland

Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20463

(7) Center for Statistical Genetics, Department of Biostatistics, School of public Health, Ann Arbor, Michigan 48109-2029
(8) Departments of Medicine and Human Genetics, Lady Davis Institute, McGill University, Montreal, QC, Canada H3 T 1E2
(9) Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom
(10) Department of Human Genetics, David Geffen School of Medicine at University of California, Los Angeles, CA 90095
(11) National Heart and Lung Institute, Imperial College London, London SW3 6LY, United Kingdom
(12) Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom Se1 7EH
(13) Department of Epidemiology, Erasmus MC, Rotterdam, 3015GE, The Netherlands
(14) Department of Medicine, Department of Human Genetics, Department of Microbiology, Immunology, and Molecular Genetics, UCLA
(15) Rosetta Inpharmatics/Merck, Inc., Seattle, WA 98109

Although genome-wide association study (GWAS) is a power tool, the identification of disease-susceptibility genes by means of statistical significance provides limited information to predict their biological processes involved in diseases' pathophysiology. To overcome this challenge, we integrated expression profiling experiments with GWAS. We first performed GWAS for osteoporosis-related traits (bone mineral density and hip geometry indices) in the Framingham Osteoporosis Study and then replicated top findings in two additional studies. Meta-analyses were performed in 7634 women and 3657 men. To identify potential biological links to bone metabolism and prioritize candidate genes, we (1) analyzed the expression QTL (eSNP) in several human tissues; (2) conducted expression profiling in cellular models for parathyroid hormone stimulated osteoclastogenesis and for osteoblastogenesis of embryonic stem cells; (3) performed likelihood-based causality model selection (LCMS) in a different mice experiment to identify genes causally related to bone phenotypes; and (4) constructed functional interaction networks based on biological information from available bioinformatics databases. We have discovered four novel loci and highlighted the efficiency of subsequent functional characterization using these experiments to prioritize candidate genes and generate new hypotheses for further investigation.

4

Fine Mapping of Colorectal Cancer Low Penetrance Susceptibility Loci

Luis G. Carvajal-Carmona (1), Malcolm Dunlop (2), Jean-Baptiste Cazier (1), Richard S. Houlston (3), Ian P.M. Tomlinson (1)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford
(2) Institute of Genetics and Molecular Medicine, University of Edinburg, United Kingdom
(3) Section of Cancer Genetics, Institute of Cancer Research, Sutton, United Kingdom

Colorectal cancer (CRC) is the third most common cancer in the western world. We have recently used genome-wide association studies to identify ten common low-penetrance CRC risk loci. These ten loci account for about 5% of the familial risk of CRC. In an attempt to identify the causal alleles tagged by some of these loci, we genotyped 875 SNPs in six of these regions in English ($n = 3000$ cases and 3000 controls) and Scottish ($n = 2000$ cases and 2000 controls) cohorts. To pinpoint the location of disease-causing variants, we used data from the Hapmap samples to generate in-silico genotypes for ~700 additional SNPs in these samples. We calculated association p-values for these SNPs in the English and Scottish samples and, using discovery phase hap550 SNP English and Scottish data ($n = 1000$ cases and 1000 controls from each population), we carried out an association meta-analysis that led to the identification of novel SNPs showing stronger associations than those detected by the original tagging SNPs.
In addition to our single-point association testing approach, we also used the haplotype analysis method incorporated in the HapCluster software in an attempt to identify the regions that could harbour the causal alleles. In the present report, we will present and compare the results obtained from both the single-point and haplotype-based analyses and discuss the advantages and pitfalls of using two-closely related European populations to fine map GWA loci.

5

A unified Mixed Effects Likelihood Framework for Detecting Associations With Rare Variants Using Sib and Unrelated Individuals With Extreme Quantitative Phenotypes: Application to Next Generation Sequencing Data.

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

In designing association studies to identify rare Quantitative Trait Loci (QTL), sampling related and unrelated individuals with extreme quantitative trait values (QTVs) are valuable for enriching rare causal variants. Existing methods to analyze common QTL are underpowered for the analysis of rare QTL (rQTL) and inflexible for modeling relative phenotypic correlations due to multiple rare causal variants loci. We propose a flexible likelihood framework with mixed effects for modeling extreme trait genetic associations with rQTL (MEGA-rQTL) for the analysis of related and unrelated individuals with extreme QTVs. MEGA-rQTL detects associations with rQTL through likelihood ratio tests, and parameters of genetic interests such as heritability and sibling residual correlation can be efficiently estimated. We investigated the power and efficiency of the MEGA-rQTL method, for 7 commonly used prospective selective sampling strategies. Simulation was carried out via coalescence theory using parameters estimated from population genetic data and models for clinically relevant traits. We demonstrate that analyzing sibpairs with extreme QTVs or using one sib per sibpair with extreme QTVs are consistently more powerful than using unrelated individuals with extreme QTVs. In conclusion, Mega-rQTL is a powerful approach to analyze next generation sequence data to map QTL due to rare variants by combining data from related and unrelated individuals with extreme phenotypes.

6

How to Account for Gene-Environment Interaction When Testing for Association With a Reference Control Panel?

Rémi Kazma (1), Marie-Claude Babron (2), Emmanuelle Génin (2)
(1) Univ. Paris-Sud, Faculté de Médecine, Le Kremlin-Bicêtre, France; Inserm UMR-S946, Paris, France
(2) Inserm UMR-S946, Paris, France; Univ. Paris-Diderot, Paris, France

The use of a reference control panel in Genome-Wide Association Studies is an interesting solution to reduce costs. In such designs, same individuals serve as controls against several sets of cases with different diseases. Relevant environmental factors (E) are usually only available in cases making it difficult to account for potential gene-environment (GxE) interactions. However, neglecting an existing interaction with E may hinder the detection of a genetic factor (G). There is thus a need to develop tests to account for GxE interaction when information on E is available only in cases.
In this context, we propose a novel method based on a multinomial logistic regression model to contrast both exposed and unexposed case samples against a population control sample with no information on E. For each case group, a genetic effect size parameter is estimated and a 2-df likelihood ratio tests jointly G and GxE interaction effects. To evaluate its performance, we simulated samples of 500 cases and 500 population controls under different models of GxE interaction. In presence of an interaction, our approach outperforms the one that ignores E both in cases and controls and only tests for G but also, and more interestingly, the approach that takes into account E in both samples and tests jointly G and GxE interaction. Furthermore, the proposed approach is particularly interesting in the situation of crossed interactions where the susceptibility allele changes depending on the E status.

7

Detecting Association with Rare Genetic Variants in Common Diseases

Yali Li (1), Tao Feng (1), Robert C Elston (1), Xiaofeng Zhu (1)
(1) Case Western Reserve University

Current Genome-Wide Association Studies (GWAS) have successfully detected many genetic variants contributing to common diseases but not rare ones. Here we propose two approaches to detect rare genetic variants based on current GWAS designs. In the first approach, we show that we can detect and classify together rare risk haplotypes using a relatively small sample, and then test association in the rest of the sample. We have applied the method to the Wellcome Trust Case Control Consortium (WTCCC) coronary artery disease and hypertension data, the latter being the only trait for which no genome-wide association evidence was reported in the original WTCCC study, and identified one gene associated with hypertension and four associated with coronary artery disease at a genome-wide significance level of 5%. The second approach is a haplotype-based truncated product method (HTPM). A *P*-value combination method from testing for the multiple hypotheses is borrowed, but is used for the purpose of clustering the information on rare risk haplotypes. Our simulation studies demonstrated that HTPM has increased power for detecting the association between rare variants and diseases, compared with the first approach. Application of HTPM to the WTCCC data has replicated the previous findings of associated gene. These results suggest that searching for rare genetic variants is feasible and can be fruitful in current genome-wide association studies, candidate gene studies or resequencing studies.

8

Analysis of Population Based Genetic Association Studies Using Generalized Propensity Scores

Huaqing Zhao (1), Timothy R Rebbeck (1), Nandita Mitra (1)
(1) University of Pennsylvania School of Medicine

Propensity scores are commonly used to address confounding in observational studies. However, they have not been widely adapted to deal with confounding such as population stratification and/or admixture stratification in genetic association studies. Recently, we developed a genomic propensity score (GPS) approach to correct for bias due to PS that considers both genetic and non-genetic factors under the assumption of a dominant genetic model. We now propose an extended GPS (eGPS) approach that allows one to estimate the effect of a genotype under a wide range of genetic models (dominant, recessive, additive) and also adjusts for confounders such as random null markers, admixture informative markers (AIMs) and patient characteristics. We validate the eGPS method by carrying out extensive simulation studies. Our results show that eGPS can adequately adjust and consistently correct for bias due to confounding. Under all simulation scenarios, the eGPS method yielded estimates with bias close to 0 (mean = 0.016, standard error = 0.011). Our method also preserves statistical properties such as coverage probability, type I error and power. We illustrate this approach in a case-control study of prostate cancer and *CYP3A* genotypes and a case-control study of testicular germ cell tumors and *KITLG* and *SPRY4* susceptibility genes. We conclude that our method provides a novel and broadly applicable analytic strategy for obtaining less biased and more valid estimates of genetic associations.

9

Legacy of Mutiny on the Bounty: Founder Effect and Admixture on Norfolk Island

Stuart Macgregor (1), Clair Bellis (2), Brian McEvoy (1), Rod A. Lea (3), Hannah Cox (2), Tom Dyer (4), John Blangero (4), Peter M. Visscher (1), Lyn R. Griffiths (2)
(1) Queensland Institute of Medical Research, Brisbane, QLD 4006, Australia
(2) Genomics Research Centre, Griffith Institute for Health and Medical Research, Griffith University, Southport, QLD 4215, Australia
(3) Kenepuru Science Centre, Institute of Environmental Science and Research, Wellington, New Zealand
(4) Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas, 78227-5301

The population of Norfolk Island, located off the eastern coast of Australia, possesses an unusual and fascinating history. Most present day islanders are related to eleven male Caucasian and six female Polynesian 'Bounty' mutineer founders. By examining a single large pedigree of 5742 individuals, spanning >200 years, we analyzed the influence of admixture and founder effect on various cardiovascular disease (CVD) related traits. On average one third of the genome of present day islanders (single large pedigree individuals) is derived from these 17 initial founders. The proportion of Polynesian ancestry in present day individuals, as estimated from the pedigree, was found to significantly influence several CVD related traits including total triglycerides, body mass index and blood pressure, with Polynesian ancestry conferring greater CVD risk. Marker derived homozygosity agreed with measures of inbreeding derived from pedigree information. Founder effect (inbreeding and marker derived homozygosity) significantly influenced height. We also examined genomic admixture proportions using ancestry informative markers (AIMs). There was a strong correlation (0.72) between genetic and pedigree based estimates of ancestry, with AIMs suggesting islanders are on average 82% Caucasian and 18% Polynesian. In conclusion, both founder effect and extreme admixture have substantially influenced the genetic architecture of a variety of CVD related traits in this population.

10

Association of Genomewide Newborn DNA Methylation Patterns with Maternal Diet, Birth Weight and SNP Variation

Ronald M. Adkins (1), Julia Krushkal (1), Fridtjof Thomas (1), John Morrison (2), Everett F. Magann (3), Fran Tylavsky (1), Grant Somes (1)
(1) University of Tennessee Health Science Center
(2) University of Mississippi Medical Center
(3) Naval Medical Center at Portsmouth

Epigenetics is important to fetal growth abnormalities, but little is known about its contribution to normal fetal growth variation. In model organisms maternal diet during gestation dramatically affects the pattern of epigenetic modifications, but how relevant this is to the dietary variation observed among pregnant women is unclear. We are conducting a longitudinal cohort study of (epi)genetics, diet, and development of 1,500 women and children from the second trimester of pregnancy through age three of the newborn. We performed genomewide SNP genotyping and DNA methylation profiling to investigate how the gene-by-gene DNA methylation patterns relate to the intake of nutrients key to DNA methylation, to size for gestational age of the newborns, and to SNP genotypic patterns. We detected highly significant correlations between birth weight and average levels of DNA methylation in several genes, many of which have not previously been associated with fetal growth control. We also found significant correlations between DNA methylation and variation in the intake of some nutrients, including at some genes implicated in newborn dysplasias. Instances of SNP genotype × DNA methylation interactions are also implicated. Our results suggest that birth weight and the extent of DNA methylation at some genes are correlated. Additionally, it appears that normal dietary variation among pregnant women is sufficient to influence the average patterns of DNA methylation at a subset of genes.

11

The Power of Family Data in Association and Linkage Studies

E. Warwick Daw (1)
(1) Washington University

The utility of using family data is examined where the "common trait, common gene" hypothesis fails. This hypothesis supposes that a moderate amount of phenotype variation is due to one mutation on a common genetic background. Alternative hypotheses include multiple rare mutations in one gene, a site mutated in many different backgrounds, or a multitude of rare mutations in different genes. In each of these situations, use of a familial sample could potentially reduce the sources of variation within the sample and thus improve power to detect genes that have variation in the sample. To examine this issue, traits were simulated with 10, 100, and 1000 genes in an interacting network with environmental factors. Environmental factors included both individual and familial factors. Trait genes had a spectrum of effect sizes: none with >5% variance, and some substantially less. We simulated sample sizes of 1000 individuals with 4 different pedigree structures: large pedigrees of 100 individuals, moderate pedigrees of 20, nuclear families of 6, and sibling pairs. These samples were analyzed with variance component linkage analysis, oligogenic segregation and linkage analysis, and familial association tests. We also simulated 1000 unrelated individuals for association analyses. We present the power to detect genes with each method under the different models, as well as the accuracy of the location, and make recommendations about when family data is most useful.

12

A Risk Prediction Algorithm for Familial Prostate Cancer Incorporating the Effects of Common Genetic Variants

Robert J. MacInnis (1), Antonis C. Antoniou (1), Rosalind A. Eeles (2), Ali A.A. Olama (1), John L. Hopper (3), Graham G. Giles (4), Douglas F. Easton (1)
(1) University of Cambridge
(2) The Institute of Cancer Research
(3) University of Melbourne
(4) The Cancer Council Victoria

GWAS have identified 15 SNPs that individually confer small relative risks of prostate cancer. This opens up the possibility for using variants in risk prediction. However, risk prediction algorithms based on SNPs alone are limited because they explain only a fraction of individual and familial risk of the disease. Using segregation analysis of families from the UK and Australia, the best model for the genetic susceptibility to prostate cancer was a mixed model of inheritance that included both a recessive major gene component and a polygenic component P that represents the total effect of a large number of genes each of small effect, where $P \sim N(0, \sigma_p^2)$. We extended this model to

incorporate the explicit effects of the 15 SNPs by decomposing the polygenic component into 2 parts: a component due to the known susceptibility loci $P_{obs} \sim N(0, \sigma^2_{obs})$, and a residual component $P_{res} \sim N(0, \sigma^2_{res})$. The resulting algorithm can be used to predict the probability of developing prostate cancer based on both SNP profiles and explicit family history information; e.g., the lifetime risk for a man aged 50 whose father was diagnosed with prostate cancer at age 60 and who is in the 95th percentile of the SNP profile distribution is 39%, compared with 18% for the same individual in the 5th percentile. This algorithm can be easily extended to incorporate further SNPs as they are identified, and can form the basis for identifying individuals at high risk of developing prostate cancer.

13

Targeted Resequencing of Over Twenty Loci Implicated in Colorectal Cancer

Mathieu Lemire (1), Vanya Peltekova (1), Quang Trinh (1), Lee Timms (1), Michelle Sam (1), Tanja Durbic (1), April Cockburn (1), Timothy Beck (1), Richard De Borja (1), Michelle Chan-Seng-Yue (1), Ada Wong (1), David D'Souza (1), Kamran Shazand (1), John D. McPherson (1), Thomas J. Hudson (1)
(1) Ontario Institute for Cancer Reseach, Toronto, Canada

Genome-wide association scans (GWAS) and meta-analyses looking for colorectal cancer (CRC) susceptibility loci have identified 10 common variants, each independently increasing the risk of developing the disease by 10%–25%. To produce an extensive catalogue of variations in CRC susceptibility regions, we resequenced 40 CRC sporadic cases, 40 controls, 25 probands from families showing the hereditary form of CRC along with 15 of their kin. We targeted 3.14 Mbp to be sequenced around the 10 GWAS loci and in genes known to be implicated in the hereditary form of CRC. DNA from the target regions were captured using microarray-based methods and were sequenced on Solexa next-generation sequencers. Prior to filtering and alignment, 3 billion paired-end reads (2×76 bp) were generated for a total of half a trillion nucleotides. After filtering of reads and only considering unique alignments, the depth of coverage in the target regions was 52.4 reads per base on average. We identified 14000 single-base substitutions and 1300 bi-allelic short indels, half of which were novel with low estimated allelic frequencies. These include 61 new non-synonymous variants, including 5 stop codons, 36 new synonymous variants and one novel splice-junction mutation. We are genotyping these variants in 2400 sporadic cases and 2400 controls on a custom high-density array, which among other things will allow the evaluation of the influence of rare variants on the risk of developing CRC.

14

Visualizing Chromosome Mosaicism and Detecting Ethnic Outliers by the Method of "Rare" Heterozygotes and Homozygotes (RHH)

Ralph McGinnis (1), Panos Deloukas (1), William McLaren (1), Michael Inouye (1)
(1) Wellcome Trust Sanger Institute

We present a novel approach for evaluating genotypes of a genome-wide association scan (GWAS) to visualize the mosaicism of ethnically admixed chromosomes and identify outliers whose ethnic ancestry is different or admixed compared to most other subjects in the GWAS. Each "ethnic outlier" is detected by counting a genomic excess of rare heterozygotes and/or homozygotes. The method also enables simple and striking visualization of non-Caucasian chromosomal DNA segments interspersed within the chromosomes of ethnically admixed individuals, thereby delineating chromosome mosaic structure. Our visualization method gives results similar to other visualization software based on hidden Markov or related models (e.g. SABER) but with much less computational time and burden. We show sensitive detection of ethnic outliers and visualization of chromosome mosaicism in subjects of the Wellcome Trust Case Control Consortium (WTCCC), in HapMap subjects and in simulated outliers of known ethnicity and admixture. The method's ability to precisely delineate chromosomal segments of non-Caucasian ethnicity has enabled us to identify previously unreported non-Caucasian admixture in several HapMap Caucasian parents and in many WTCCC subjects. Its simple visual discrimination of discrete chromosomal segments of different ethnicity implies that this method of rare heterozygotes and homozygotes (RHH) is likely to have diverse and important applications.

15

A Computational Evolution System for Open-Ended Automated Learning of Complex Genetic Relationships

Jason H. Moore (1), Doug Hill (1), Casey S. Greene (1)
(1) Dartmouth College

The failure of genome-wide association studies to reveal the genetic architecture of common diseases suggests that it is time to embrace the full complexity of the problem. To this end, we have developed an open-ended computational evolution system (CES) that makes no assumptions about the underlying genetic model and can learn through evolution by natural selection how to solve a particular genetic modeling problem. This is accomplished by providing the basic mathematical building blocks (e.g. +, −, *, /, LOG, <, >, =, AND, OR, NOT etc.) for models that can take any shape or form and the basic building blocks for algorithmic functions (e.g. ADD, DELETE, COPY, etc.) that can manipulate genetic models in a manner that is dependent on expert statistical and biological knowledge or prior modeling experience. Here, we introduce an additional layer to our CES approach that introduces noise into the training data (5%, 10%, 15% and 20%) to drive the discovery process toward models that are more likely to generalize. We show using simulated epistatic relationships in genome-wide data that the CES leads to significantly smaller models ($P < 0.001$) thus reducing false-positives and overfitting while maintaining a power of 97% to 100%. These results are important because they show how introduced noise in the data can yield more parsimonious models and reduce overfitting without the need for computationally expensive cross-validation.

16

Informing Disease Via Integrative Genetics and Systems Approaches in Human Postmortem Brain Tissue

Cliona Molony (1), Tao Xie (2), Joshua McElwee (2), Judy Zhong (2), Patrick Loerch (2), Keith Tanis (3), Jun Zhu (2), Chunsheng Zhang (2), Manikandan Narayanan (2), Valur Emilsson (2), David J. Stone (3), Eric E. Schadt (2)
(1) Rosetta Inpharmatics
(2) Rosetta Inpharmatics
(3) Merck Research Laboratories

We are identifying all SNPs that affect gene expression in the human brain through the most extensive postmortem brain study to date, genotyping 650k SNPs and profiling RNA in 3 regions from 1000 individuals. eSNPs that are known to affect RNA expression are more likely to be clinically relevant, given they are already associated with a biologically relevant phenotype. Thus by identifying SNPs that affect gene expression (eSNPs), we demonstrate that we can increase power to identify associations with clinical phenotypes in a genome-wide setting. Here, we obtained 3 regions of postmortem brain (Prefrontal Cortex, Visual cortex, and cerebellum) from 1000 individuals from the Harvard Brain Tissue Resource Center. Patient diagnoses included Alzheimer's ($n = 600$), Huntington's ($n = 200$), or control ($n = 200$), all confirmed by a neuropathology report. Signature analysis showed that these 3 regions in combination express 98% of the genes detectable in hippocampus. Gene-expression data in PFC detected an increase of inflammation related genes ($P = 5.3 \times 10–79$) and a decrease in synaptic transmission related genes ($P = 2.0 \times 10–53$) in AD cases compared to controls, whereas the cerebellum did not show these differences. Therefore the effect of SNPs on gene expression can be examined in both the normal and diseased setting. The resulting SNP set has been used to increase power in genome-wide scans and provides functional information on genes associated with disease.

17

A Natural Aggregation Function for Pathway/Network-Based Approaches to GWAS and Gene Expression Analysis

David Heckerman (1), Carl Kadie (1), Xiang Zhang (1), Jennifer Schymick (2), John Ravits (3), Bryan Traynor (2), Jennifer Listgarten (1)
(1) Microsoft Research
(2) NIH
(3) Benaroya Research Institute

We present an approach for identifying significant *sets* of SNPs in GWAS in a similar spirit to that of GSEA. We quantify the relationship between a set of SNPs and a phenotype by how well those SNPs predict the phenotype. Given SNP and phenotype data from a collection of individuals, we train a probabilistic model that predicts the phenotype based on SNPs using data from some individuals, and apply this trained model on the remaining data, yielding a probability distribution over the phenotype for each individual. We then test for associations between aspects of this distribution and the actual phenotype observations using standard methods to obtain a *p*-value for the SNP set. E.g. when the phenotype is binary, we can test for an association between the probability of having the phenotype and actually having the phenotype. The approach can also be applied to GSEA, where gene expression rather than SNPs are used to predict phenotype. We have applied this approach to GWAS with binary phenotypes using data from KEGG to define SNP sets, 50/50 train/test splits, lasso logistic regression for the prediction model, and the Mann-Whitney test for association. The approach yields uniformly distributed p-values on null data. When applied to a GWAS data set from the US for the disease ALS, three pathways are significant after correction for multiple tests: neuroactive ligand-receptor, complement, and axon guidance. These pathways are also significant in a data set from Italy.

18

Overcoming Data Quality and Copy Number Detection Issues in Genome-Wide CNV Association Studies

Christophe Lambert (1), Greta Linse (1), James Grover (1), Josh Forsythe (1), Doug Hawkins (2)
(1) Golden Helix, Inc.
(2) School of Statistics, University of Minnesota

Incorporating copy number variations into genome-wide association studies promises to explain more of the heritability of common diseases than that accounted for by SNPs alone. This potential goldmine however, has been plagued by myriad of technical and experimental challenges. We examine the most persistent issues observed in over 20 CNV GWAS studies conducted by us and our collaborators. These include huge batch effects, genomic waves, mosaicism, T-cell artifacts and poor signal-to-noise ratios, all of which can lead to false positives, negative CNV detection and subsequent association findings. To address these issues we describe a novel principal component analysis approach that simultaneously corrects for batch and wave effects and population stratification, while significantly improving signal-to-noise ratios. We address the challenge of lingering batch effects in CNV regions, and we describe optimal segmentation methods using dynamic programming to detect CN segment boundaries on either a per-sample (univariate) or a multi-sample (multivariate) basis. Unlike Hidden Markov Model methods, which assume the means of different CN states are consistent, optimal segmenting methods properly delineate segment boundaries in the presence of mosaicism, even at a single probe level, and with superior sensitivity and false discovery rates. We then outline several approaches to genome-wide scans for CNV association, demonstrating the utility of these methods on a series of large-scale GWAS.

19

Importance of Sequencing Rare Variants After a Genome-Wide Association Study (GWAS): the MC1R Gene, 16q24 Region and Melanoma Story (the GenoMEL Consortium)

Florence Demenais (1), Eve Corda (2), Jennifer Barrett (3), Mark Iles (3), Elizabeth M. Gillanders (4), Alisa M. Goldstein (5), Peter A. Kanetsky (6), Egbert Bakker (7), Timothy Bishop (3), Julia A. Newton-Bishop (3), Nelleke A. Gruis (7)

(1) INSERM U946, Paris
(2) Fondation Jean Dausset-CEPH, Paris
(3) Leeds Institute of Molecular Medicine, Leeds, United Kingdom
(4) NHGRI, NIH, Baltimore, MD
(5) NCI, NIH, Baltimore, MD
(6) University of Pennsylvania
(7) Leiden University Medical Centre, Leiden, The Netherlands

A melanoma GWAS identified association with 16q24, a region harbouring candidate genes (*CDK10*, *MC1R*). Three SNPs had independent effects, rs258322 (*CDK10*), rs4785763 (*AFG3L1*)and rs8059973 (*DBNDD1*) but none of the non-synonymous (NS) *MC1R* variants was present on the chip. To investigate whether the association signals might be accounted for by *MC1R*, this gene was sequenced in 1,805 GWAS subjects (918 cases, 887 controls). We used logistic regression to compare the strength of association when examining each SNP with/without each *MC1R* variant in the model and, conversely, for *MC1R* variant/SNP pairs. It was followed by stepwise regression and haplotype analysis with *MC1R* variants and SNPs. Among 75 *MC1R* variants, 9 NS variants had allele frequency $\geq 1\%$. Three variants, R151C, R160W and D294 H, had significant effect ($1.8 \times 10^{-11} \leq P \leq 1.5 \times 10^{-3}$). There was no longer evidence for association with rs258322 in presence of R151C and decreased evidence with rs4785763 in presence of R151C or R160W. Conversely, the association with R151C was reduced by rs258322 or rs4785763 and with R160W by rs4785763. Stepwise regression showed significant effect of R151C, R160W and D294 H ($3.1 \times 10^{-13} \leq P \leq 4.6 \times 10^{-5}$). Haplotype analysis demonstrated that the rs258322 signal was accounted for by R151C and the rs4785763 signal by R151C and R160W. Thus, ignoring rare variants can lead to incorrect inferences on the potential role of candidate genes carrying common SNPs identified by GWAS.

20

Lessons to be Learned From Genome-wide Interaction Analysis (GWIA)

Manuel Mattheisen (1), Michael Steffens (1), Tim Becker (1), Thomas Sander (2), Rolf Fimmers (1), Christine Herold (1), Daniela A. Holler (1), Costin Leu (2), Stefan Herms (3), Sven Cichon (4), Bastian Bohn (5), Thomas Gerstner (5), Michael Griebel (5), Markus M. Nöthen (4), Max P. Baur (1), Thomas F. Wienker (1)
(1) Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany
(2) Cologne Center for Genomics, University of Cologne, Germany
(3) Institute for Human Genetics, Department for Genomics, Life & Brain Center, Bonn, Germany
(4) Institute for Human Genetics, University of Bonn, Germany
(5) Institute for Numerical Simulation, University of Bonn, Germany

Here we present our experience in a full genome-wide interaction analysis (GWIA) gene × gene interaction analysis (1,559 individuals and 620,726 SNPs) from a case-control study of idiopathic generalized epilepsy. In particular we focus on two main issues: feasibility of analysis and pitfalls in SNP selection prior to the analysis.
Major bottlenecks in modern computers are memory and disk access. In a pre-processing step, the genome data was converted to binary form using 2 bits for each marker, resulting in a raw binary stream that consists of approximately 240 Mbytes, which can be easily stored in main memory. In addition the computational burden has been parallelized and the analysis performed on a multi-processor system. On our parallel system with its 256 processors, the complete two marker interaction analysis only required seven hours.
Notably, by far the most SNPs included in our top results of the GWIA did not show strong marginal effects and thus would have been lost if SNPs had been pre-selected on the basis of detected single-marker effects in the run-up of our analysis. Guided by these observations, one should be encouraged to make use of all available SNPs in the GWIA (without pre-selection of SNPs).
Results from GWAS are beginning to reveal a ''missing heritability'' in complex traits and diseases. Systematic, hypothesis-free analysis of epistatic interaction (GWIA) may help to close, at least in part, this increasingly recognized gap in heritability.

21

A Large Genome-wide Association Study of Glycated Hemoglobin Identifies Ten Common Variants not Mediated Through BMI

Eleanor Wheeler (1), Nicole Soranzo (1), Serena Sanna (2), Christian Gieger (3), Dörte Radke (4), Josee Dupuis (5), Elliot Stolerman (6), Nabila Bouatia-Naji (7), Claudia Langenberg (8), Inga Prokopenko (9), Manjinder S. Sandhu (10), Linda Kao (11), N. J. Wareham (8), Jose C. Florez (12), Manuela Uda (2), Inês Barroso (1), James B. Meigs (13), for the MAGIC investigators (14)
(1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom
(2) Istituto di Neurogenetica e Neurofarmacologia del CNR, Monserrato, Cagliari, Italy
(3) Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
(4) Institute for Community Medicine, Ernst Moritz Arndt University Greifswald, Greifswald, Germany
(5) Boston University School of Public Health, Boston, MA and National Heart, Lung, and Blood Institute's Framingham, MA
(6) Center for Human Genetic Research, and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston
(7) CNRS-UMR-8090, Institut Pasteur de Lille and Lille 2 University, Lille, France
(8) MRC Epidemiology Unit, Cambridge, United Kingdom
(9) WTCHG and OCDEM, Oxford, United Kingdom
(10) University of Cambridge, Cambridge, United Kingdom
(11) Department of Epidemiology and Medicine, Johns Hopkins University, Baltimore
(12) Massachusetts General Hospital and Harvard Medical School, Boston, MA and Broad Institute, Cambridge, MA, United States
(13) General Medicine Division, Massachusetts General Hospital, Boston

(14) Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)

Glycated hemoglobin (HbA1C), which results from the non-enzymatic glycation of hemoglobin molecules, is widely used in the clinical monitoring of diabetes control, and is increasingly used for the diagnosis of diabetes. Though the estimated heritability of HbA1C in populations of European ancestry is relatively high (~47%), known genetic factors only account for ~1.5% of the variance. To identify novel loci influencing HbA1C levels, we conducted a meta-analysis of ~2.5 M directly genotyped or imputed autosomal SNPs from 10 genome-wide association studies (Stage 1 cohorts), totaling 14,898 non-diabetic adults of European descent. When we examined the top 20 independent signals in 30,233 additional samples (Stage 2 cohorts) we obtained strong evidence for association ($P < 5 \times 10^{-8}$) at five loci, including *HK1* ($P = 1.53 \times 10^{-50}$), *GCK* ($P = 8.80 \times 10^{-20}$) and *G6PC2* ($P = 9.36 \times 10^{-18}$). Combining all available data from the Stage 1 and Stage 2 cohorts (45,131 individuals from 29 cohorts) we identified five additional loci reaching genome-wide significance, including *MTNR1B* ($P = 8.35 \times 10^{-11}$), a locus recently associated with fasting glucose and Type 2 diabetes risk. Further adjustment for BMI did not attenuate the associations. Of the ten association signals, five are in or near genes implicated in the regulation of glycemia, two more in iron homeostasis, and the remainder have unknown function. Our results could influence the use of HbA_{1C} in clinical practice.

22

GWAS Meets Microarray: Are the Results of Genome-Wide Association Studies and Gene-Expression Profiling Consistent? Prostate Cancer as an Example

Ivan P. Gorlov (1), Christopher Amos (1), Olga Gorlova (1), Christopher J. Logothetis (1)
(1) The University of Texas M. D. Anderson Cancer Center

Genome-wide association studies (GWASs) and global gene expression profiling (GEP) are two major technological breakthroughs that allow hypothesis-free identification of candidate genes associated with tumorigenesis. It is not obvious whether there is a consistency between the candidate genes identified by GWAS (GWAS genes) and those identified by GEP (GEP genes). We used the Cancer Genetic Markers Susceptibility database to retrieve single nucleotide polymorphisms from candidate genes for prostate cancer idenfied by two GWAS. In addition, we conducted a large meta-analysis of GEP data in normal prostate and prostate tumor tissue from eight studies. We identified 13,905 genes that were interrogated by both GWASs and GEPs. On the basis of P values < 0.01 from at least one GWAS, we selected 1,649 most significantly associated genes for functional annotation by the Database for Annotation, Visualization and Integrated Discovery (DAVID). We also conducted functional annotation analysis using the same top genes identified in meta-analysis of the gene expression data. The number of genes significant in both GWASs and GEPs was significantly higher than expected by chance. Genes involved in cell adhesion were overrepresented among both the GWAS and GEP genes. Results of these analyses suggest that combining GWAS and GEP microarray data is a more effective approach for identifying causal genes and pathways influencing cancer development than analyzing individual datasets.

23

Finding that Elusive Gene-Environment or Gene-Gene Interaction: Prioritizing SNPs for Quantitative Trait Interaction Testing

Guillaume Pare (1), Nancy R. Cook (2), Paul M. Ridker (2), Daniel I. Chasman (2)
(1) McMaster University
(2) Brigham and Women's Hospital

Whole genome association studies have identified many common genetic determinants of complex traits. However, few gene-gene and gene-environment interactions have been identified and validated. One of the main obstacles to interaction testing stems from multiple hypothesis testing. In this report, we show that under plausible interaction scenarios, the variance of a quantitative trait is expected to differ between the three possible genotypes of a biallelic SNP. Leveraging this observation with Levene's test of equality of variance, we propose a novel method to prioritize SNPs for subsequent gene-gene and gene-environment testing. Prioritization is independent of subsequent interaction testing and substantially reduces the burden of multiple hypothesis testing. We first use simulations to demonstrate that our method has increased power over conventional ones. Then, using data from our ongoing genome scan ($n = 20{,}628$) of two inflammatory markers, C-reactive protein (CRP) and soluble ICAM-1 (sICAM-1), we successfully apply our method to identify two novel interactions which replicate (combined P-values of 4.2×10^{-10} and 4.4×10^{-8}, respectively). Both CRP and sICAM-1 are plasma markers used in the prediction of cardiovascular disease and diabetes. The first interaction involves the leptin receptor SNP rs12753193 and body mass index (BMI) in the prediction of CRP. The second interaction involves the PNPLA3 SNP rs738409 and BMI in the prediction of sICAM-1.

24

Trait Prediction Using Multi-Locus Information: Psoriasis as a Model for Complex Disease Prognostics.

Steven J. Schrodi (1), Yonghong Li (1), Monica Chang (1), Veronica E. Garcia (1), Kristina Callis Duffin (2), Rajan P. Nair (3), Anne M. Bowcock (4), James T. Elder (3), Gerald G. Krueger (5), Charles M. Rowland (1)
(1) Celera Corp
(2) Univeristy of Utah
(3) University of Michigan
(4) Washington University
(5) University of Utah

Gene variants may modify an individual's risk of disease or treatment response. We present a Bayes method for calculating posterior probabilities of a binary trait for all multi-locus genotype combinations across susceptible loci and a novel statistic, the summed Kullback-Leibler divergence (SKL), which measures the departure of the posterior probabilities of a trait from the prior probabilities, weighted by the frequency of each multilocus genotype

combination. We show that the SKL is monotone increasing with the number of loci, and is well-described by a power function. The SKL is a concave function of disease risk, approximately peaking at the average frequency of the genotypes used. Applying the SKL to a 1,446 case/1,432 control study, we built a prognostic using *HLA-C* and the cytokine-related genes *IL13*, *IL12B* and *IL23R*. With a 3% prevalence, these results showed up to an 11-fold increase in psoriasis risk across individuals (SKL = 0.0029), a 30% increase over *HLA-C* alone. The signature was then applied to 2 additional psoriasis case/control sets, replicating the findings (SKL = 0.0035 (SKLa = 0.01 = 2.7E-04), SKL = 0.0025 (SKLa = 0.01 = 9.2E-05)). With 30% prevalence the SKL increases 5-fold. These results compare favorably to those from other studies: the Framingham risk score applied to a British heart disease study (SKL = 0.0049) and using 11 replicated type-2 diabetes SNPs (SKL = 0.0023). We conclude that genotypes at these 4 psoriasis genes substantially change psoriasis risk.

25

Using Pathway Information to Detect Higher Order Interactions in Complex Traits

Gary K. Chen (1), Duncan C. Thomas (1)
(1) University of Southern California, Department of Preventive Medicine, Division of Biostatistics

Genetic mapping studies of common diseases have revealed that the SNPs with the most significant p-values generally have modest marginal effect sizes (e.g. OR < 1.3). The small effect sizes may reflect the fact that these SNPs are not independent, in which case modeling them jointly using a model selection method (e.g. stepwise regression) would be appropriate. We prefer an alternative known as Bayes model averaging, which averages across all models to make inferences about the significance of a model or a model variable. Our MCMC algorithm computes a posterior probability at each candidate model by placing a prior on the association statistic of the observed data (e.g. β) as MVN(Zπ,T). The prior mean and variance-covariance structures are constructed from external biological knowledge (e.g. pathway information). Candidate models are proposed using a reversible jump Metropolis-Hastings proposal density. In the context of folate and methionine metabolism, we applied this method to a simulated dataset consisting of a case-control observational study with bio-markers measured on a subset of the subjects. Model averaged estimates of effect size and significance for main and higher-order effects showed good concordance to the best model from an AIC-based step wise regression. In a real-life example from a GWAS of breast cancer, we show how we can apply this method to infer the biologically plausible models by incorporating Gene Ontology knowledge into the prior.

26

Investigation of maternal effects, maternal-foetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring

Heather J Cordell (1), Holly F Ainsworth (1), Jennifer Unwin (1)
(1) Newcastle University

Many complex genetic influences, including epigenetic effects, may be expected to operate via mechanisms in the inter-uterine environment. A popular design for the investigation of such effects, including effects of maternal genotype and maternal-foetal interactions, is to collect DNA from affected offspring and their mothers. This design can also be used for the investigation of parent-of-origin (imprinting) effects, although greater efficiency may be achievable via the additional collection of DNA from fathers, where possible. Here we propose a novel multinomial modelling approach that allows the estimation of complex genetic effects of this type using data from either case/mother duos or case/parent trios. Through the incorporation of additional assumptions (such as Hardy-Weinberg equilibrium, random mating and known allele frequencies) and/or the incorporation of additional control samples (such as unrelated controls, controls and their mothers, or both parents of controls), we show that the parameters of interest are identifiable and well-estimated by our method. We investigate the required sample sizes and data structures necessary to provide accurate estimation of such effects and high power for their detection. Our method is illustrated by application to data on several candidate genes involved in congenital cardiovascular malformation.

27

Exploring Common Genetic Regulators of Gene Expressions Related to Insulin Resistance Syndrome: A Coupled Application of Factor Analysis and Linkage Analysis to GAW15 Problem 1 Data

Kyeezu Kim (1), Jin-Young Min (2), Kyunga Kim (3), Joohon Sung (1), Sung-Il Cho (1)
(1) School of Public Health, Seoul National University
(2) Institute of Health & Environment, Seoul National University
(3) Department of Statistics, Seoul National University

In this study, we aimed to explore the loci for a common regulator of the gene expressions related to insulin resistance. We selected 49 genes, and investigated their expression levels as expression quantitative traits (eQTs). Using factor analysis, we identified latent eQTs that characterize the clustering of 49 eQTs. VC linkage analysis was performed to explore the expression loci of each eQTs. Three resulting factors were regarded as insulin-related, lipid-related, and glucose-related latent eQTs, respectively. The highest linkage peaks were observed on the chromosome 5 and chromosome 2, for two eQTs of ADFP and IRS2. Insulin-related and lipid-related latent eQTs were located with comparatively higher LOD scores than most single original eQTs. Also, we found a considerable marker on chromosome 13, rs1319730, which were located on the peak locus of lipid-related latent eQT. The physical location of this SNP marker turned out practically agreed with the original location of IRS2. The results supported that characterizing and locating latent eQTs can provide an effective way to explore common regulators of genes related to insulin resistance. Rs1319730, has not been included in any specific gene, but we suggest the marker could be considered as a part of IRS2 gene or as a single functional marker. IRS2 gene could have a role of gene regulator among the lipid-related eQTs.

Factor analysis can be utilized for exploring common regulator loci.

28

Using Expression Data in Genetic Association Studies: an Integrated Bayesian Approach to Determine Genetic Pathways

Sharon M. Lutz (1), Christopher Paciorek (1), Christoph Lange (1)
(1) Harvard School of Public Health, Dept. of Biostatistics

Based on the hypothesis that the road from the disease locus to the phenotype of interest leads through expression data, the combination of genetic association data and expression data should provide a unique opportunity for pathway analysis and, ultimately, the understanding of the genetic causes of the disease. However, while expression data becomes more and more available in genetic association studies, its integration in the statistical analysis is not trivial. Given both the observed, genetic association between the genetic locus and the phenotype and the observed, genomic association between the expression profile and the phenotype, we want to conclude that the genetic association is completely explained by the genomic association, establishing a path from the genetic locus to the expression data. As we will show, this question can not be answered in the framework of standard statistical hypothesis testing. We develop here a Bayesian approach that is to able assess the genetic association in the presence of the genomic association. Using simulation studies, we verify that the proposed approach has the desired properties and sufficient statistical power. The approach is illustrated by an application to a real data set.

29

Transcriptomic Analysis of Quantitative Traits Related to Cardiovascular Disease

Eugene Drigalenko (1), Joanne E. Curran (1), Matthew P. Johnson (1), Jac Charlesworth (1), Thomas D. Dyer (1), Melanie Carless (1), Shelley A. Cole (1), Laura Almasy (1), Mahaney C. Mahaney (1), David L. Rainwater (1), Eric K. Moses (1), John Blangero (1), Harald H. Göring (1)
(1) Southwest Foundation for Biomedical Research

Microarray technology makes it now possible to quantify the expression levels of essentially all known genes. The expression profile can be used as a candidate gene discovery tool by identifying genes whose expression levels are correlated with traits assessed in the same individuals. We have applied this approach on 178 quantitative traits (QTs) of the San Antonio Family Heart Study. They include anthropometrics, obesity, diabetes, blood pressure, lipids, inflammation, oxidative stress, and thyroid function. Expression profiles were available for lymphocytes from 1,240 Mexican Americans. Using bivariate variance components analysis based on the pattern of covariation among relatives in our extended pedigree sample, we computed the genetic correlations between all 178 QTs and 22,413 autosomal transcripts found to be expressed, after adjusting for the affects of sex and age on both the QTs and the expression levels. We used the false discovery rate (FDR) approach to correct for multiple testing, calculating tail-based FDR (q-value) and local FDR for every QT–transcript pair. We found that for 80 QTs at least one transcript is significantly genetically correlated with $q \leq 0.05$. The largest number of significant transcripts was obtained with free thyroxine concentration in blood (FT4), where 7,273 transcripts were found to be significantly correlated with the trait. We use association analysis of the QTs and SNPs located within or near the correlated transcripts.

30

Multiple Component Linear Mixed Models to Correct for both Population Structure and Expression Heterogeneity

Jennifer Listgarten (1), Carl Kadie (1), David Heckerman (1)
(1) Microsoft Research

The presence of hidden structure is known to be a significant confounding factor in association studies, leading to spurious signal and loss of power when not properly accounted for. Because of the large scale of GWAS, these factors can often be learned from the data itself and then corrected for, using, for example, PCA based approaches or probabilistic, generative models such as mixed models.
Such methods have shown large success in a variety of experimental set-ups. On the one hand, some studies have used these approaches to account for hidden population structure such as race, or family and/or cryptic relatedness, where this structure is defined in the space of SNP data. On the other hand, some studies have used these approaches to account for ''expression heterogeneity'', that is, hidden structure defined in the space of microarray expression data.
In eQTL studies, where one is looking for association between SNPs and expression phenotypes, both types of hidden structure can be present and thus both need to be accounted for. We propose a novel linear mixed model approach that uses two variance components to account for these two confounding factors. We show that on data containing both of these kinds of structure that our model performs better in terms of calibration of the test statistic (using quantile-quantile plots) and in terms of power of detection (using ROC curves), than models which tackle only one of these two confounding factors.

31

Haplotyle-Based Tests for Imprinting Using Case-Parents Trios and Case-Parent Pairs

Wing K. Fung (1)
(1) Department of Statistics and Actuarial Science, The University of Hong Kong

Genomic imprinting is important in genetic trait study. Considerable research effort has been devoted to detection of imprinting effects. Recently there have been increasing interests in genetic studies involving several closely linked loci. This is mainly due to the fact that complex diseases are often associated with multiple markers and haplotype analysis is generally regarded as advantageous over single-marker analysis. The haplotype-based parental-asymmetry test (HAP-PAT) is constructed to test for imprinting using

multiple tightly linked markers based on case-parents trios. It is not uncommon in practice that one parent is missing due to some reasons such as late onset and case-parents trios are thus reduced to case-parent pairs, and discarding such kind of data certainly leads to a severe loss of information. Taking the information on case-parent pairs into account in genetic study is addressed in this study, the statistic HAP-1-PAT based on case-parent pairs is thus proposed to detect imprinting. Furthermore, the combined statistic HAP-C-PAT is developed to jointly use case-parents trios and case-parent pairs. Simulation studies are conducted to investigate the validity and the power of the proposed tests. We also find that the proposed statistics are robust to population structure.

32

A new Family-based Association Test for multiple tightly linked markers

Yilin Dai (1), Jianping Dong (1), Renfang Jiang (1)
(1) Michigan Technology University

Multi-marker tests usually work better to detect an underlying genetic variant over a genomic region than the single test, especially for the detection of complex diseases. In this report, the novel multi-marker family-based association test (FBAT) is based on the genotype scores after wavelet smoothing used to extract the common variation from the multiple tightly linked markers. When multiple markers are correlated because of linkage disequilibrium (LD), we might expect that identification of the common variation would capture more of the genetic signal encoded in this region. The new FBAT regards the multi-marker genotypes in the specified region as a signal. After wavelet transformation, the signal would be transformed into the time-frequency space and then be compressed into the modified genotype scores using an empirical Bayesian thresholding. Using wavelet smoothing can suppress noise automatically and enable a significant amplification of the genetic variation, while it keeps the genetic signal contained in the spatial ordering of SNPs. The new FBAT would be a potentially powerful method for multiple tightly linked markers and can also provide an alterative tool for the detection of underlying causative genetics variants. In the simulation study, we examine the type-I error and compare the power with other FBAT tests under different LD patterns. It has the correct type-I error rate and is more powerful than single-marker test with Bonferroni correction.

33

A Bayesian Approach to Genetic Association Studies With Family-based Designs

Melissa G. Naylor (1), Scott T. Weiss (2), Christoph Lange (1)
(1) Harvard School of Public Health
(2) Channing Laboratory, Brigham and Women's Hospital

For genomewide association studies with family-based designs, we propose a Bayesian approach. We show that standard TDT/FBAT statistics can naturally be implemented in a Bayesian framework. We construct a Bayes factor conditional on the offspring phenotype and parental genotype data and then use the data we conditioned on to inform the prior odds for each marker. For the construction of the prior odds, the evidence for association for each single marker is obtained at the population-level by estimating the genetic effect size in the conditional mean model. Since such genetic effect size estimates are statistically independent of the effect size estimation within the families, the actual data set can inform the construction of the prior odds without any statistical penalty. In contrast to Bayesian approaches that have recently been proposed for genomewide association studies, our approach does not require assumptions about the genetic effect size; this makes the proposed method entirely data-driven. The power of the approach was assessed through simulation. We then applied the approach to a genomewide association scan to assess association between single nucleotide polymorphisms and body mass index in the Childhood Asthma Management dataset.

34

Gene-environment Interaction Testing in Family-based Association Studies With Phenotypically Ascertained Samples: A Causal Inference Approach

David W. Fardo (1), Dawn L. DeMeo (2), Edwin K. Silverman (2), Stijn Vansteelandt (3)
(1) University of Kentucky College of Public Health
(2) Brigham and Women's Hospital and Harvard Medical School
(3) Ghent University

The class of family-based association tests (FBATs) provides strategies for testing main genetic effects that are robust to undetected/unaccounted for population substructure. Conditioning on parental/founder genotypes (or the corresponding sufficient statistics when parental genotypes are missing) insulates these testing strategies from the bias due to ancestry-driven confounding. However, once a main genetic effect must be estimated, as in the case of testing for $G \times E$ and gene-gene ($G \times G$) interactions, ascertainment conditions for sample recruitment must appropriately be taken into account.
The calculus of directed acyclic graphs (DAGs), specifically rules of d-separation, helps identify estimating equations that can properly incorporate ascertainment criteria. We employ the concept of principal stratification and G-estimation techniques to estimate main genetic effects consistently and are able, then, to test for gene-environment interactions. The resulting test maintains robustness to population stratification, avoids assumptions on the phenotypic and allele frequency distributions and accounts for sample ascertainment. We assess the performance of this test empirically through extensive simulation studies. To illustrate the approach in practice, we also apply these new techniques to a study of chronic obstructive pulmonary disease.

35

The Extended MFG Test: Improving a Test for Disease-Related Maternal-fetal Genotyping Incompatibilities to Allow for Arbitrary Family Structures

Erica Childs (1), Kenneth Lange (1), Christina GS Palmer (1), Janet S Sinsheimer (1)
(1) UCLA

Maternal-fetal genotype (MFG) incompatibility is an interaction between genes of a mother and child at a locus that adversely affects the fetus by inducing a maternal immunological attack, thereby increasing risk to disease. Statistical methods to examine MFG incompatibility as a risk factor for disease using a nuclear family based candidate gene approach have been developed1. Since families in a study can be large and complex we extend the MFG test to allow for arbitrary family structures. We modify the test by replacing the nuclear-family based mating type approach with Ott's pedigree likelihood, and change the Mendelian transmission probability to include a MFG incompatibility parameter. We implement our modification in Mendel 9.02 using an option that allows the user to fit their own likelihood. The modified test allows for inclusion of offspring-specific covariates such as offspring allelic effects and sex through the penetrance function, and to model multi-allelic loci. This implementation is more user friendly than earlier software versions. To extend the test it was necessary to make a slightly more stringent assumption of random mating than was necessary in the nuclear family MFG tests. Effects of violating random mating and improvements in statistical power arising from the ability to analyze more family genetic information will be discussed.

[1]Euro J Hum Genet 2004;12:192–198
[2]Am J Hum Genet 2001;69(supp.) A1886

36
WITHDRAWN

37
Is Correcting for Relatedness Necessary for Family-based Data? Evidence Against the Common Wisdom

Stacey Knight (1), Nicola J. Camp (1)
(1) University of Utah

We examined validity and power of family-based data in association studies. Different family selection criteria (random/ascertained), controls (familial/independent) and statistical methods were considered. Methods were: naïve (relatedness ignored); weighted (weights to account for relatedness); and variance inflation factor (VIF) adjustment. Simulated data were created for families and singletons under null and alternate models. For random family sampling, families were selected without regard to the number of cases in a family. For ascertained sampling, families were chosen based on a minimum number of cases. Controls were chosen to be in the families (familial) or independent individuals from the singleton simulations (independent). For each design 1000 cases and 1000 controls were generated and a trend test performed, amended by weighting or VIF. Results suggested that the naïve method was valid for randomly selected families. However, for ascertained families it appeared anti-conservative for independent controls, but conservative for familial controls. The weighting method was valid for independent controls (for either family selection), but was conservative otherwise. The VIF method was always valid. Independent controls were more powerful than familial controls, although the gain was small. In summary, contrary to common wisdom it appears that ignoring relationships in an association analysis may be valid, or perhaps even conservative, in some situations.

38
Incorporating Evidence for Population Stratification Bias in Combined Analyses of Case-Control and Case-Trio Data

Lucia Mirea (1), Lei Sun (1), James E. Stafford (1), Claire Infante-Rivard (2), Shelley B. Bull (3)
(1) Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
(2) Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine, McGill University, Montreal, Canada
(3) Samuel Lunenfeld Research Institute, University of Toronto, Toronto, Canada

Integrated analyses using case trios (CT) and independent case-control (CC) individuals can increase statistical power to identify susceptibility loci, however population stratification bias (PSB) among the latter remains a serious concern. Existing methods initially test for PSB by comparing results from CC and CT using an arbitrary testing level αPSB, typically 5%. Combined analyses are performed if no PSB is detected, otherwise analyses are restricted to CT. As a novel alternative, we propose to employ the PSB *P*-value (p*PSB*) and (1−*pPSB*) directly as weights to combine estimates from CC and CT, respectively. The weighted approach generalizes the method of Chen and Lin (*Genetic Epidemiol* 32(6): 520–7, 2008) by using a continuous weighting function that depends on p*PSB* instead of a binary one that depends on αPSB. To standardize the weighted estimate, we consider both a bootstrap variance estimate and a variance approximation assuming p*PSB* is a constant. Using simulations, we show that in comparison to Chen and Lin [2008], the weighted approach has reduced 5% type I error, increased (decreased) accuracy for larger (smaller) PSB levels, and overall increased positive predictive value. The resulting PSB adjustment is SNP-specific and can be applied to either candidate gene or genome-wide studies. We illustrate application of the weighted approach using a candidate gene study of childhood leukemia.

39
Family-based Association Tests: Accounting for Sib-sib Correlation, Linkage Effect, and Gene-environment Interaction.

Andrea Callegaro (1), Jenine Houwing-Duistermaat (1)
(1) Leiden University Medical Center

Family-based tests of association provide the opportunity to test for association avoiding false-positive results produced by population stratification. To test for family-based genetic association a class of score statistics has been proposed. These statistics are linear combinations of offspring genotypes and weights, where the weights are functions of the offspring traits (Genetic Epi, 19(Suppl 1), 2000, S36–S42).
In this work we propose three different extensions of the existing score statistics. First, we derive a new class of weights for general phenotypes incorporating the variance-covariance matrix of the random effects. Second, in order to test for association in the presence of linkage, we derive the variance of the score statistic taking into account the number of alleles shared identical-by-descent.

Finally, we adjust the family-based score statistics for gene-environmental interaction using a linear regression method which maintains power for small/weak interaction effects.
The performance of the proposed extensions is studied by simulations. As illustration, the proposed methods are applied to data on Rheumatoid Arthritis from the NARAC study (Genetic Epi, 31 (Suppl 1), 2007, S1-S148). Our results show that adjusting the score tests for additional information such as the sib-sib correlation, the linkage effect, and gene-environment interaction can considerably increase the power, as compared to the standard family-based score tests.

40

Detection of Foeto-maternal Genetic Effects in Early-onset Diseases: an Evaluation of the Methods

Mathieu Bourgey (1), Jasmine Healy (1), Marie-Hélène Roy-Gagnon (1), Daniel Sinnett (1)
(1) CHU Sainte-Justine research center / University of Montreal

Diseases with early-onset involve both the affected individual's inherited genotype and parentally-mediated mechanisms. In particular, a mother can influence the offspring's risk of disease not only as a genetic donor but also because she provides the fetal environment. However, differentiating between inherited and maternal genetic effects is not straightforward. We present a comparative analysis of three analytical approaches for the study of genetic associations in the context of early-onset diseases. We compared Weinberg et'al.'s hybrid design using case triads and parents of controls and a modified form of this log-linear, likelihood-based approach using case triads augmented by a set of unrelated cases and unrelated controls to Cordell et'al.'s conditional logistic regression approach which was used in combination with a classic case-control genotypic test for association. Using simulation experiments, we assessed type I error rates and power to detect genetic effects driven by the child, mother, or both under mating symmetry or asymmetry. We derived a new statistic to quantify the level of mating asymmetry. Finally, we applied the methods to a mixed data set consisting of childhood acute lymphoblastic leukemia patients, case-parent triads, and unrelated controls. This study will aid in maximizing analytical efficiency to account for the underlying genetic complexities of early-onset disorders and help refine our understanding of the etiology of pediatric disorders.

41

Genome-wide Association Analysis for Mixed Design Under Population Stratification in Genome-wide Association

Sungho Won (1), Nan Laird (1), Christoph Lange (1)
(1) Harvard School of Public Health

For family-based association analysis, we have two choice; population-based analysis incorporating correlation structure between family-member vs transmission-based test(TDT) such as FBAT. While population-based approach is sensitive to population stratification and phenotype distribution, TDT-type approach suffers from the insufficient power. Recently we proposed to combine FBAT with Wald test from between-family component and we showed that it can have virtually same power as population-based analysis in trio design.
In this report, we extend it to the general family design and propose a new method to adjust the population stratification based on principle component analysis for population-based approach in a general family design. To correct the correlations between family members we also provide the random effect model for family-based design in a population-based approach. Our results show that, first, the proposed method adjusts population stratification in a family design well. Second, the proposed FBAT-based approach can achieve virtually same power as population-based approach for both continuous and binary traits in a general family-design, and at the same time it preserves the same robustness as FBAT. As a result, we can use either way for family/mixed design under population stratification, but FBAT-based approach may be preferred because of its complete robustness to both population stratification and phenotypic distribution.

42

Fine-mapping of JAZF1 Region Associated With Prostate Cancer Susceptibility

Ludmila Prokunina-Olsson (1), Patricia Porter-Gill (1), McAnthony Tarway (1), Allison Burrell (1), Wei Tang (1), Gilles Thomas (2), Meredith Yeager (2), Demetrius Albanes (3)
(1) Laboratory of Translational Genomics, NCI/NIH
(2) CGF/DCEG/NCI/NIH
(3) NEB/DCEG/NCI/NIH

The SNP rs10486567 located in intron 2 of *JAZF1* has been identified as a novel risk factor for prostate cancer (PC) susceptibility (Thomas et'al., Nat Gen, 2008). We confirm now that this SNP is associated with prostate cancer susceptibility with the observed p-value of 1.83×10^{-11} in 9,331 cases and 10,558 controls of European origin. The rs10486567 is located 203 Kb from and has low correlation ($D' = 0.34$, $r^2 = 0.038$) with another *JazF1* SNP, rs864745, associated with susceptibility to type 2 diabetes (Zeggini et'al., Nat Gen, 2008). To explore the genetic landscape surrounding the PC-associated SNP, we first sequenced a region of 56 Kb in genomic DNA from 5 individuals (3 prostate cancer cases and 2 HapMap controls) and genotyped 11 SNPs (novel and variants from dbSNP with unknown frequencies) in three HapMap populations (European, Asian and African). Based on LD pattern observed in the HapMap samples of European descent (CEU), we selected 15 SNPs in $r^2 > 0.5$ with the original SNP rs10486567 and genotyped this set of markers in 500 controls and 500 prostate cancer cases from Finland (ATBC study). We observed that the pattern of LD in this region differed from the HapMap samples substantially. To further explore the LD structure and genetic variation in this region, we are conducting deep resequencing with the next-gen technology (454 Roche) over a 127 Kb region (chr7:27,899,757-28,027,351) in 96 samples representing several ethnic groups.

43

Inferences of Disease Polymorphisms from Case-control Genotype Association Data

Leeyoung Park (1)
(1) Yonsei University

After finding genes associated with diseases, the next step would be to obtain disease polymorphisms in the gene region. However, it is usual to observe several associated polymorphisms in a gene region from case-control association studies. In many cases, the most significantly associated polymorphism is considered as the disease polymorphism. However, it cannot be excluded that there are more than one disease polymorphisms in a gene region. To improve the ability to distinguish true disease polymorphisms from markers in linkage disequilibrium, a method is proposed, which can detect the disease polymorphisms regardless of their numbers in a gene region. Relying on the linkage disequilibrium between polymorphisms in controls, the proposed method utilizes model-based tests for finding disease polymorphisms. This method showed the acceptable type I and type II error rates, when sample sizes are large enough. However, depending on odds ratios and linkage disequilibrium, it is possible that there are situations demanding extremely large sample sizes for acceptable error rates.

44

A High-resolution Scan in the MHC Region for Psoriatic Arthritis Reveals Associations that are not Confounded by Previously Known HLA Risk Alleles

Nicole M. Roslin (1), Mathieu Lemire (2), Fawnda J. Pellett (3), Andrew D. Paterson (1), Joseph Beyene (1), Lynette Peddle (4), Angela Pope (4), Celia M.T. Greenwood (1), Proton Rahman (4), Dafna D. Gladman (3)
(1) The Hospital for Sick Children Research Institute
(2) Ontario Institute for Cancer Research
(3) Toronto Western Hospital, University Health Network
(4) Memorial University of Newfoundland

Psoriatic arthritis (PsA) is an inflammatory arthritis associated with psoriasis. Genetic factors are involved, with an estimated sibling recurrence risk ratio of approximately 30. Alleles at HLA-B, HLA-Cw and HLA-DRB1 have been shown to be associated with PsA. In order to determine if additional loci in the MHC region contribute to PsA, 909 individuals (422 cases, 487 controls) were genotyped for 2299 SNPs in a 5 Mb region on chromosome 6 encompassing the MHC region, using Illumina's MHC Panel Sets. Individuals were also genotyped at HLA-B, -Cw and -DRB1 to two-digit accuracy using sequence-specific oligonucleotide probes or sequence-specific primers. Since substantial linkage disequilibrium (LD) exists across the MHC region, association analysis of the SNPs was carried out conditioning on HLA risk alleles, using UNPHASED 3.1.3. rs1150735 was most strongly associated ($P = 8.98 \times 10-6$). Carriers of haplotypes that included any non-risk allele at all of HLA-B, -Cw and -DRB1 had a two-fold increase in risk when their haplotype contained the A allele at rs1150735 compared to the G allele (OR for the A allele: 2.19, 95% confidence interval: 1.43–3.36, $P = 0.00052$). rs1150735 is approximately 1.5 kb upstream of RNF39 (ring finger protein 39). Four additional SNPs were also associated with PsA conditional on HLA risk alleles ($P < 10-4$). By taking known risk alleles and LD into account, we were able to identify new associations with PsA in the MHC region.

45

A Bayesian Model for fine Mapping Following a Genome-wide Association Study

Jennifer A. Sinnott (1), Peter Kraft (1)
(1) Harvard School of Public Health

After a genome-wide association study (GWAS) has located a small region where a causal locus is likely to be found, subsequent "fine mapping" studies seek to identify SNPs that are most likely to be causal by genotyping an exhaustive set of markers in the associated region. Frequentist analyses of such studies rely on p-values to choose these SNPs, but it is not necessarily clear how to act on the basis of *P*-values—for example, it's not clear how to choose a p-value cut-off for declaring a SNP in this denser mapping worthy of follow-up. We propose instead a Bayesian approach, which gives each SNP in the region an equal prior probability of being the true causal locus, and uses the data to update this probability for each locus. This approach provides the posterior probability that each SNP is the causal SNP. We applied this model to a dense set of 640 SNPs in the FGFR2 gene in 1145 breast cancer cases and 1142 controls, following a significant, replicated association in this gene from a GWAS. This yielded results that were consistent with the frequentist approach, but more interpretable: it provided a credible set of 10 SNPs that contains the causal variant with 95% probability. We also applied this model to several other gene regions to evaluate its performance in different settings. This approach is easy to implement, and yields results that are more intuitive and actionable than a list of *P*-values.

46

Fine-Mapping in a Genomewide Linkage Scan of Prostate Cancer Susceptibility in Finland.

Claire L. Simpson (1), Cheryl D. Cropp (1), Tiina Wahlfors (2), Asha George (3), Ha Nati (2), Teuvo Tammela Tammela (2), Johanna Schleutker (2), Joan E Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health
(2) Institute of Medical Technology, University of Tampere and Tampere University Hospital
(3) National Human Genome Research Institute, National Institutes of Health/Fox Chase Cancer Center

Prostate cancer (PRCA) is the most common non-skin cancer in men and although mortality has been decreasing in developed countries due to early detection and better treatments, mortality is on the increase in some Asian countries. A number of risk factors have been identified for PRCA including age, ethnicity and family history of PRCA. Many genome-wide linkage studies, candidate gene association and genome-wide association studies have been performed to try to elucidate the genetic basis of susceptibility to PRCA. However, so far these studies have

produced mixed results because of disparate populations, genetic heterogeneity and high sporadic PRCA incidence. Previous genome-wide linkage studies using microsatellite loci in 54 Finnish PRCA families suggested 9 loci: 17q21–22, 10q22, 14q32, 4q22–23, 4q25, 3q25–26.3, 15q26, 13q34, 6q12–16, including genome-wide significant linkage at 17q21–22. For fine-mapping, 59 Finnish families with multiple men affected with PRCA were genotyped using the Illumina HumanLinkage-12 Marker Panel 6090 SNPs with average spacing every 0.58 cM. The data were checked for Mendelian inconsistencies and low call rate and the marker allele frequencies were estimated from the data. We then identified marker pairs that had intermarker LD $> r^2 = 0.1$ and dropped the less polymorphic markers of the pairs. Linkage analyses of PRCA to the remaining SNPs and to a combined microsatellite/SNP dataset are ongoing using GENEHUNTER-PLUS and will be presented.

47

Deep Sequencing of LIPC Reveals Multiple Variants Influencing HDL Size Distribution.

Phillip E. Melton (1), Karin Haack (1), Tom D. Dyer (1), Michael C. Mahaney (1), Jean W. MacCluer (1), David R. Rainwater (1), John Blangero (1), Laura Almasy (1), Shelley A. Cole (1)
(1) Southwest Foundation for Biomedical Research

The chromosome 15q21 region near the hepatic lipase structural gene (*LIPC*) has been implicated as a potential regulator of HDL-C concentration and particle size. Previous analysis has focused on a *LIPC* promoter variant (−514C→T). However, conditional linkage suggests that −514C→T does not explain the observed QTL.
We sequenced coding and conserved non-coding regions of LIPC in 182 founders and examined the 629 SNP genotypes identified in *LIPC* in 1,336 participants from the San Antonio Family Heart Study. Median diameters were measured for HDL particles stained for apoA1 (A1), apoA2, unesterified cholesterol (UC) and esterified cholesterol.
All HDL size phenotypes exhibited linkage to *LIPC* (LODs UC = 1.80 to A1 = 5.65). Measured genotype analyses on A1 indicate that 16 SNPs in the region demonstrate stronger association with particle size than LIPC-514. We also find evidence that −514C→T is in high LD with 4 other SNPs. The best SNP, rs11858164 ($P = 6.69 \times 10-5$), has also been identified in GWAS for triglyceride levels and HDL-C. We are currently conducting Bayesian Quantitative Trait Nucleotide analyses to identify potential functional variants from within the many correlated SNPs in *LIPC*. Preliminary results suggest multiple functional variants within *LIPC* are responsible for the observed 15q21 QTL. To our knowledge, this is the first comprehensive study of markers in *LIPC* and variation in lipoprotein size.

48

Exploring Rare Variants and Lung Cancer Risk from the Nicotinic Acetylcholine Receptor Region on Chromosome 15q25 in African-Americans

Chris I. Amos (1), Chongjuan Wei (1), Isaac Wun (1), Qiong Dong (1), Margaret R. Spitz (1), Marsha L. Frazier (1)
(1) U.T. M.D. Anderson Cancer Center

Last year, three independent groups including ours identified a region of chromosome 15q24-25.1 that associates with lung cancer risk for all histologies. In CEPH cell lines we noted SNPs associating with lung cancer risk also associate with expression of exons 3–5 of the nicotinic acetylcholinergic receptor, *CHRNA5*. To identify additional variation that may be associated with lung cancer risk we sequenced all of the exons for *CHRNA3*, *CHRNA5*, *PSMA4 and LOC123688*. Results of sequencing identified many new variants, and there were significantly more rare variants in cases than controls for *CHRNA5*. In addition, we identified insertion/deletion variations in the promoters of *CHRNA3* and *CHRNA5*. In a case control analysis of rare variants that were identified by resequencing none of the novel variants we identified were significantly more common in 467 AA cases than in 388 controls. The previously identified asparagine variant of rs16966968 was significantly more common in AA cases than controls ($P = 0.003$, OR = 1.98), but the variant with the highest odds ratio was an Alanine to Proline change in rs11551779 in PSMA4 yielding an OR of 5.0 ($P = 0.017$) for the major allele, with a rare minor allele. We saw very little effect of SNPs on smoking behavior in these cases and controls. The combined results suggest that both common and rare variants contribute to lung cancer risk for a region of chromosome 15q24–25.1.

49

A New Strategy for Linkage Analysis Under Epistasis Taking into Account Genetic Heterogeneity

Alexandre Bureau (1), Chantal Mérette (1), Jordie Croteau (1), Alain Fournier (1), Yvon C Chagnon (1), Marc-André Roy (1), Michel Maziade (1)
(1) Centre de recherche Université Laval - Robert-Giffard

Epistasis, the biological interaction of multiple genes modulating their individual effects, is likely omnipresent in complex diseases, and modelling epistasis in linkage studies can help detect loci with little marginal effect and detect epistatic effects themselves. We propose a complete three-step strategy for parametric linkage analysis under epistasis and heterogeneity in extended pedigrees. (1) Loci most likely involved in epistatic interactions are pre-screened using two-locus one-marker analyses. (2) Among selected loci, linkage to each locus is evaluated conditionally on linkage information at another locus under two-locus epistatic models. Linkage statistics are maximized over a space of epistatic models to avoid misspecification of model parameters. (3) Families linked to the conditioning locus are selected to deal with heterogeneity between pairs of epistatically interacting loci and other unlinked loci. Properties of conditional linkage statistics prevent the introduction of bias. Simulations reveal important gains in power to detect a locus with weak marginal effect involved in epistatic interaction. Application of our methods to schizophrenia and bipolar disorder in Eastern Quebec kindreds suggests epistasis between three locus pairs for bipolar disorder: 8p11–16p13, 15q11–16p13 and 18q12–15q11. These results suggest that the proposed strategy is powerful for tackling complex phenotypes involving epistasis and heterogeneity.

50

Power and Design Considerations for Detecting Gene-gene Interactions: Analytical and Simulation Comparison of GMDR and MDR

Guo-Bo Chen (1), Xiang-Yang Lou (2), Hai-Ming Xu (1), Yi Xu (1), Jun Zhu (1), Ming D. Li (2)
(1) Institute for Bioinformatics, Zhejiang Unversity, Hangzhou, Zhejiang, China
(2) Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, VA

This study intends to address population-based experimental design issues to detect gene-gene (G × G) interactions by using multifactor dimensionality reduction (MDR) and generalized multifactor dimensionality reduction (GMDR) methods. A literature review on detection of interactions suggested the testing accuracy ranged from 0.50 to 0.70 with a predicted heritability of 0.01∼0.05. The mathematical expectation of accuracy was derived and applied as a statistic indicator for both methods in power assessment within the practical range of accuracy. GMDR had a power of >80% with a sample size of ∼2000, when the accuracy ranges from 0.56 to 0.62, and when the accuracy is below 0.56, a sample size of 4,000 or more is required to yield sufficient power. GMDR outperformed MDR when the sample size is between 1,000 and 2,000 and the accuracy ranges from 0.56 to 0.62 for all simulated models, and they performed similarly when the accuracy is greater or the sample size increases such that the power becomes nearly 100%. Together, our simulation studies and the literature survey indicate that a sample size of 1000∼2000 should be sufficient to achieve most experimental goals in detection of moderate G × G interactions with heritability of 0.01 ∼0.05. (This study is being supported by NIH grants DA-12844 and DA-025095.)

51

Multifactor Dimensionality Reduction 2.1: Open-Source Genetic Analysis Software for Embracing the Complexity of Common Human Diseases

Peter C Andrews (1), Jason H Moore (1)
(1) Dartmouth College

Multifactor dimensionality reduction (MDR) was designed as a genetic model-free approach to identifying 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 2.1 of the open-source MDR software package that has been made freely available to the genetic epidemiology and bioinformatics communities since February of 2005. This new version of MDR has been significantly updated to allow users to load and analyze genome-wide associations study (GWAS) data. Improved data loading procedures and memory management techniques make it possible to carry out an MDR analysis on a GWAS data set. We expect this new version of MDR will open the door to routine epistasis analysis with GWAS data when combined with stochastic search methods such as the included estimation of distribution algorithm (EDA) that has the important 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.

52

A Learning Classifier System Approach to Detecting and Modeling Genetic Heterogeneity in the Presence of Epistasis

Ryan J. Urbanowicz (1), Jason H. Moore (1)
(1) Dartmouth College

Genetic heterogeneity (GH) greatly complicates the relationship between genotype and phenotype in genetic association studies. We propose here the development of Learning Classifier Systems (LCS) that combine machine learning with evolutionary computing and other heuristics to produce an adaptive system that simultaneously learns to solve different parts of a particular problem. The solution evolved by an LCS is represented as a population of independent rules or models which are utilized collectively to make decisions or predictions. We selected LCS for study here due to their inherent focus on multiple models that each explain different subsets of the data but collectively form a solution to the problem. Three LCS algorithms of differing architectures (XCS, MCS, and GALE) were implemented, evaluated and compared using simulated genetic heterogeneity data in the presence of epistasis. We specifically compared (1) the power to correctly detect predictive polymorphisms, (2) classification accuracy on testing data, (3) computational time, and (4) model generality. This analysis indicated than an LCS is able to achieve better than 80% power to detect all of the attributes involved in the underlying GH model. Additionally, XCS was demonstrated to significantly outperform the other tested LCS methods ($P < 0.001$). The results of this study demonstrate the potential of utilizing an LCS algorithm to address the problem of GH.

53

An Immersive 3-D Visualization Environment for Exploring High-dimensional Genetic Analysis Results

Doug Hill (1), Richard Cowper (1), Jason H. Moore (1)
(1) Dartmouth College

This is a challenging time due to the bioinformatics needs associated with storing, managing, analyzing and interpreting 'omics' data. While we have made progress with data analysis, the bioinformatics methods for knowledge discovery in the large volumes of statistical results are in their infancy. To address this challenge, we have developed a 3-D visualization approach to the exploration of statistical results from genome-wide association studies that capitalizes on the power of human visual perception and our evolved ability to recognize patterns. The goal of this study is to replace the traditional approach of sifting through thousands of rows of p-values in an Excel spreadsheet with a visual approach that presents the results in an interactive 3-D graphical format that makes

important patterns much easier to identify. To implement this we have harnessed the power of computer graphics technology in the form of video card hardware and a 3-D video game engine (Unity3D) that is designed for real time rendering of 3-D visual environments. Here, we map multiple different analytical results for each gene onto a 3-D structure such as a tree or a tropical fish. This allows the user to visually explore forested landscapes or schools of tropical fish for visually interesting patterns that reflect important multivariate information. This immersive 3-D virtual environment puts the user in a visually appealing environment that enhances scientific discovery.

54

Computationally Efficient Relief-based Algorithms for Detecting Epistasis in Genome-wide Association Studies

Casey S. Greene (1), Jeff Kiralis (1), Jason H. Moore (1)
(1) Dartmouth College

The detection of epistasis in genome-wide association study (GWAS) data is an enormous computational challenge due the combinatorial nature of the problem. We have previously shown that the ReliefF family of algorithms is able to detect epistasis in GWAS data without the need for a combinatorial algorithm. We previously developed an improved Relief algorithm called Spatially Uniform ReliefF (SURF) that significantly increases the power to detect interacting attributes in this domain. Here we provide a novel alternative to the nearest neighbor approach that instead uses all instances and weights genetic variants differently depending on whether each instance (i.e. subject) is close to or far away from the instance in question. We show using simulated genome-wide data that this new algorithm (SURF*) significantly outperforms ReliefF and SURF for genetic analysis in the presence of gene-gene interactions. For example, the power of ReliefF, SURF and SURF* to rank two interacting genetic variants in the top 95% of 1000 total variants is 5%, 10% and 80%, respectively, for an epistatic model with a heritability of 0.2 and a sample size of 1600 subjects. This study provides for the first time a computationally efficient algorithm that can detect epistatic loci with low to moderate genetic effect sizes in genome-wide data in a non-combinatorial manner.

55

Epistasis and the Genetic Architecture of Arterial Thrombosis in West Africans

Nadia Penrod (1), Folkert W. Asselbergs (2), Kwabena Poku (3), Jason H. Moore (1), Scott M. Williams (4)
(1) Dartmouth College
(2) University of Groningen
(3) University of Ghana
(4) Vanderbilt University

The enzymes, tissue plasminogen activator (t-PA) and plasminogen activator inhibitor 1 (PAI-1) act in concert to counterbalance thrombus formation and degradation in a homeostatic defense against the development of arterial thrombosis and excessive bleeding. Here we explore the involvement of epistatic interactions between polymorphisms in central genes of the renin-angiotensin (RAS) and fibrinolytic systems on plasma t-PA and PAI-1 levels within a sample of Blacks from Ghana ($n = 992$). Statistical modeling of epistatic interactions between polymorphisms in the ETNK2, REN, ACE, PAI-1, t-PA, and AGT genes was carried out using two-way ANOVA to assess their interaction effects on plasma levels of t-PA and PAI-1. We found the strongest interactions between polymorphisms of REN with the TPA I/D variant ($P = 0.001$) for females and the PAI 4G/5G variant ($P = 0.009$) for males, indicative of PAI-1 levels. Similarly, t-PA levels associated with ETNK2 variations in association with the PAI 4G/5G variant ($P = 0.029$) for females and the REN G/T polymorphism ($P = 0.024$) for males. These results represent only a few of the many significant interactions detected in this analysis. The outcome of this analysis confirms the notion that the underlying genetic architecture of cardiovascular disease is complex and, as such, it is necessary to consider the relationship between interacting polymorphisms of the pathway specific genes that predict t-PA and PAI-1 levels.

56

A Novel Approach for Detecting Gene-gene Interaction Using Multiple Traits

Xiaoqi Cui (1), Huann-Sheng Chen (1)
(1) Michigan Technological University

Recently, there is evidence suggesting that gene-gene interaction may play an important role in the complex disease. To detect gene-gene interactions, some methods have been proposed [Nelson et'al., 2001; Ritchie et'al., 2001; Lou et'al., 2007], but these methods only work on a single trait at a time. For a complex disease, the data is often collected with information on several traits. To take advantage of the information from multiple traits, we propose a Multivariate Combinatorial Searching Method (MCSM) that utilizes multiple phenotypes at a time. MCSM is conducted by incorporating the multiple phenotypess with various techniques of feature selection to search for a set of disease-susceptibility genes that may have interaction. We conducted a two-step screening procedure to select the significant marker sets that best explain the variation of multiple phenotypes. In simulation, under a variety of models, we demonstrate that incorporating multiple traits do gain higher power than using single trait only.

57

Genome-Wide Association Tests by Two-Stage Approaches with Seventeen Two-Locus Models

Zhaogong Zhang (1), Adan Niu (2), Shuanglin Zhang (1), Qiuying Sha (2)
(1) Department of Mathematical Sciences, Michigan Technological University;Heilongjiang University
(2) Department of Mathematical Sciences, Michigan Technological University

In this paper, a novel method, STORM, is proposed to search effectively and efficiently the two-locus effects which are jointly associated with the disease status in genome-wide association (GWA) studies. Since an exhaustive pairwise method is impossible for the two-locus model in GWA, we use the two-stage strategy. In the

second stage, we only retain hundreds of promising markers which have largest marginal effects at the first-stage then test on the retained eliminate the impacts of those high significant single markers for the reason that the high significant single locus being jointed with a noise signal locus can produce a very strong signal of the two-locus effects. It disturbs to test the significance of the two-locus effects. This algorithm adopts the two-locus models which have the biological meaning in the second-stage to get the significance of the statistics in two-stage analysis. Due to its effective and efficient, this method can be applied to genome-wide data. In order to evaluate STORM, we analysis its power and type I error by simulation data. Results demonstrate that our algorithm is more powerful than SMT and the two-locus genotypic test with 8 degrees of freedom for all two-marker models designed scenarios and its type I error have the nominal level.

58

Improved Ranking and Selection of Single Nucleotide Polymorphisms in Association Studies

Holger Schwender (1), Katja Ickstadt (2), Ingo Ruczinski (1)
(1) Departmen of Biostatistics, Johns Hopkins Bloomberg School of Public Health
(2) Department of Statistics, TU Dortmund University

A major goal in association studies is the detection of single nucleotide polymorphisms (SNPs) exhibiting an impact on the risk of developing a disease. The typical strategy for identifying such SNPs is to carry out SNP-specific (marginal) tests, and rank the SNPs by their univariate p-values.
Although this certainly is an acceptable strategy for a first pass through the data, it does not take the multivariate data structure of the SNPs into account, and therefore does not allow to identify SNPs that do not exhibit a main effect, but show an impact on the disease risk when interacting with other SNPs.
A solution to these problems is to employ regression or discrimination methods such as logic regression that explicitly search for interactions of variables, and thus enable the detection of disease-associated SNPs without a marginal effect. Based on the output of such procedures the importance of each SNP for a good prediction of the response of a study can be quantified, which in turn can be used to test the SNPs.
In this presentation, we will propose a testing method that makes essential use of this idea to enable a more appropriate ranking of SNPs than univariate testing, and thus to improve the detection of disease-associated SNPs. We will furthermore show how the proposed procedure can be extended to test biological sets of SNPs (e.g., SNPs belonging to the same LD-block or gene) and other types of responses (e.g., quantitative and multi-categorical responses).

59

Detection of SNP-SNP Interactions in Case-Parent Trios

Ingo Ruczinski (1), Qing Li (1), Thomas A. Louis (1), Daniele Fallin (1), Ann E. Pulver (1)
(1) Johns Hopkins University

Statistical approaches to detect higher order SNP-SNP (and possibly SNP-environment) interactions are critical in genetic association studies, as susceptibility to complex disease is likely related to the interaction of multiple SNPs (and environmental factors). We present a novel method to detect SNP-SNP interactions in trios with affected probands. The interactions are assessed in a regression framework, and become part of the model search space. The approach accounts for the linkage disequilibrium (LD) structure in the genotype data, and accommodates missing genotypes via haplotype-based imputation. We present a case study using case-parent trios from a study of schizophrenia and schizo-affective disorder, which revealed a genotype-phenotype association that includes an allele without marginal effect. We also introduce the open source software, which includes an efficient algorithm to simulate case-parent trios where genetic risk is determined by epistatic interactions.

60

Lasso Penalized Regression as a Screening Tool to Identify DNA Repair SNP-SNP Interactions in Familial Breast Cancer

Mary E. Sehl (1), Brittany N. Duncan (1), Patricia A. Ganz (1), Shehnaz K. Hussain (1), Zuo-Feng Zhang (1), Kenneth L. Lange (1), Janet S. Sinsheimer (1)
(1) University of California, Los Angeles

Background: SNPs involved in double strand break repair (DSBR) may modulate risk of familial breast cancer (BC) either singly or jointly with other SNPs or environmental factors. To explore this problem, we examined the association of BC with 173 DSBR SNPs in 399 Caucasian women in the UCLA family cancer registry. Analyzing interactions for 173 SNPs leads to a sparse data problem where the number of predictors ($>$10,000) far exceeds population size. Our approach involves using Lasso penalized regression (LPR) as a screening tool and stepwise logistic regression (SLR) for model refinement. Methods: LPR was performed using a log-additive model in the SNP Association option of Mendel 9.0, adjusting the penalty tuning constant to select the top 20 predictors and 40 interactions. Covariates included age, Ashkenazi Jewish heritage, and education. Selected predictors with both marginal p-values and leave-one-out indices <0.1 were introduced into SLR using R software. Results: Lasso predictors meeting the above criteria included age, 7 SNPs, one age$*$SNP and 5 SNP$*$SNP interactions. SLR supported an association between BC and age, and SNPs from XRCC3, TP53, RAD52, BRIP1, ZNF350, BRCA2, MRE11A, and RAD51. Conclusions: LPR is an efficient tool for identifying cumulative associations and interactions. Combining LPR screening and SLR to fully characterize models is an ideal method for identifying gene-gene and gene-environment interactions in groups with high BC risk.

61

Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction.

Manuel Garcia-Magarinos (1), Iñaki Lopez-de-Ullibarri (2), Ricardo Cao (2), Antonio Salas (1)

(1) Unidad de Genetica, Facultad de Medicina, Universidad de Santiago de Compostela
(2) Departamento de Matematicas, Universidade da Coruña

Most common diseases are likely to have complex etiologies. Methods of analysis that allow for the phenomenon of epistasis are of growing interest in the genetic dissection of complex diseases. By allowing for epistatic interactions between potential disease loci, genetic variants that might otherwise have remained undetected could be identified. Here we analyzed the ability of logistic regression (LR), classification trees (CART) and random forest (RF), to detect epistasis. Multifactor-dimensionality reduction (MDR) was used for comparison. Our approach involves simulation of case-control SNP datasets containing a two-loci interaction and 98 noise SNPs. We modelled interactions under different scenarios of sample size, missing data, minor allele frequencies (MAF) and several penetrance models, some of them involving pure interaction. CART, RF, and LR yield similar results in detection of association, but with better classification error results for CART and RF. In conclusion, tree-based methods and LR are important statistical tools for the detection of unknown interactions among SNPs with marginal effects and in the presence of a significant number of noise SNPs. In pure interaction models, RF performs reasonably well in the presence of large sample sizes and low percentages of missing data. However, when the study design is suboptimal, there is a high chance of detecting spurious associations. (Published in Ann Hum Genet 2009, DOI: 10.1111/j.1469–1809.2009.00511.x).

62
Model Selection Via Penalized Logistic Regression
Kristin L. Ayers (1), Heather J. Cordell (1)
(1) Institute of Human Genetics, Newcastle University

Penalized regression methods are an attractive alternative to single marker regression in association analysis. In underdetermined problems, regression methods are overwhelmed as the model becomes saturated and over fitting occurs. Penalized regression methods shrink the size of the coefficient of a marker with little effect on the trait of interest, ideally leading to a subset of markers most associated with the disease. Here we explore the advantages and pitfalls of penalization in selecting predictors in genetic association studies.
We investigated the performance of several software packages for performing penalized logistic regression, using computer simulations. We simulated data under a variety of disease locus effect size and linkage disequilibrium patterns. We compared several penalized methods, such as the elastic net, the lasso, the MCP, and the Bayesian inspired NEG shrinkage prior, to standard single locus analysis (the ATT), and simple forward stepwise regression. We investigated how markers enter the model as penalties and *P*-values are varied, and report the sensitivity and specificity for each of the methods. Preliminary results show similar performance in penalized methods and the ATT, with the main difference being penalized methods generally do not allow correlated variables to enter the model, leading to a sparser model in which most of the explanatory markers are accounted for. Forward stepwise regression performs poorly with a higher false positive rate.

63
Identification of Epistatic Effects Using Protein-protein Interaction Database
Yan V. Sun (1), Sharon L.R. Kardia (1)
(1) University of Michigan

Epistasis has long been recognized as an important mechanism underlying the complexity of the genetic architecture of common human traits. In many cases, the statistical detection of epistasis does not map onto or relate to the physical interactions between genetic variations that may impact human phenotypes through their combined influence on gene expression or through their interactions at the gene product (i.e. protein) level. We present a way to incorporate the physical protein-protein interaction (PPI) information into the statistical genetic analysis strategy. To illustrate this method we focused first on identifying epistatic intereations in gene expression that could be related to PPIs and copy number variations (CNVs). Among the 23,640 pairs of known human PPIs and the 1,141 common CNVs detected among HapMap samples using Affymetrix 6.0 chip, we identified 39 pairs of CNVs in both genes of a PPI pair. Two CNV pairs that provide sufficient genotype variation to illustrate our epistatic analysis due to the small HapMap sample size and the allele frequencies of these CNVs. Using 47,294 gene expression probes as the outcomes, 5 epistatic effects were identified with *P*-value less than 10^{-6}. Notably we found a CNV-CNV interaction significantly associated with gene expression of TP53TG3 with *p*-value of 2×10^{-20}. These significant associations suggest that the PPI data can assist in statistical analysis and detection of epistatic effects that reflect molecular mechanisms.

64
A Likelihood-based Approach for Detection of Gene-gene Interaction in a Case-control Study
Saonli Basu (1)
(1) University of Minnesota

Genome-wide association studies (GWAS) are a powerful approach for detection of genes associated with complex diseases. Although, new SNPs are found to be associated with these diseases through GWAS, still only very little of the genetic risk can be explained by individual SNP. A possible reason could be that each SNP alone may have little effect on the risk of the disease, but together (gene-gene interaction) may increase the risk substantially. In this paper, we propose a parsimonious model to assess the joint effects of a group of SNPs incorporating the possibility of interactions among them. The proposal aims to implement the data reduction strategy in Multifactor Dimensionality Reduction Method (MDR) within a likelihood framework and use a likelihood-ratio test to assess joint effects a group of SNPs. The advantage of our approach over MDR is that our approach does not use the adhoc criterion in MDR to decide if a genotype

combination is high-risk and low-risk. In the proposed approach, a likelihood-ratio test is used to assess the statistical significance, which is way less computationally intensive than the permutation test in MDR. Moreover, this proposed approach provides a way to capture the uncertainties regarding the choice of the model, which most of the current approaches thrives to capture. The proposed method was compared with two other approaches using simulation studies and appeared to outperform these two approaches in our simulation studies.

65

Modeling Multiplicative SNP Interactions in the Presence of an Additive Genetic Risk Score

Benjamin D. Horne (1), Nicola J. Camp (2), John F. Carlquist (1), Jeffrey L. Anderson (1)
(1) Intermountain Medical Center, University of Utah
(2) University of Utah, Intermountain Medical Center

A Genetic risk score (GRS) is a polygenic metric often computed as a sum of risk alleles (0–2) of each single nucleotide polymorphism (SNP) in a study. GRSs measure the joint effect of multiple SNPs at many loci, a shared aim of multiplicative SNP interactions. A GRS has the comparative benefit of avoiding small sample sizes, but the effect of a GRS on multiplicative interactions has not been evaluated. In a coronary angiography population ($N = 1{,}513$ [66% coronary disease]), 28 validated risk SNPs in 20 genes were studied for 2-way multiplicative SNP interactions by logistic regression. A total of 89 (23.5%) of 378 possible interactions had $P \leq 0.20$ (24 [6.3%] had $P \leq 0.05$) and addition of a 28-SNP GRS (risk allele range: 17–36) to each model produced a mean change in the interaction's β-coefficient of $0.67\% \pm 3.03\%$ (range: -8.4 to 8.8%). In an independent population ($N = 1{,}437$ [64% coronary disease]), validation analyses evaluated the 45 interactions with initial $P \leq 0.10$. Only 2 of the 45 replicated with effects in the same direction (*CETP* rs1800776 × *LPL* rs328, p-interaction = 0.08 and 0.028 in the initial and validation sets; *CRP* rs2794520 × *LIPG* rs2156552, $P = 0.025$ and 0.06). Multiplicative interactions were only mildly affected by a GRS, suggesting that multiplicative SNP models and additive GRS variables account for separate genetic effects across multiple loci. Furthermore, this indicates that both additive and multiplicative effects require continued study.

66

Reconstructability Analysis as a Tool for Identifying Gene-Gene Interactions in Studies of Human Diseases

Stephen Shervais (1), Patricia Kramer (2), Shawn Westaway (2), Nancy J Cox (3), Martin Zwick (4)
(1) Eastern Washington University, Cheney, WA
(2) Oregon Health and Sciences University, Portland, OR
(3) University of Chicago, Chicago, IL
(4) Portland State University, Portland, OR

There are a number of common human diseases for which the genetic component may include an epistatic interaction of multiple genes. Detecting these interactions with standard statistical tools is difficult because there may be an interaction effect, but minimal or no main effect. Reconstructability analysis (RA) uses Shannon's information theory to detect relationships between variables in categorical datasets. We applied RA to simulated data for five different models of gene-gene interaction, and find that even with heritability levels as low as 0.008, and with the inclusion of 50 non-associated genes in the dataset, we can identify the interacting gene pairs with an accuracy of $\geq 80\%$. We applied RA to a real dataset of type 2 non-insulin-dependent diabetes (NIDDM) cases and controls, and closely approximated the results of more conventional single SNP disease association studies. In addition, we replicated prior evidence for epistatic interactions between SNPs on chromosomes 2 and 15.

67

The Taiwan Schizophrenia Genetic Interaction Study

Hsin-Chou Yang (1), Jia-Wei Chen (1), Chih-Min Liu (2), Chun-Chiang Wen (2), Yu-Li Liu (3), Chun-Houh Chen (1), Hai-Gwo Hwu (2)
(1) Institute of Statistical Science, Academia Sinica
(2) Department of Psychiatry, National Taiwan University Hospital
(3) Division of Mental Health and Addiction Medicine, National Health Research Institute

This study aims to identify genetic interactions to schizophrenia using SNPs on 10 schizophrenia candidate genes in the Han Chinese population of Taiwan. We collected 912 cases and 600 controls. One hundred of SNPs on the 10 candidate genes (DISC1, LMBRD1, DPYSL2, TRIM35, PTK2B, NRG1, DAO, G72, RASD2 and CACNG2) were studied. Two neuropsychological indices, continuous performance tests (Zd' and Zmd'), were measured for schizophrenia patients when they were interviewed. Stratified analyses were carried out to reduce sample heterogeneity and genetic heterogeneity. We propose a multi-stage procedure for an identification of gene-gene/SNP-SNP interactions. Firstly, we identify within-gene interactions. Secondly, we identify between-gene interactions. Thirdly, a permutation test is performed to validate the within-gene and between-gene interactions. Fourthly, we calculate a 10-fold cross-validation consistency. Fifthly, logistic regressions with main effects and interactive effects are fitted for result confirmation and interpretation. The unstratified and CPT-stratified analyses construct the interaction networks, which clearly describe complex interplay of SNPs and genes to schizophrenia. The interaction networks show that DISC1, TRIM35, NRG1, and G72 are master gene nodes. Master SNP nodes are found within gene DISC1, DPYSL2, TRIM35, NRG1, and G72. Three interaction hotspots, TRIM35_SNP5∗NRG1_SNP7, TRIM35_SNP5∗G72_SNP2, and NRG1_SNP7∗G72_SNP2, are identified.

68

A Family-Based Association Test to Detect Gene-Gene Interactions in the Presence of Linkage

Lizzy De Lobel (1), Hans De Meyer (1), Lutgarde Thijs (2), Tatiana Kouznetsova (2), Jan Staessen (2), Kristel Van Steen (3)
(1) University of Ghent
(2) KULeuven
(3) University of Liège

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 models simultaneously linkage and association. It splits up the association effect in a between- and a within-family component to adjust and test for population stratification and includes a variance components method to model linkage.
We extend this approach to detect gene-gene interactions between two unlinked QTLs by adjusting the definition of the between- and within-family component and the variance components included in the model. To capture the influence of population stratification, we derive the bias of the estimated interaction effect and discuss the influence on type I error rates. We simulate data to investigate the influence of the epistatis model, LD patterns between the markers and the QTLs, family structures and allele frequencies on the power and type I error rates of the approach. Results show that for some of the investigated settings, power gains are obtained in comparison with other techniques (e.g. FBAT-LC). We conclude that our approach shows promising results for studies where too few markers are available to correct for population stratification using standard methods (e.g. EIGENSTRAT). The proposed method is applied to real-life data on hypertension.

69
Allelic Based Gene-Gene Interaction in Longitudinal Data
Jeesun Jung (1)
(1) Indiana University School of Medicine

In longitudinal data where repeated measures are present, there is lack of statistical methods to identify genetic association through interaction caused by multiple single nucleotide polymorphisms (SNPs) within a gene as well as by SNPs at unlinked genes contributing to risk of a disease. A novel statistical approach is proposed to detect the gene-gene interactions at the allelic level under semiparametric framework. With a new allelic score inferred from the observed genotypes at two or more unlinked SNPs, we derive profiled score statistics which are unbiased and asymptotically efficient using profile likelihood function. By testing for the association, the interaction can be assessed both in cases where the SNP association can be detected and cannot be detected as a main effect in single SNP approach. The analytical power and type I error rates over 6, two-way interaction models are investigated based on simulation study. Simulation studies demonstrate that (1) the profiled score statistic follows chi-square distribution on three degrees of freedom in two unlinked genes and (2) the allelic based method provides higher power than a genotypic based methods.

70
Evaluating Gene-gene and Gene-environment Interactions in AMD
Päivi Onkamo (1), Sanna Seitsonen (2), Ilkka Immonen (2), Irma Järvelä (3)
(1) Department of Biological and Environmental Sciences, University of Helsinki
(2) Department of Ophthalmology, University of Helsinki
(3) Department of Medical Genetics, Haartman Institute, University of Helsinki

Age-related macular degeneration (AMD) is an eye disease of the elderly, signs of which appear after the age of 50. Currently, the genetics of the syndrome is among the best known of any multifactorial diseases: A few major genes from complement cascade, and an environmental factor, smoking, assert over 60% of population attributable risk. Shared metabolic pathways could plausibly result in gene-gene interactions. Indeed, several studies have identified potential interactions between the major genes in AMD, but unfortunately, replications of these interactions have been almost non-existent. In this study, CFH, HTRA1, LOC387715, C3, and properdin genotypes as well as sex and tobacco smoking were assessed in 331 cases and 108 non-AMD controls. We evaluated gene-gene and gene-environment interactions between these factors and sex, with several different methods: logistic regression, synergy index (departure-from-additivity model), mutual interaction (MI) and Multifactor Dimensionality Reduction (MDR). Evidence in favour of a gene-gene interaction between the two major AMD-associated loci, LOC387715 and CFH, was found with three first mentioned methods. Smoking (ever vs. never) exerted an extra risk for AMD, but in our data, only in connection with other factors such as sex and C3 genotype, but this was not confirmed by all methods. No interaction between CFH Y402H and smoking or LOC387715 A69S and smoking could be detected, except for MDR.

71
A Pattern-counting Based Sequential Permutation Method for Detecting Epistasis in Disease Association Studies
Li Ma (1), Themistocles L. Assimes (1), Narges B. Asadi (1), Carlos Iribarren (2), Mark Hlatky (1), Thomas Quertermous (1), Wing H. Wong (1)
(1) Stanford University
(2) Kaiser Permanente of Northern California

Epistasis has long been speculated to play important roles in the epidemiology of common diseases. However, current approaches to testing genetic interactions in large association studies have achieved limited success, and are most ineffective in the presence of genetic heterogeneity—when people have (clinically) the same disease for different genetic reasons--which is likely the norm for complex diseases. In this work, we introduce a method for testing interactions that is robust to genetic heterogeneity, and yet achieves computational efficiency and statistical rigor. The method consists of two modules--one is a pattern counting algorithm designed for efficiently evaluating the risk significance of marker combinations, and the other is a sequential permutation scheme for multiple testing correction. We demonstrate the work of our method in analyzing a dataset for cardiovascular and coronary diseases with a spiked-in signal. The data involve 1,634 samples (977 cases and 657 controls) and 580 SNPs covering about 100 candidate genes. The injected signal, a three-gene interaction, is present in about 5% of the observations (7.9% of the cases and 1.1% of the controls),

imitating the effect of genetic heterogeneity. While both the logistic regression approach and the multifactor dimensionality reduction (MDR) method fail to detect the injected signal, our method picks it up with no ambiguity (a permutation adjusted p-value of 0), along with a number of other signals.

72

Causal vs. Mathematical Models of Two-locus Penetrance

Ann M Madsen (1), Susan E Hodge (2), Ruth Ottman (3)
(1) Department of Epidemiology, Columbia University
(2) New York State Psychiatric Institute, Department of Biostatistics, Columbia University
(3) Department of Neurology, G.H. Sergievsky Center, Department of Epidemiology, Columbia University

Background: Two-locus penetrance models based on mathematical models, e.g. multiplicative, have been described, but the two-locus penetrances predicted by various causal models have not; so the validity of existing two-locus penetrance models for etiologic questions is unproven. Here, we derive two-locus penetrances from a sufficient-component cause model (SCCM) and compare these with previous derivations.
Methods: We define a SCCM for a complex disease with two unlinked genetic causes, incorporating reduced penetrance, phenocopies, genetic heterogeneity and gene-gene interaction; determine the corresponding two-locus penetrances, and compare these two-locus penetrances with previously used mathematical model-based penetrances.
Results: Assuming two binary genetic variants, a binary outcome, independent distribution of genetic and non-genetic causes, and our definitions of causal gene-gene interaction and causal genetic heterogeneity, the mathematical "heterogeneity" penetrance model is the appropriate basis for modeling causal genetic heterogeneity and gene-gene interaction. Additive two-locus penetrance models do not generally describe a disease characterized by genetic heterogeneity, nor does a multiplicative two-locus penetrance model generally describe a disease characterized by gene-gene interaction.
Conclusion: Some mathematical models do have biologic interpretations, but only under certain assumptions and explicit definitions.

73

Modifications to the ReliefF GWAS Filter for Detecting Gene-Gene Interactions: Inclusion of Covariate and Pathway Information

Nicholas M. Pajewski (1), Vinodh Srinivasasainagendra (2), Brett A McKinney (3)
(1) University of Alabama at Birmingham
(2) University of Alabama at Birmingham
(3) University of Tulsa

Recent research has turned to the role that interactive effects may play in the development of complex traits. Machine learning approaches, such as ReliefF, have been recently popularized as powerful methods for detecting these effects. However, an issue that has not been fully explored is the desire to incorporate additional data beyond genotype and phenotype, such as covariate or pathway information. This could occur if one needed to account for population stratification due to a racially diverse sample, or if one wanted to highlight interactive effects operating within specific pathways. Using simulated data based on Phase 3 of the HapMap Project, we first illustrate the deleterious impact of failing to account for covariate information upon applying the ReliefF filter. We then introduce a modified algorithm, termed ReliefFStrat, which incorporates stratification/covariate information in constructing the ReliefF score for each locus. We also illustrate the use of ReliefF scores as a post-processing metric in order to illuminate interactive effects occuring within specific pathways.

74

A General Statistical Framework for Genome-wide Association Studies (GWAS) Based on Bayesian Graphical Modeling

Laurent Briollais (1), Jinnan Liu (1), Adrian Dobra (2), Helene Massam (3)
(1) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Canada
(2) Dept. of Stat., University of Washington
(3) Dept. of Stat. and Math., York University, Canada

The actual paradigm to analyze GWAS is to perform an exhausting testing of all single SNP associations with the response variable with the major drawback that the selected subset of SNPs has in general a very low predictive value. As a shift to the usual approach, we propose here a general statistical framework for GWAS based on Bayesian graphical modeling (Massam et'al., 2008, Annals of Stat.) and able to: (1) Assess the joint effect of multiple SNPs (linked or unlinked); (2) Explore the model space efficiently using the stochastic search algorithm MOSS (Dobra and Massam, 2008, Stat. Method.); (3) Incorporate expert prior knowledge in the model search. We illustrate the methodology through a GWAS of 42 NCI cell lines classified as resistant ($n = 27$) or sensitive ($n = 15$) after exposure to estrogen using 25K SNPs. Our algorithm selected 17 SNPs embedded in multilocus models with high posterior probabilities. Most of the selected SNPs have a biological interest and their predictive value is perfect in our sample. Interestingly, many of them would not have been detected in the single-SNP testing approach. Finally, we discuss the impact of informative priors in this statistical framework.

75

Microsimulation of Populations With Realistic Genetic and Environmental Risk Factors of Lung Cancer

Bo Peng (1), Christopher I. Amos (1)
(1) the University of Texas, M. D. Anderson Cencer Center

Effective use of simulated datasets in genetic epidemiology studies requires close resemblance between simulated and real datasets. Although a number of methods have been proposed to simulate datasets that match real datasets in terms of marker density, allele frequency and linkage disequilibrium patterns, none of them incorporates realistic patterns of environmental factors and their

interactions with genetic risk factors. In order to determine effective statistical methods in the detection of genetic susceptibilities of lung cancer in the presence of strong environmental factors such as smoking, we developed a forward-time population genetics simulation algorithm to simulate large populations that resemble real human populations in terms of both individual-level genotype and smoking behaviors and population-level properties such as patterns of smoking and lung cancer incidence. We drew different samples from these populations and studied the relative efficiency of these ascertainment methods. For example, compared to a case-control design with both smokers and non-smokers, a smoker-only case-control design eliminates the impact of a nicotine-addition gene, but it does not improve the power of detecting other disease predisposing loci. We demonstrated the potential applications of this simulation technique in public health genomics with an preliminary analysis of the cost-effectiveness of genetic-guided screening of lung cancer.

76

Comparison of Three Machine Learning Approaches to Examine the Genetic and Environmental Predictors of Vitamin D Levels

Corinne D. Engelman (1), Justin Lo (1), Martha O'Brien (1), Carl D. Langefeld (2), Tasha E. Fingerlin (3), Jill M. Norris (3)
(1) University of Wisconsin
(2) Wake Forest University
(3) University of Colorado

Vitamin D status, determined by blood levels of 25-hydroxyvitamin D (25[OH]D), is associated with many adverse health outcomes. The role of interactions between genes and with environmental factors has not been examined. Machine learning approaches can screen large amounts of data and take interaction effects as well as main effects into account without requiring model specification. We measured 25[OH]D in 504 Hispanics from San Antonio, TX; 513 Hispanics from the San Luis Valley, CO; and 513 African Americans from Los Angeles, CA, recruited by the IRAS Family Study. Gender, age, ethnicity, solar radiation, BMI, physical activity, and 30 SNPs in 3 vitamin D-related genes (*VDR*, *CYP27B1*, and *DBP*) were included in 3 machine learning approaches: multifactor dimensionality reduction (MDR); generalized, unbiased, interaction detection and estimation (GUIDE); and random forests (RF). Using MDR with 10-fold cross-validation, the best model included ethnicity and solar radiation. Using GUIDE with 10-fold cross-validation, the first split variable was age, followed by solar radiation, ethnicity, *DBP* SNP rs7041, gender, *CYP27B1* SNP rs703842, BMI, and *VDR* SNP rs2239185. In 5 runs of RF, solar radiation and ethnicity were always the first and second most important variables, respectively, with *VDR* SNP rs7302235, *DBP* SNP rs4588, and *VDR* SNP rs10783219 ranking 3-5. Machine learning offers an alternative to selecting particular interactions to evaluate.

77

Studying Case-parent Triads to Identify Haplotype-by-Exposure Interactions

David M. Umbach (1), Min Shi (1), Clarice R. Weinberg (1)
(1) NIEHS

Joint analysis of multiple SNP markers can be informative, but studying joint effects of haplotypes and exposures is particularly challenging. Population structure can involve both genes and exposures: If exposure, disease, and haplotype frequencies covary across subpopulations, a case-control study may wrongly suggest haplotype-by-exposure interaction. Suppose a particular haplotype has no influence on propensity to exposure. Then, under a no-interaction null hypothesis, its transmission from parents to affected offspring can violate Mendelian expectations but should be independent of exposure. This insight underpins our test of no interaction. Defining strata by sums of parental genotypes, we base our test statistic on approximating the mean of stratum-specific covariances between exposure and a score reflecting transmission. We generate its null distribution by permuting exposures within each stratum. Its desirable features include: investigators need not know or estimate haplotypes, their frequencies or phases; genetic main effects are unrestricted; Hardy-Weinberg disequilibrium and exposure-related population stratification are tolerated, as are missing SNPs; and many markers can be studied. Simulations suggest our proposed test does not exceed its nominal Type I error rate and provides good power under a variety of scenarios. We illustrate by examining whether SNP variants in GSTP1 modify the association between maternal smoking and oral clefting, a birth defect.

78

A Comparison of Sample Size and Power in Case-only Association Studies of Gene-Environment Interaction

Geraldine M. Clarke (1), Andrew P. Morris (1)
(1) University of Oxford

Assuming continuous environmental and categorical genotype variables, the authors compare six case-only tests of association for gene-environment interaction. Novel tests modelling the environmental exposure as response, a non dominant coding of the genetic exposure as response, as well as the traditional test with its dominant coding of the genetic exposure as response, are included. The authors show that tests imposing the same genotypic pattern of inheritance perform similarly regardless of whether genotype is modelled as response or predictor. These novel tests with genetic exposure as response are advantageous as they are robust to non-normally distributed environmental exposures. Dominance deviance, deviation from additivity in the main or interaction effects, is key to test performance: when zero or modest, tests that search for a trend with increasing risk alleles, are optimal; when large, tests for genotypic effects are optimal. However, the authors show that when testing at a proxy locus, common in genome-wide association studies, a large dominance deviance is rare and so genotypic, as opposed to trend, tests are not generally recommended. Traditional tests assuming a dominant pattern of inheritance can suffer from serious losses of power in the presence of any recessive, or modest dominant, effects.

79

Efficient Testing of Gene-Environment Interaction in Genome-wide Association Studies

Cassandra E. Murcray (1), Juan Pablo Lewinger (1), W. James Gauderman (1)
(1) University of Southern California

High-volume genomic data available in genome-wide association studies (GWAS) offers a unique opportunity to discover new gene-environment (G × E) interactions related to complex-disease etiology. However, because of the large sample size requirements to detect interactions using standard tests and the large penalty for multiple testing, new powerful statistical methods are needed to detect G × E interactions on the genome-wide scale. In the context of a case-control GWAS, we describe an efficient two-step analysis method to identify SNPs that are likely to be involved in G × E interactions. The method uses a case-only style test in Step 1 to screen all SNPs for potential G × E interaction. This screening step reduces the number of markers that are formally tested for interaction in Step 2, reducing the penalty of multiple testing. We demonstrate that our 2-step test is substantially more powerful than a traditional test for interaction under a wide range of scenarios. For example, for a GWAS of 500,000 markers in 1,000 cases and 1,000 controls, our 2-step method achieves 80% power to detect an interaction OR of 2.5 compared to only 33% power for a traditional test for interaction. We describe a tool to calculate optimal power and sample size for the 2-step procedure that can be used in the design of future studies. We also apply the 2-step procedure to a GWAS of asthma conducted in the Children's Health Study to scan for GxE interactions with in-utero tobacco smoke exposure.

80

Can Gene-by-environment Interaction be Inferred From Parent-case Transmission Rates?

Ji-Hyung Shin (1), Brad McNeney (1), Jinko Graham (1)
(1) Simon Fraser University

Most complex diseases result from an interplay between genes (*G*) and environmental attributes (*E*). Statistical interaction between *G* and *E* occurs when genotype relative risks of disease vary with *E*. Since GxE interaction leads to variation with *E* in transmission rates of the risk allele, transmission rates have been used to make inference about GxE interaction. We investigate the validity of this practice by deriving theoretical transmission rates under standard population genetic assumptions and show that variation in transmission rates with *E* does not, in general, reflect the truth about GxE interaction.

81

Meta-Analysis of Gene × Environment Interaction: Joint Synthesis of SNP and SNP × E Regression Coefficients.

Alisa K. Manning (1), Elliot Stolerman (2), Jose C. Florez (2), Michael Lavalley (1), James B. Meigs (3), L. Adrienne Cupples (1), Josee Dupuis (1)
(1) School of Public Health, Boston University
(2) Massachusetts General Hospital; Broad Institute; Harvard Medical School
(3) Massachusetts General Hospital; Harvard Medical School

Introduction: The validation of associations in genome-wide association scans (GWAS) has evolved from single study replication to consortia using meta-analysis of large numbers of studies. Because a risk factor may interact with genetic components, investigation of gene × environment effects in these large consortia may yield additional susceptibility loci. We introduce a method of joint meta-analysis of SNP and SNP × Environment (SNPxE) regression coefficients.
Methods: In GWAS of gene × environment interaction, a two degree of freedom test has been proposed to identify genetic variants with an influence on the trait of interest. This approach has the advantage of being able to detect marginal and interaction effects when the association between the trait and the SNP is not homogeneous across different levels of the environmental factor. We propose a method to jointly meta-analyze the SNP and SNPxE coefficients using multivariate generalized least squares. Our approach provides confidence intervals, a joint significance test for SNP and SNPxE terms, and a test of homogeneity across samples.
Results: We present simulation studies exploring the utility of this method to continuous and binary outcomes, and an application of this method to a meta-analysis of the association of the K121Q allele of the ENPP1 gene with type 2 diabetes∗, accounting for a possible interaction with BMI.
∗ the ENPP1 Consortium

82

Exploiting the Gene-environment Independence Assumption to Assess Genetic Modification of Menopausal Hormone Therapy Associated Postmenopausal Breast Cancer Risk

Rebecca Hein (1), Sascha Abbas (1), Dieter Flesch-Janys (2), Jenny Chang-Claude (1)
(1) German Cancer Research Center (DKFZ)
(2) University Medical Center Hamburg-Eppendorf

Few studies have investigated potential effect modification of menopausal hormone therapy associated breast cancer (BC) risk by single nucleotide polymorphisms (SNP) in the progesterone pathway.
We used a population-based case-control study from Germany with 2,502 postmenopausal BC patients and 4,833 controls to asses statistical interactions between five functional SNPs in the progesterone metabolizing enzymes, *AKR1C3*, *AKR1C4* and *SRD5A1*, and estrogen-progestagen combination therapy as well as estrogen monotherapy with regard to postmenopausal BC risk. Multivariate logistic regression (LR) and an empirical-Bayes (EB) procedure that tests for interaction using a weighted combination of the case-control and case-only estimator were applied (Mukherjee and Chatterjee, *Biometrics*, 2008, 64(3):685–694).
BC risk associated with duration of combination therapy was found to be significantly modified by *SRD5A1*_rs3736316 ($P = 0.02$ using LR, $P = 0.005$ using the EB method). The risk associated with duration of use of estrogen monotherapy was also significantly modified by *AKR1C3*_rs7741 ($P = 0.03$ using LR, $P = 0.06$ using the EB method) and two SNPs in *AKR1C4* (rs3829125: $P = 0.01$ using LR, $P = 0.04$ using the EB method; rs17134592: $P = 0.02$ using LR, $P = 0.06$ using the EB method). After Bonferroni

correction for multiple testing only *SRD5A1*_rs3736316 assessed using EB was still significant ($P = 0.03$).
The EB method exploits the G-E independence assumption and thus can have increased power to detect interactions.

83

NAT2 Haplotypes Modify the Effects of Smoking, Alcohol, and Caffeine on Fertility

Kira C. Taylor (1), Lauren E. Murray (1), Chanley M Small (1), Malania Wilson (2), Weining Tang (2), Mark Bouzyk (2), Michele Marcus (1)
(1) Emory University Department of Epidemiology
(2) Emory University Biomarker Service Center

Background. The enzyme N-acetyltransferase 2 (*NAT2*) is responsible for metabolizing and detoxifying xenobiotics such as caffeine and tobacco smoke. Common polymorphisms in the *NAT2* gene determine haplotypes that have slow or fast acetylator phenotypes. The slow haplotypes have been associated with increased risk of bladder cancer and other conditions. We investigated whether *NAT2* haplotypes affected time to pregnancy and/or modified the effects of xenobiotic exposures on time to pregnancy. Methods. We conducted a prospective cohort study investigating time to pregnancy in a population of 470 women office workers who were at risk for pregnancy. Urine samples served as the source of DNA. Three *NAT2* polymorphisms were genotyped and discrete survival analysis was used to determine whether *NAT2* haplotypes modified any effects of alcohol, smoking, or caffeine on time to pregnancy. Results. Increasing levels of alcohol, smoking, and caffeine were all associated with increased time to pregnancy in a dose-response manner. There was no main effect of *NAT2* haplotype or genotypes on time to pregnancy. Interaction was observed between the *NAT2* haplotype and smoking, alcohol, and caffeine; slow acetylators were more susceptible to the effects of all three exposures. Conclusion. When studying the effects of xenobiotics on human health, it may be of scientific importance to incorporate genetic information about relevant metabolic enzymes that may alter susceptibility.

84

Measuring Genetic Association by Familial Relative Risks Attributable to Imputed Genotypes in Order to Boost the Identification of Causal Variants

Justo Lorenzo Bermejo (1), Abigail G. Matthews (2)
(1) Institute of Medical Biometry and Informatics, University Hospital Heidelberg
(2) Ott Laboratory, Rockefeller University

Current genotyping platforms are designed to investigate common variants (polymorphisms). Most associations from genome-wide scans have a small impact on the observed familial relative risks of disease, commonly represented by λ. Previously, we have shown that when identified polymorphisms are markers of rarer causal variants, the rare variants explain a larger proportion of λ than the identified polymorphisms [1]. Here, we investigate the possible benefit of the representation of genetic association by the λ attributable to imputed genotypes in order to facilitate the detection of causal variants.

We first simulated case-control data based on HapMap. We assumed that a particular SNP in the set of haplotypes from HapMap was a causal variant, predefined relative risks for homozygous and heterozygous carriers of the susceptibility allele and inferred data at flanking SNPs conditional upon HapMap haplotypes. We then removed the causal SNP from the simulated dataset, imputed genotypes for the causal SNP based on HapMap and summarized the association between imputed genotypes and disease by the attributable λ. The power of this statistic was compared with the power of the Bayes Factor under different parameterizations (frequency, penetrance and inheritance mode of the causal SNP, and its linkage with flanking SNPs).
[1] Hemminki K, Forsti A, Lorenzo Bermejo J: The 'common disease-common variant' hypothesis and familial risks. *PLoS ONE* 2008, 3:e2504.

85

The Impact of Pedigree Structure on Heritability Estimates

Claus T Ekstrøm (1)
(1) University of Copenhagen, Faculty of Life Sciences.

Heritability measures the familial aggregation of a disease or trait and a non-zero heritability suggests that a genetic component may be present. Reliable heritability estimates are necessary in the planning phase of a linkage or genetic association study but often these estimates are obtained from other studies where the composition of pedigrees may be different from the study that is prepared.
The impact of pedigree structure on precision and accuracy of heritability estimates is examined for data and models both with and without dominance effects. Analytical and simulation results find that for purely additive genetic effects all but the simplest pedigree structures provide the same information about the heritability of a quantitative trait. In the presence of dominance effects there is a substantial difference in the precision obtained by different pedigree structures.

86

SNP-SNP Interactions Dominate the Genetic Architecture of Candidate Genes Associated with Left Ventricular Mass in African-Americans of the GENOA Study

Kristin J. Meyers (1), Jian Chu (1), Thomas Mosley (2), Sharon L.R. Kardia (1)
(1) University of Michigan Department of Epidemiology
(2) University of Mississippi Medical Center

Left ventricular mass (LVM) is a strong, independent predictor of heart disease incidence and mortality. This research attempts to characterize the genetic architecture of LVM in an African-American population by examining the main and interactive effects of candidate gene SNPs and conventional risk factors on LVM. We used least-squares linear regression to investigate 1,878 SNPs from 268 candidate genes for associations with LVM in 1,368 African-Americans from the Genetic Epidemiology Network of Arteriopathy (GENOA) study after adjustment for admixture. We reduced the probability of false positive

results by implementing three analytic strategies: false discovery rate (FDR), testing for internal replication of results, and four-fold cross-validation. A multivariable model was built using forward selection with any SNP main or interactive effects that passed all three criteria based on pre-determined thresholds. No SNP main effects or SNP-covariate interactions passed all three criteria. 409 SNP-SNP interactions were found to be significant after FDR adjustment, internal replication, and cross-validation. A multivariable model including 4 SNP-SNP interactions explained 11.3% of the variation in LVM in the full GENOA sample and 5.6% of LVM variation in independent test sets. The results of this research underscore that context dependent effects, specifically SNP-SNP interactions, may dominate genetic contributions to variation in complex traits such as LVM.

87

Effects of Inheritance and Shared Family Environments on Dietary Patterns — in a Twin-family Study of Korea

In Kyoung Kim (1), Sung-Il Cho (1), Dong-hun Lee (1), Joohon Sung (1)
(1) Graduate School of Public Health, Seoul National University, Korea

We studied the relative importance of genetic and environmental effects on dietary patterns in a Korean twin-family study.
The Healthy Twin Study recruited same sex adult twins and their family members. A total of 2029 subjects (373 twin pairs, 1283 family members) between 2005 and 2008 were informative for analyses. Dietary intakes were estimated from a semi-quantitative FFQ of 103 food items classified to 17 categories by food ingredient table. Factor analysis was performed to identify eating patterns. After reconstructed factors, heritability (H2) and shared familial environmental effects (C2) were estimated by SOLAR.
Three independent eating patterns were identified: the "**traditional pattern (TP)**" consisting of vegetables, fish/seafood, legumes, kimchi, seaweed, mushrooms, potatoes; the "**Westernized pattern (WP)**" consisting of cereal, meat, eggs, beverages, chips; "**snack pattern (SP)**" consisting of milk and dairy products, fruit, cookies/cracker, potatoes, seeds, nuts. Genetic effects estimated for each consumption patterns (TP 0.25, WP 0.22, SP 0.26, $P<0.001$, adjusting for age, sex). When C2 were included, H2 and C2 were 0.19, 0.05 ($P=0.01$, 0.18) for TP, 0.18, 0.03 (0.01, 0.27) for WP, 0.11, 0.11 (0.08, 0.02) for SP, respectively.
In this study population, all three identified dietary patterns showed significant genetic effects. Only snack-eating pattern was significantly influenced by C2.

88

Heritability of Bone Mineral Density and Body Composition and Their Relationship in Adult Korean Women

Tae-Hun Kim (1), Donghun Lee (1), Joohon Sung (1), Sung-il Cho (1)
(1) Graduate School of Public Health, Seoul National University

Genetic and environmental factors affect bone mineral density (BMD). Twin and family studies have estimated that 38-88% of the variance in BMD at different skeletal sites is genetically determined. Studies showed that in the premenopausal women, lean mass (LM) was the main predictor of BMD, whereas in postmenopausal women, fat mass (FM) predicted BMD better than LM. The objective of this study was to examine the association of FM and LM with BMDs of weight-bearing bones: leg, pelvis, and spine. This study included 872 healthy women of 309 families, aged 30–70 years. In premenopausal women, age-adjusted correlation between LM and BMD at leg, pelvis, and spine were 0.49, 0.35, and 0.31, respectively ($P<0.001$). FM was weakly correlated with BMD at spine ($r=0.13$, adjusted for age, $P<0.03$). In postmenopausal women, age-adjusted correlation between LM and BMD at leg, pelvis, and spine were 0.30, 0.24, and 0.33, respectively ($P<0.001$). FM was correlated with BMD at pelvis ($r=0.18$, adjusted for age, $P<0.001$). Heritabilities of LM, FM, and those of BMD at the leg, pelvis, and spine were all significant after adjusting for age and menopausal status (0.84, 0.64, 0.85, 0.76, and 0.78, respectively). There were stronger association of BMD with LM than with FM. Heritability of LM was also greater than FM. Genetic effects on BMD appeared to be partly through the genetic effects on LM, suggesting that there may be common genes associated with BMD and LM in women.

89

Coevolution Causes Allelic Association Between Physically Unlinked Gamete Receptor Genes

Rori V. Rohlfs (1), Willie J. Swanson (1), Bruce S. Weir (1)
(1) University of Washington

Coevolving interacting genes undergo complementary mutations to maintain their interaction. Different allele combinations between polymorphic coevolving genes may interact differently, conferring varying degrees of fitness. This differential fitness would result in selection for allele matching, which could be observed in a population as linkage disequilibrium (LD)-like allelic association.
Gametic LD is not an appropriate measure of allelic association between physically unlinked coevolving genes. Instead, we propose both standard composite linkage disequilibrium (CLD) and novel genotype association (GA). Using a simple selective model, we simulated loci and calculated power for CLD and GA, showing that the tests can feasibly detect allelic association. We apply CLD and GA tests to the putatively coevolving, polymorphic, and physically unlinked human gamete recognition genes *ZP3* and *ZP3R*.
We observe unusual allelic association not attributable to population structure between *ZP3* and *ZP3R*. This study shows that selection for allele matching can drive allelic association in a present-day human population and that selection can be detected using CLD and GA tests. The observation of this selection is surprising and novel, but reasonable in the system of fertilization. We are currently applying the CLD and GA tests genome-wide to estimate the number of loci associated due to selection for allele matching and to identify classes of coevolving genes.

90

Gene-gene and Gene-environment Interactions Unlikely to Account for Much "Missing Heritability"

Peter Kraft (1)
(1) Harvard School of Public Health

Although genome-wide association studies (GWAS) have identified scores of common genetic markers associated with complex human traits, these markers account for a small proportion of the genetic variance for most traits. Some hypothesize that gene-gene and gene-environment interactions can account for a large proportion of this ''missing heritability.'' Using a version of the polygenic model proposed by Fisher [1918] informed by single-locus marginal effects observed in GWAS, I show that gene-gene and gene-environment interactions involving known loci are unlikely to add much to the total genetic variance, except in extreme scenarios. For example, for a normally distributed trait with twenty known markers, each of which explains 0.5% of the overall trait variance marginally, and interactions that eliminate the marker-trait association among exposed (25% of the population) for ten markers, the interaction component accounts for 25% of the genetic variance, or about 1.2% of the total trait variance. These results imply that in the absence of easily-detectable interactions involving known loci, most of the remaining genetic contribution to complex traits is due to as-yet unknown factors.

91

Segregation Analyses of Familial Barrett's Esophagus

Xiangqing Sun (1), Robert C. Elston (1), Amitabh Chak (1), Gary Falk (2), William Grady (3), Sumeet Mittal (4), Margaret Kinnard (1), Sanford Markowitz (1), Joseph E. Willis (1), Jill Barnholtz-Sloan (1)
(1) Case Western Reserve University
(2) Cleveland Clinic
(3) University of Washington
(4) Creighton University

Esophageal adenocarcinomas (EAC), esophagogastric junction adenocarcinomas (EGJAC), and their precursor Barrett's esophagus (BE), have been shown to aggregate in families. Individuals with familial BE are believed to have an inherited genetic susceptibility to the development of BE and/or EAC/EGJAC. In order to identify the inheritance model for genetic variants that predispose individuals to develop BE or EAC/EGJAC, complex genetic segregation analyses of 881 singly ascertained pedigrees were conducted with the SEGREG program in S.A.G.E. 6.01. Individuals with BE, EAC and/or EGJAC were defined as affected, and the likelihoods were conditioned on proband and founder phenotypes. Results indicated that the model of no major gene segregating was rejected, and that a multifactorial component was insufficient to fully explain the familial BE, EAC and/or EGJAC. For a dominant model with incomplete penetrance, allowing for a sex difference, no heterogeneity between generations was detected. The estimated sex-specific prevalences were 0.011 for males and 0.003 for females — close to what was expected based on published data. For a recessive model with incomplete penetrance, allowing for a sex difference, intergenerational heterogeneity existed even after adjusting susceptibility for founder status. For this model, the estimated sex-specific prevalences were 0.016 for males and 0.004 for females (assuming homogeneity between generations).

92

hzAnalyzer: Detection, Uantification, and Visualization of Contiguous Homozygosity in Eleven Human Sample Populations from High-density Genotyping Datasets Using R and Java

Todd A Johnson (1), Yoshihito Niimura (2), Tatsuhiko Tsunoda (1)
(1) Laboratory for Medical Informatics, Center for Genomic Medicine, RIKEN Yokohama Institute, Yokohama, Kanagawa-ken, Japan
(2) Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

Human genetic variation is often non-randomly organized into regions of restricted diversity in which a limited range of haplotypes can be observed. To examine such patterns, we developed hzAnalyzer, a suite of programs using R and Java, to analyze the genome-wide and localized extent of contiguous homozygosity, which represents one simple means for reducing high-density genotyping datasets into a form by which populations can be compared for their relative haplotype diversity. We analyzed the HapMap phase 3 rel. 2 dataset which includes approximately 1.5 million SNPs for 1,184 individuals from 11 sample populations representing founders from Africa, East Asia, India, and Europe. Our analysis suggests that between closely related populations such as CHB and JPT, much of the genome possesses quite similar characteristics with respect to the localized extent and frequency of contiguous homozygosity, and that our methods can be useful for extracting regions that harbor extended haplotypes that are differentiated between such populations. Similarly, we are developing methods that could be employed in the context of other closely similar populations such as the case and control samples used in genome-wide association studies to detect regions that harbor recessive disease associated variants.

93

Using Mendelian Randomisation to Investigate the Relationship Between Blood Pressure and The Severity of Bbstructive Sleep Apnea

Matthew N. Cooper (1), Gemma Cadby (1), Jessica D. Lee (1), Annette C. Fedson (1), Laila Simpson (1), Kim L. Ward (1), David R. Hillman (2), Sutapa Mukherjee (2), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, The University of Western Australia, Perth, Australia
(2) Western Australian Sleep Disorders Research Institute, Queen Elizabeth II Medical Centre, Perth, Australia

Increased blood pressure (BP) and hypertension have been consistently associated with increased severity of obstructive sleep apnea (OSA), as measured by the apnea-hypopnea index (AHI). Our aim in this study was to investigate the nature of the association between BP and OSA using the SLC30A8 (solute carrier family 30 [zinc transporter], member 8) gene and the techniques of Mendelian Randomization (MR).
Analysis was conducted using the SLC30A8 SNP (rs13266634) as an instrumental variable for systolic BP (SBP). Overnight polysomnography, clinical and questionnaire data was obtained from 653 physician-diagnosed

Caucasian OSA patients as part of the Western Australian Sleep Health Study and was used to explore the 3-way associations between genotype, SBP and OSA severity.
SLC30A8 was a good instrument for SBP ($F > 9$, $P = 0.002$), independent of other risk factors (sex, age, body mass index). Increased SBP was associated with increased loge (AHI) ($P = 0.03$). A formal MR analysis using SLC30A8 as an instrumental variable suggested that SBP was not associated with AHI ($P = 0.47$).
These results suggest that SBP may not be on the etiological pathway to determination of OSA-severity, and that this relationship may reflect confounding or reverse causation. Modulation of this phenotype may therefore have little impact upon the severity of OSA. Our findings also suggest that the search for OSA-severity related genes might usefully exclude known BP and hypertension-related genes.

94

Power Estimates for Mendelian Randomization Studies Using Multiple Genetic Variants in Two-stage Least Squares Regression

Brandon L. Pierce (1), Habibul Ahsan (1), VanderWeele J. Tyler (1)
(1) University of Chicago

Mendelian Randomization (MR) studies assess the causality of exposure (X)-disease (Y) associations using genetic determinants (i.e., instruments) of X. Power and instrument strength requirements for MR studies using multiple genetic variants have not been explored. We simulated datasets consisting of a biallelic locus (G), a normally-distributed X affected by G, and a normally-distributed Y affected by X. We estimated power to detect an effect of X on Y for varying allele frequencies, effect sizes, and samples sizes (using two-stage least squares regression on 1,000 datasets—stage 1 is a regression of X on G; stage 2 is a regression of Y on X). We also report first-stage F statistics; an $F < 10$ indicates a ''weak instrument'' (WI) and biased effect estimates. Similar analyses were conducted using multiple Gs (5, 10, 20) as both independent and combined instruments. Our results suggest that power depends primarily on sample size and the first-stage $R2$. Well-powered MR studies using a single G are unlikely to have WI problems. For a fixed $R2$, F decreases as the number of instrument increases; combining Gs into fewer instruments results in modest power decreases, but maintains higher F statistics. Ideal methods for combining Gs (to avoid WIs) depend upon knowledge of the genetic architecture of X. Based on such knowledge for several candidate exposures, large sample sizes will be required ($n > 1000$).

95

Sampling Ancestries at a Genomic Location Conditional on Data From Surrounding Genetic Markers

Kelly M. Burkett (1), Brad McNeney (1), Jinko Graham (1)
(1) Department of Statistics and Actuarial Science, Simon Fraser University

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. We implemented a Markov Chain Monte Carlo sampler, based on an approach outlined in Zöllner and Pritchard [2005], that samples genealogies compatible with observed haplotype data from unrelated individuals. This presentation will describe the implementation and its extension to unphased genotype data. We will illustrate the use of the sampler by applying it to data from an association study.
Zöllner S and Pritchard JK (2005). Genetics 169:1071–1092.

96

High Resolution Detection of Identity by Descent With Linkage Disequilibrium Modelling

Sharon R. Browning (1), Brian L. Browning (1)
(1) University of Auckland

Individuals are identical by descent (IBD) at a locus if they share genetic material due to co-inheritance from a common ancestor. The IBD concept is applicable to individuals without known relationship (''unrelated'' individuals), as all pairs of individuals share common ancestors at some time in the past. Detection of identity by descent (IBD) in ''unrelated'' individuals has important, wide-ranging applications including relationship inference, population-based linkage analysis (IBD mapping), improved haplotype inference, imputation of ungenotyped variants, genotype error detection, and detection of deletion structural variants. Existing approaches to IBD detection either require markers to be in linkage equilibrium or rely on observed length of identity by state. Typically, such methods can detect IBD regions of length > 5 cM. We have recently developed a new approach to IBD detection that is based on the localized haplotype clustering (BEAGLE) model (Am J Hum Genet 2007 81:1084–1097). Our method estimates posterior probabilities of IBD based on haplotype probabilities, and thus accounts for linkage disequilibrium. We use 1958 British Birth Cohort data to show that our method can reliably detect very small (≤ 2 cM) IBD regions in data from the Illumina 500 K or Affymetrix 500 K platforms. There is a high level of overlap between the detected regions found using the two platforms. With data from a 1 M SNP platform, even smaller regions ($<$ 1cM) can be detected.

97

Population Stratification Analysis Based on Allele Sharing Distance

Xiaoyi Gao (1)
(1) Washington University in St. Louis

Researchers have been using allele sharing distance (ASD) and closely related metrics for population stratification analysis. However, the theory for this practical usage has not been reported. In this work, we describe the theoretical background for using ASD on single nucleotide polymorphism (SNP) genetic data for human population stratification analysis. In showing the proof, we lay the ground work for a general distance-based method for classifying subpopualtions using SNPs and ASD. We also compare its performance to other closely related methods using HapMap Phase I SNP data.

98

Population Structure in Brazilian and other Worldwide Human Populations Revealed by SNP arrays

Suely R. Giolo (1), Júlia M.P. Soler (2), Mariza de Andrade (3), José E. Krieger (4), Alexandre C. Pereira (4)
(1) Department of Statistics, Federal University of Parana and Laboratory of Genetics and Molecular Cardiology, USP, Brazil
(2) Department of Statistics, University of Sao Paulo, Brazil
(3) Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN
(4) Laboratory of Genetics and Molecular Cardiology, Heart Institute, University of Sao Paulo, Brazil

Brazilians are one of the most admixture populations in the world, formed by extensive interethnic crosses between Amerindians, Europeans, Africans and Asians. Although there are several studies that have investigated the genetic structure of several populations, the worldwide coverage remains incomplete with, for instance, populations from South America being underrepresented in databases of human genetic variation. In this work, we have analyzed patterns of genetic variation across 365,116 SNPs genotyped using Affymetrix SNP array 6.0 and/or Illumina 1M in 1,129 unrelated individuals from 12 worldwide populations: one Brazilian and 11 of the HapMap Project, Phase III. In our analysis we have considered only the SNPs that all these populations had in common, excluding those with more than 5% missing entries, not in H-W equilibrium and with minor allele frequency smaller than 1%. PCA applied to this data revealed discernible genetic differentiation among these populations and also that the large degrees of racial miscegenation experienced by the Brazilian population over centuries play an important role in its pattern of genetic variation. We have also analyzed whether small panels of markers could effectively capture the genetic variation revealed by all markers. In this regard, we found that around 500 SNPs could efficiently reproduce the revealed population structure.

99

Assessing Population Stratification Using Mating Type Frequencies

John S. Grove (1), Dongmei Li (1), Norikazu Yasuda (2), Loic LeMarchand (3)
(1) John A. Burns School of Medicine, Univ. of Hawaii
(2) National Institute of Radiological Science, Japan
(3) Cancer Research Center of Hawaii, Univ. of Hawaii

N. Yasuda (Hum Hered 1969 19:444–456) proposed using the distribution of serotypes in husband-wife pairs to estimate population inbreeding, demonstrating that husband-wife pairs had three times the information on inbreeding as independent individuals for codominant systems. For the null hypothesis of random mating, a multivariate version of Yasuda's score test for inbreeding can be constructed to test for population stratification. Loci can be screened for informativeness and the reduced set used in a cluster analysis to search for strata. Five commonly available methods of cluster analysis were compared using simulated data for three thousand husband-wife pairs or six thousand independent individuals with two subpopulations. Genotype frequencies were fixed at their expected values for each set of allele frequencies. The Hierarchical Average method was found to be fairly effective with both equal and unequal stratum sizes and was robust to overstating the number of strata when marital pairs were used instead of individuals. For the extreme case of two subpopulations with two typed SNPs with allele frequencies of (0.2, 0.8) in one stratum and (0.8, 0.2) in the other for both loci, the proportions of couples with misclassified population were 0.003, 0.013, 0.038, and 0.062 for number of clusters set as 2 (correct), 3, 4, and 5. The corresponding misclassification risks were 0.028, 0.12, 0.19, and 0.27 for classifying strata based on individuals.

100

Developing Admixture Mapping Panels for African Americans from Commercial High Density SNP Chips

Guan jie Chen (1), Jie Zhou (1), Ayo Doumatey (1), Daniel Shriner (1), Yuanxiu Chen (2), Noman Gerry (3), Alan Herbert (4), Michael Christman (3), Charles Rotimi (1), Adebowale Adeyemo (1)
(1) CRGGH/NHGRI/NIH
(2) National Human Genome Center at Howard University
(3) The Coriell Institute for Biomedical Research
(4) Boston University

Mapping by admixture mapping works as a powerful tool for identifying genetic variants involved in human disease by exploiting the unique genomic structure in recently admixed populations. While several panels of ancestry informative markers for admixture (AIM) mapping exist, these markers have to be genotyped afresh with each research question. The increasing availability of dense SNP data from genome wide association studies has made it feasible to develop such AIM panels from commercial SNP chips. In this communication, we demonstrate that AIM panels for African Americans can be developed from SNPs on the Affymetrix 6.0 chip and that such panels are as informative for ancestry as existing published AIM panels. Developing AIM panels for admixture mapping from existing dense SNP data offers several advantages, 1. no fresh genotyping needs to be done, thereby saving costs; 2. the markers can be filtered for various quality measures and replaced to minimize gaps; 3. replacement markers are available at no additional cost; 4. several non-overlapping panels can be developed from the same SNP dataset and each used in the admixture mapping project, thereby providing a way to validate findings of significant

genomic regions. The ability to develop AIM panels from commercial SNP GWAS chips provides fresh opportunities to conduct admixture mapping for disease genes in admixed populations, especially where GWAS data exist or are planned.

101
Clustering Based on Genetic Ancestry
Nadia Timofeev (1), Steven H. Hartley (1), Clinton T. Baldwin (2), Daniel A. Dworkis (3), Lindsay A. Farrer (3), Mark Gladwin (4), Elizabeth S. Klings (3), Jacqueline N. Milton (1), Thomas T. Perls (3), Martin H. Steinberg (3), Paola Sebastiani (1)
(1) Boston University, Department of Biostatistics
(2) Boston University School of Medicine, Center for Human Genetics
(3) Boston University School of Medicine, Department of Medicine
(4) Vascular Medicine Branch, National Heart, Lung and Blood Institute

Population stratification is a major confounder causing spurious associations in genome wide association studies. Principal components analysis (PCA) is a widely used approach to detect population substructure and to adjust for stratification by including the top principal components (PCs) in the analysis. An alternative solution to the population stratification problem is genetic matching of cases and controls that requires, however, well defined population strata for appropriate selection of cases and controls. With this objective in mind, we developed a novel algorithm to cluster individuals into groups based upon similar ancestral backgrounds using the PCs computed by PCA. Our algorithm utilizes *k*-means clustering of the most informative PCs to iteratively group individuals into clusters and a novel index that includes accuracy, stability and between-cluster distance to choose the appropriate number of clusters. We tested the algorithm on 7 African populations in the Human Genome Diversity Project to demonstrate that the algorithm accurately clusters individuals. Our algorithm also discriminated fine groupings of Caucasians in a large cohort of centenarians. In addition, examination of the ancestral substructure in a large cohort of African American sickle cell disease (SCD) patients using this algorithm revealed that African Americans with SCD are less admixed than African Americans without SCD and have ancestries similar to the Yoruban, Mandenka or Bantu populations.

102
Evaluation of Different Case-control Matching Designs in Genome-wide Association Studies
Marie-Claude Babron (1), Hervé Perdry (2), Rémi Kazma (3), Simon Heath (4), Mark Lathrop (4), Emmanuelle Génin (1)
(1) INSERM UMR-S946, Paris, France; Univ Paris-Diderot, Paris, France
(2) Univ Paris-Sud, Le Kremlin-Bicêtre, France; INSERM UMR-S535, Villejuif, France
(3) Univ Paris-Sud, Le Kremlin-Bicêtre, France; INSERM UMR-S946, Paris, France
(4) Centre National de Génotypage (CNG), CEA, Evry, France; Fondation Jean Dausset-CEPH, Paris, France

Large-scale genome-wide association studies are increasingly used in multifactorial disease studies. An attractive idea to minimise cost and work load is to gather a common Reference Control Panel (RCP). Each centre only needs to collect and genotype its own case sample, and compare it to the RCP. However, controls must be carefully selected to avoid false positives due to population stratification.
We focus on local studies where cases originate from one country and compare two control matching designs: global matching where the set of cases is compared to a set of controls, and individual matching where each case is compared to one or two genetically matched controls. We investigated three selection schemes: at-random, based on the IBS, and on the Principal Components Analysis (PCA). From the RCP set up by the CNG, 6,000 controls from 13 European countries genotyped on an Illumina 317 K chip, we simulated samples of 600 French cases, based on their genotypes for chosen disease susceptibility SNPs. Up to 1,200 controls were selected among the remaining individuals. The chosen SNPs represent the different stratification characteristics observed across the European populations.
Selection at-random or based on IBS does not perform well. Overall, selecting controls on the basis of the results of PCA is better, both in terms of Type I error and power to detect association. However, even that approach does not correct well when the SNP under study exhibit important stratification.

103
Genomic Inbreeding Coefficients and Runs of Homozygosity by Descent in the HGDP-CEPH Panel of World-wide Populations
Anne-Louise Leutenegger (1), Mourad Sahbatou (2), Howard Cann (2), Emmanuelle Génin (1)
(1) Inserm U946, Paris
(2) Fondation Jean Dausset, CEPH, Paris

We present inbreeding coefficients estimated from genotypes at 642,914 autosomal SNPs for 940 unrelated HGDP-CEPH individuals from 51 populations (Li et'al., Science, 319:1100–4, 2008).
FEstim (Leutenegger et'al., AJHG, 73:516–23, 2003) was used to estimate the inbreeding coefficients from these individuals' genomic information. It is a maximum likelihood method that uses a hidden Markov chain to model marker genotype and homozygosity-by-descent (HBD) dependencies along the genome. To do so, a marker map without linkage disequilibrium (LD) is required; this is not the case for very dense SNP maps. We developed a procedure to generate multiple LD free sub-maps and to combine their information. 75% of the original map could be captured using 1,000 sub-maps.
This approach also allows inference of likely HBD genomic regions. Based on these inferences, we propose a test to identify genomic regions with excess HBD among the HGDP-CEPH individuals. Regions with excess HBD can be the consequence of recent selective pressures and will be compared to the ones already identified using other methods.

104

Local Structures and Ancestry-informative Markers in the Quebec Population

Ilija P Kovac (1), Michael Phillips (2), Yassamin Feroz Zada (1), Louis-Philippe Lemieux Perreault (2), Jean-Claude Tardif (2), Marie-Pierre Dubé (2)
(1) Montreal Heart Institute Research Center, QC, Canada
(2) Montreal Heart Institute, Research Center, QC, Canada; Faculty of Medicine, Université de Montréal

We examine local populations in Quebec (QC) to identify ancestry informative markers (AIMS), to improve genetic studies. The 174 subjects are selected for grandparent birthplace from a large ongoing study on statin pharmacogenomics in QC. We defined four groups, with four grandparents from the Saguenay Lac Saint Jean region (SLSJ, $n = 67$), Montreal city (MTL, $N = 52$), Quebec City (VQC, $N = 24$), and admixed with grandparents from QC and world regions (QC_ET, $N = 31$). Subjects were genotyped on the Illumina Human 1 M duo_v3 platform. We used 993692 polymorphic SNPs with $>95\%$ call rate. Further checks included plating effects, heterozygosity, sex, and IBS relative check. We excluded 1 subject from each of 2 sib pairs (SLSJ and QC_ET), retaining 172 samples. We did principal component analysis using independent SNPs in our 172 samples combined with 59 genotyped CEU parents. The first component separated the SLSJ, the second component separated CEU from MTL and QC_ET. The SLSJ-MTL and CEU-MTL differentiation were Fst = 0.004 and 0.001 respectively. We identified candidate AIMs from 993692 SNPs, by the absolute allelic frequency difference. For SLSJ-MTL, we identified 44 SNPs with difference ≥ 0.3 and 365 SNPs >0.25. For CEU-MTL, we identified 18 SNPs with difference ≥ 0.3 and 115 SNPs >0.25. These results demonstrate local differentiation. Candidate AIMs are expected to improve genetic studies in the Quebec population. We are proceeding with further analyses.

105

Genetic Distance and Population Structure Analysis of Parents Drawn from a Family Based Genome Wide Association Study of Oral Clefts.

Tanda Murray (1), Beaty H. Terri (2), Hetmanski B. Jacqueline (2), Scott F. Alan (3), Liang Kung-Yee (2), Ruczinski Ingo (2), Wu Tao (2), Redetta A. Richard (3), Marazita L. Mary (4), Murray C. Jeffery (5), Munger G. Ronald (6), Wilcox Allen (7), Lie T. Rolv (8), Wu-Chou Yah-Huei (9), Wang Hong (10), Huang Shangzhi (11), Yeow Vincent (12), Chong S. Samuel (13), Jee Sun Ha (14), Christensen Kaare (15), GENEVA Consortium (16)
(1) Johns Hopkins University, School of Public Health
(2) Johns Hopkins University,School of Public Health
(3) Johns Hopkins University, School of Medicine
(4) University of Pittsburgh, Health Sciences
(5) University of Iowa, Children's Hospital
(6) Utah State Universiy
(7) NIEHS/NIH, Epidemiology Branch
(8) University of Bergen
(9) Chang Gung Memorial Hospital
(10) Peking University Health Science Center
(11) Union Medical College
(12) Wuhan University, School of Stomatology
(13) National University of Singapore
(14) Yonsei University, Epidemiology and Health Promotion
(15) University of Southern Denmark
(16) NIDCR/NHGRI

Using marker data on 4599 parents of oral cleft cases drawn from a genome wide association study, we investigated the structure of 13 different populations using principal components analysis (PCA) and estimated Wright's FST to measure genetic distance. This study used a case-parent study design where cases with isolated clefts were recruited from 13 populations in 9 countries. Approximately 45.1% these parents were of European ancestry and 51.9% were of Asian ancestry, with modest representation from other racial groups. From the 580,307 single nucleotide polymorphic (SNP) markers, 40,000 highly polymorphic, independent autosomal SNPs were selected for PCA [with pairwise linkage disequilibrium $r^2 < 0.12$]. Asian-only PCA revealed tight clustering among the 3 sub-populations in China, and contrasting the first principal component (PC1), with the third (PC3) revealed a north-south genetic cline from Korea through China to Taiwan and Singapore. Filipinos formed a distinct cluster along PC1. Genetic distances (FST) reflected these structural similarities among Asians with values between 0.0-0.006 for Asian populations, and greater distances from Filipinos ($0.01 < FST < 0.02$). European Americans (Utah, Pittsburg, Iowa, and Maryland) clustered with the Norwegian and Danish populations. Genetic distances between Europeans and Asians were substantial ($0.08 < FST < 0.11$). These genetic distances between Asian and European populations should be considered in analyses of GWAS.

106

Accurate IBD Inference Identifies Cryptic Relatedness in 9 HapMap Populations.

Apostolos Dimitromanolakis (1), Andrew D Paterson (2), Lei Sun (1)
(1) Dalla Lana School of Public Health, University of Toronto
(2) Genetics and Genomic Biology, The Hospital for Sick Children, Toronto

The HapMap project has been very successful in mapping human genetic variation. If numerous cryptic relationships exist among founder individuals in the dataset, the accuracy of allele frequency calculations, estimation of LD patterns, selection of tag SNPs, population stratification analysis and association studies could all be adversely affected. We extended our previous work [McPeek and Sun, 2000], developed for genome-wide linkage data, with PREST-plus, suitable for both population and family based genome-wide association studies, and we applied an accurate likelihood-based IBD inference via the EM algorithm to the most recent release (27, phase 3) of HapMap data. Cryptic relatedness was detected in 195 of the 1002 founders (20%). Among the 271 founder pairs that were shown to share more that 10% of their genome (p.IBD.0 <0.9), we found strong evidence for first-degree relatives, including 21 parent-offspring and 27 full-sibling pairs. The results were validated using the method of

moments IBD inference algorithm implemented in PLINK [Purcell et'al., 2007]. We would like to emphasize the importance of checking for cryptic relatedness in genetic studies and provide an outline of the necessary steps to achieve this, by the use of PREST-plus and companion post-processing R scripts assisting the interpretation of the results.

107
Ancestry Informative Markers and Family-Based Association
Albert M. Levin (1), Indrani Datta (1), James Yang (1), Micheal C. Iannuzzi (2), Paul M. McKeigue (3), Benjamin A. Rybicki (1), Courtney L. Gray-McGuire (4)
(1) Henry Ford Health System
(2) State University of New York Upstate Medical University
(3) University of Edinburgh Western General Hospital
(4) Oklahoma Medical Research Foundation

Population differences in allele frequencies and linkage disequilibrium are the premises for admixture mapping. Such mapping relies on the fact that ancestry informative markers (AIMs) have stricking differences among populations. By comparing AIMs in an admixed population to that of their founder populations, one can identify the ancestral population from which a given genomic segment originates. Because the existing methods (and accompanying software) for admixture analysis are limited to independent observations, there has been little to no discussion about the use of AIMs with family data. However, because many linkage and family-based association studies have been conducted in samples amenable to admixture mapping, studies like ours for mapping sarcoidosis susceptibility genes in African Americans are likely to arise. We therefore present practical issues that arise when using AIMs with family data, including difficulties in estimating relationships and how the skewness of allele sharing can affect analysis. For example, using the relationship inference software RELPAIR, which assumes that the observed genomes derive from a single ancestral population, we identified a large group of individuals which were unrelated, but $>50\%$ Caucasian and thus appeared to be first-degree relatives. This finding is not only relevant to relationship checking with AIMs but also in performing family-based association in admixed populations.

108
Adjustment for Population Structure in Association Studies Using Markov Kernels
Omar De la Cruz (1)
(1) Stanford University

Spectral methods, like principal components analysis (PCA), have been proposed for detecting and quantifying population structure and adjusting statistical tests of association. Through the use of kernels, non-linear alternatives to PCA can be reduced to the linear case; however, it is not clear which kernel is the most appropriate in the case of population structure.

In this work we describe a kernel that has a probabilistic interpretation as a Markov kernel, measuring the probability of flow of genetic information between different points of a synthetic landscape that represents the population structure (if such structure is due only to geographic distribution, that landscape would correspond to the geography). This way, we can interpret our selected axes of variation as a low dimensional basis for approximate solutions to an integral equation that generalizes the Hardy-Weinberg equilibrium. Thus, elements of this basis can be easily incorporated into association tests to account for the population structure.

109
Simultaneous Genotype Calling and Haplotype Phase Inference Improves Genotype Accuracy and Reduces False Positive Associations for Genome-wide Association Studies
Brian L. Browning (1), Zhaoxia Yu (2)
(1) The University of Auckland, Auckland, New Zealand
(2) University of California Irvine, Irvine, California

We present a new method that performs simultaneous genotype calling and haplotype phase inference. Our method employs the computationally efficient BEAGLE haplotype frequency model and can be applied to large-scale studies with millions of markers and thousands of samples. We compare our method to state of the art genotype calling methods using genotype data called with the GenCall, Illuminus, Chiamo, and Birdseed genotype-calling algorithms from the Illumina 550 K and 1 M arrays and Affymetrix 500 K and 6.0 arrays. For Affymetrix data, our method reduces discordance rates with high-quality Illumina genotypes by a factor of 3 or more. For Illumina data, our method improves genotype accuracy and reduces missing data by an order of magnitude.
We have re-called genotype data for the Wellcome Trust Case Control Consortium Bipolar Disease study (Nature 2007;447:661–678). More than 90% of the known false positive association signals caused by genotyping artefacts in the original study are eliminated when using genotype calls from our new method. The phased haplotypes produced by our genotype calling method also eliminate a similar proportion of the false positives association signals that occur in a BEAGLE haplotypic analyses of these data (Hum Genet 2008;123:273–280). Our new genotype calling methods are freely available from www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html.

110
What's the Best Statistic for a Simple Test of Genetic Association in a Case-control Study?
Chia-Ling Kuo (1), Eleanor Feingold (2)
(1) University of Pittsburgh, Department of Biostatistics
(2) University of Pittsburgh, Department of Human Genetics and Biostatistics

Genome-wide genetic association studies typically start with univariate statistical tests of each marker. In principle, this single-SNP scanning is statistically straightforward—the testing is done with standard

methods (e.g. chi-squared tests, regression) that have been well-studied for decades. However, a number of different tests and testing procedures can be used. In a case-control study, one can use a 1 df allele-based test, a 1 df or 2 df genotype-based test, or a compound procedure that combines two or more of these statistics. Additionally, most of the tests can be performed with or without covariates included in the model. While there are a number of statistical papers that make power comparisons among subsets of these methods, none has comprehensively tackled the question of which of the methods in common use is best suited to univariate scanning in a genome-wide association study. In our work, we consider a wide variety of realistic test procedures, and first compare the power of the different procedures to detect a single locus under different genetic models. We then address the question of whether or when it is a good idea to include covariates in the analysis. We conclude that the most commonly-used approach to handling covariates — modeling covariate main effects but not interactions — is almost never a good idea. Finally, we consider the performance of the statistics in a genome scan context.

111

Automated Evaluation of Signal Intensity Plots — Cluster Validity Measures are Great

Arne Schillert (1), Olof-Joachim Frahm (1), Tanja Zeller (2), Daniel F Schwarz (1), Stefan Blankenberg (2), Andreas Ziegler (1)
(1) Universitaet zu Luebeck, Inst. f. Med. Biometrie u. Statistik
(2) Johannes Gutenberg University Mainz, Dept. Of Medicine II

The visual inspection of signal intensity plots by two experienced independent readers is standard in the quality control of genome-wide association studies (GWA). Because of high costs this process is restricted to candidate single nucleotide polymorphisms (SNPs) in practice. Thus, the valid judgement of signal intensities for all SNPs from a GWA through an automated procedure would be helpful. Even more so, meta-analyses which require imputed genotypes and machine learning approaches like random forests would greatly profit from this. So far, only few approaches for automatically evaluating signal intensities have been proposed.
We embed the problem of signal intensity plot inspection in the well-developed theory of measuring cluster validity in cluster analyses. We propose various measures for judging cluster compactness, cluster homogeneity, cluster connectedness, cluster separability and combinations of these criteria. We jointly evaluate these criteria through a random forest approach and propose a simple combination of the aforementioned criteria. The criteria are evaluated using 3300 samples from the Gutenberg Heart Study. The gold standard for signal intensity plots is provided through ratings from two independent readings of experienced readers for 5000 SNPs which successfully passed the standard quality criteria.
We show that a simple combination of cluster validity measures show satisfactory agreement between visual and automated inspection of signal intensity plots.

112

GWAMA: Software Tool for Meta-analysis and Visualization of Whole Genome Association Data

Reedik Magi (1), Andrew P. Morris (2)
(1) Wellcome Trust Centre for Human Genetics; Oxford Centre for Diabetes, Endocrinology and Metabolism
(2) Wellcome Trust Centre for Human Genetics

We have developed the GWAMA (Genome-Wide Association Meta-Analysis) software to perform meta-analysis of the results of genome-wide association studies of binary or quantitative phenotypes. Fixed-effects meta-analyses are performed for both directly genotyped and imputed SNPs using estimates of the allelic odds ratio and 95% confidence interval for binary traits, and estimates of the allelic effect size and standard error for quantitative phenotypes. The software incorporates error trapping facilities to identify strand alignment errors and allele flipping, and performs tests of heterogeneity of effects between studies. Comprehensive log files of all errors and warnings are composed with unique error codes for each exception. This allows researchers to quickly and easily discover all problems in their dataset.
The software package also includes R scripts for creating publication quality Manhattan- and QQ-plots from the output data to summarise results. Both Manhattan- and QQ-plots can be drawn for illustrating the results.
GWAMA is an open source software package and can be downloaded from http://www.well.ox.ac.uk/gwama/. A user-guide and example files are provided with the software.

113

Coronary ARtery DIsease Genome-wide Replication And Meta-Analysis (CARDIoGRAM) — Design of a prospective meta-analysis of 14 genome-wide association studies

Inke R. König (1), John R. Thompson (2), Michael Preuss (3), Themistocles L Assimes (4), Stefan Blankenberg (5), Eric Boerwinkle (6), Adrienne Cupples (7), Stephen Epstein (8), Alistair Hall (9), Christian Hengstenberg (10), Sekar Kathiresan (11), Reijo Laaksonen (12), Winfried März (13), Ruth McPherson (14), Christopher J O'Donnell (15), Thomas Quertermous (4), Daniel Rader (16), Muredach Reilly (16), Robert Roberts (14), Alex Stewart (14), Unnur Thorsteinsdottir (17), Andreas Ziegler (1), Jeanette Erdmann (18), Nilesh J Samani (19), Heribert Schunkert (18), on behalf of CARDIoGRAM (20)
(1) Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Germany
(2) Department of Health Sciences and Genetics, University of Leicester, United Kingdom
(3) Institut für Medizinische Biometrie und Statistik and Medizinische Klinik II, Universität zu Lübeck, Germany
(4) Department of Medicine, Stanford University School of Medicine
(5) Department of Medicine II, Johannes Gutenberg-University Mainz, Germany
(6) Human Genetics Center, University of Texas Health Science Center at Houston
(7) Department of Biostatistics and Epidemiology, Boston University
(8) Cardiovascular Research Institute, MedStar Research Institute, Washington Hospital Center

(9) LIGHT, University of Leeds, United Kingdom
(10) Klinik und Poliklinik für Innere Medizin II, Universität Regensburg, Germany
(11) Cardiovascular Research Center and Center for Human Genetic Research, Massachusetts General Hospital
(12) Science Center, Tampere University Hospital, Finland
(13) Synlab Medizinisches Versorgungszentrum für Labordiagnostik Heidelberg, Germany
(14) Division of Cardiology, University of Ottawa Heart Institute, Canada
(15) National Heart, Lung, and Blood Institute and its Framingham Heart Study, National Institutes of Health
(16) Institute for Translational Medicine and Therapeutics and Cardiovascular Institute, University of Pennsylvania School of Medicin
(17) deCODE Genetics, Iceland
(18) Medizinische Klinik II, Universität zu Lübeck, Germany
(19) Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, United Kingdom
(20).

Coronary artery disease (CAD) is the leading cause of death in U.S. and Europe and is a heritable trait. Classical genome-wide association studies (GWAS) uncovered at least associated 13 loci. However, each variant confers a modest effect and explains a small fraction of inter-individual variation only.
We assembled the CARDIoGRAM Consortium pooling data from ADVANCE, CADomics, CHARGE, deCode, GerMIFS I-III (KORA), LURIC/AtheroRemo, MedStar/PennCath, MIGen, OHGS, and WTCCC. It comprises >22,000 well-characterized cases and >60,000 controls. In each study, genome-wide genotyping was carried out, and imputation was conducted to generate genotypes for ~2.2 million SNPs in each study.
Standard operating procedures were generated to harmonize data analyses. Extensive quality control was performed study-wise and centrally for standardized data formats. With the assembled sample size, the power to detect modest effects is substantially increased; for genome-wide significance, the power is about 80% for an odds ratio of 1.1, if the minor allele frequency is at least 10%.
Meta-analyses for the CAD phenotype and for important subgroups including myocardial infarction and early-onset CAD will be carried out. Following this, wet lab replication genotyping of top results will be sought in >15,000 additional cases and 15,000 controls. CARDIoGRAM brings together an enormous wealth of GWAs data, thus representing the largest study to date to uncover the inherited basis of CAD.

114
Withdrawn

115
Genome-wide Association Study on HDL Cholesterol Level in the Marshfield Personalized Medicine Research Project as Part of the eMERGE Network

Stephen D. Turner (1), Catherine A. McCarty (2), Yuki Bradford (1), Dick Berg (3), Peggy Peissig (3), Jim Linneman (3), Justin Starren (3), Russell A Wilke (4), Marylyn D. Ritchie (1)
(1) Center for Human Genetics Research, Vanderbilt University
(2) The Center for Human Genetics, Marshfield Clinic Research Foundation
(3) Biomedical Informatics Research Center, Marshfield Clinic Research Foundation
(4) Department of Medicine and Department of Pharmacology and Toxicology, Medical College of Wisconsin

Individuals having below-normal levels of HDL cholesterol are at increased risk for cardiovascular disease. To more thoroughly understand the genetic basis of this complex trait, we have conducted an initial GWAS in 3900 participants in the Marshfield Clinic Personalized Medicine Research Project (PMRP) as part of the Electronic Medical Records and Genomics (eMERGE) network. Self reported race within the PMRP cohort indicates the population is 98.2% white Caucasian of central and northern European descent, making it ideal for a genetic association study since risk of population stratification is minimized. Clinical lab HDL cholesterol levels were electronically harvested from subject EMRs in 3900 individuals. All samples were genotyped with the Illumina 660-Quad platform. Each SNP passing rigorous quality control measures was tested for association with HDL cholesterol level using linear regression, assuming an additive model. Here we report several highly significant findings, shedding more light on the genetic contribution to a complex lipid phenotype of substantial public health interest. In addition, we compare and contrast these results with the many already published SNPs associated with lipid levels. Finally, we demonstrate the success of using an EMR-derived phenotype to perform genetic analysis of complex disease. This study along with the other eMERGE projects provide evidence supporting the use of biobanks linked to EMRs for genomic studies.

116
Using Ascertainment with Two Stage Genome Wide Association Studies can Save Resources

Michael D. Swartz (1), Bo Peng (1), Sanjay Shete (1)
(1) University of Texas M. D. Anderson Cancer Center

Researchers continue to use Genome wide association studies (GWAS) to find the genetic markers associated with disease. Although the cost of genotyping continues to go down, GWAS require large samples; and methodologies for GWAS must continue to strike a balance between cost and power. To this end, we propose a two stage design that can reduce the cost of a GWAS study without sacrificing power. We introduce an ascertainment scheme where we ascertain only those cases and controls with the presence of a risk allele to move forward to stage two. We simulate complex diseases with multiple causal SNPs to evaluate our method versus a typical two stage design. The simulation studies show that by ascertaining individuals from stage one, researchers can substantially reduce the cost of the study, without increasing false positives and only a modest decrease in power compared to standard two stage designs.

117

Identifying Ancestry and Sample Integrity Issues Using Study-wide Pairwise Concordance

Dan J. Serie (1), Brooke L. Fridley (1), William Bamlet (1), Tom A. Sellers (2), John D. Potter (3), Ellen L. Goode (1)
(1) Mayo Clinic
(2) H. Lee Moffitt Cancer Center & Research Institute
(3) Fred Hutchinson Cancer Research Center

Assurance of sample integrity is necessary in any genetic study. Numerous programs test for Mendelian errors; however, sample switches can be more difficult to distinguish. To this end, we integrated pairwise concordance estimates into a comprehensive quality-control framework using data from the Colorectal Cancer Family Registry (1855 individuals; 327 families; Affy10k) and a Mayo Clinic case-control ovarian cancer study (396 cases; 469 controls; 1479 SNPs). Family analysis used PedCheck, PREST, and EIGENSTRAT; and using principal components, we were able to specify a unique race for each family with missing/discordant race data at the individual level. To augment these tests, we estimated genotype concordance for all sample pairs using SAS and PLINK. Among families with Mendelian errors, concordance estimates $\sim$0.5 identified unrelated individuals, and estimates >0.65 identified labeling errors amongst presumed-unrelated individuals. Thus, the use of pairwise concordance allowed for the inclusion of an additional 12 families (24 individuals) which would have been excluded from family-based analysis. This approach also has utility in association studies, where cryptic relatedness can impair variance estimation. We used this method in an ovarian cancer study to reveal six pairs of individuals with sibling-level relatedness that was then accounted for in association-testing. We conclude that simple techniques can help maximize use of valuable samples in genetic epidemiology.

118

Integrating Large-scale Genetic and Monocyte Expression Data Reveals Major Trans Regulators of Biological Processes

Maxime Rotival (1), Tanja Zeller (2), Philipp Wild (2), Silke Szymczak (3), Arne Schillert (3), Thomas Munzel (2), Françoois Cambien (1), Andreas Ziegler (3), Laurence Tiret (1), Stephan Blankenberg (2)
(1) INSERM UMRS 937, Paris France; Université Pierre et Marie Curie Paris, France
(2) II. Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universität Mainz, Mainz, Germany
(3) Institut für Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany

In a large population-based cohort ($n = 1{,}490$), we have analyzed 12,808 well characterized genes expressed in monocytes — a key player in the development of atherosclerosis and the immune response — in relation to genome-wide variability (675,350 Single Nucleotide Polymorphisms — SNPs). We applied first a method of extraction of expression patterns — Independent Component Analysis — allowing us to identify 91 transcriptional modules of co-regulated genes, several of them being significantly enriched in functional pathways. These patterns were then related to SNPs. At a study-wise significance threshold of 8.14×10^{-10}, we found 8 blocks of SNPs associated with expression patterns. One block mapped to the ERBB3 gene previously identified as a susceptibility gene for type 1 diabetes (T1D). As the associated pattern did not contain ERBB3, an effect via ERBB3 expression could be excluded. The associated pattern suggested MADCAM1, an endothelial adhesion cell molecule, as a plausible mediator linking the ERBB3 locus to T1D. Another block mapped to the ARHGEF3 gene which is involved in Rho signaling pathway activation. The associated pattern was strongly enriched in genes involved in cell adhesion and platelet activation. This study demonstrates how coupling genome-wide expression and genetic variability in large cohorts may help to identify major trans regulators.

119

Rare Variants with Recessive Effects Identified in a Genome-wide Association Study of Parkinson's Disease

Jeanne C. Latourelle (1), Audrey E. Hendricks (2), Jemma B. Wilk (1), Anita L. DeStefano (1), Nathan Pankratz (3), Tatiana Foroud (3), Richard H Myers (1)
(1) Boston University School of Medicine
(2) Boston University School of Public Health
(3) Indiana University

Genomewide association studies (GWAS), based on the common disease/common variant (CDCV) hypothesis, are powerful tools for finding common variants with modest effects, but are not ideal for detecting rare variants. Little evidence supports the CDCV hypothesis in Parkinson's Disease (PD), while known genes such as SNCA and Parkin show clear evidence of the influence of rare mutations. Typical GWAS may not appropriately test for rare variants with recessive effects. In a recent PD GWAS, 3711 SNPs were excluded from study under a recessive model because of a lack of controls homozygous for the rare allele. One alternative method for studying rare variants uses Hardy-Weinberg equilibrium (HWE) to predict the expected number of rare genotypes at a locus. By identifying SNPs that depart from the expected genotype frequency, we can identify associations that may be missed by conventional analysis. To identify possible PD-associated rare variants we flagged SNPs with ≥ 5 rare homozygote cases, but no rare homozygote controls. We examined SNPs in a second PD GWAS by Fung et'al. and excluded previously flagged SNPs that showed any rare homozygotes in that control sample, identifying 91 SNPs with potential recessive effects on PD. We then calculated the expected number of rare homozygotes in the case sample based on allele frequency in both cases and controls and an exact binomial probability to test whether the observed frequency was greater in the PD cases than expected under HWE.

120

A Minimum Encoding Approach to Analysing GWAS Data

Enes Makalic (1), Daniel F. Schmidt (1), Mark Jenkins (1), John Hopper (1)
(1) University of Melbourne, Centre for Molecular, Environmental, Genetic and Analytic Epidemiology

Aims and Objectives: A standard approach to GWAS analysis to minimise false positive findings (Type I error) is to apply a suitable correction for multiple hypothesis testing. We present an alternative method that aims to minimise the sum of Type I and Type II errors by using the data to determine the optimal significance level.
Methods: Based on information theory and data compression (Minimum Description Length), we developed a method using the MATLAB software environment. For each SNP, the phenotype data was split into two sets based on the SNP genotype. The model tests whether the two sets of phenotype data compressed more efficiently than the single combined set. This is true when there exists an association between the SNP and the phenotype. An advantage of the resulting method is that it provides a posterior distribution over the tested hypotheses which may be integrated into a decision theoretic post-testing analysis.
Results: The developed framework for multiple hypothesis testing was tested on real and simulated GWAS data. The results suggest that the new method is equivalent to standard methods for SNPs with a high MAF in large studies, but appears to offer an improved power to detect associations for small GWAS studies and SNPs with rare MAF.
Conclusion: This is useful for studying rare diseases where obtaining a large sample is infeasible. Further, it appears to have superior power to detect SNPs that may be causal or linked to rare causal genetic factors.

121
The Western Australian Melanoma Health Study (WAMHS)
Sarah V. Ward (1), Gemma Cadby (1), Judith M. Cole (2), Fiona M. Wood (3), Michael Millward (1), Lyle J. Palmer (1)
(1) The University of Western Australia
(2) Western Australian Melanoma Advisory Service
(3) McComb Foundation

Melanoma is the most aggressive form of skin cancer and Western Australia (WA) has one of the highest rates in the world. The major environmental and host risk factors have been identified but little is known about its genetic causes or how they interact with environmental factors.
The Western Australian Melanoma Health Study (WAMHS) is a population-based database and biospecimen resource enabling investigation into the clinical and genetic epidemiology of melanoma, and the natural history of scarring post melanoma excision. The study aims to recruit all new adult incidence cases of invasive cutaneous melanoma in WA from January 2006 onwards. We aim to collect data on at least 2000 individuals, with over 1250 consented to date. Data includes family history, medical history, an overview of lifetime sun exposure, skin characteristics and scar management techniques. Biospecimens (DNA, RNA and serum), clinical pathology data and scar assessment measurements are also collected.
Planned analyses include a GWAS and gene-environment interaction studies to identify new genes underlying melanoma risk and natural history, and those underlying scar progression and healing. The WAMHS will be critical for understanding the importance and functional roles of melanoma genes in the general Australian population, and their relationship to environmental factors. It will also lead to more specific interventions to reduce the impact of scarring on patients.

122
Unbiased Estimation and Inference for Replicated Associations Following a Genome Scan
Jack Bowden (1), Frank Dudbridge (1)
(1) MRC Biostatistics Unit

It is well-known that standard estimates of effect sizes are upwardly biased for the strongest associations in a genome scan, but are unbiased in replication samples. Therefore estimates are often obtained only from replication data, but measures of significance may be combined across the scan and replication data, leading to inconsistencies between confidence intervals and *P*-values. We recently proposed a method for obtaining unbiased estimates of odds ratios, combining data from a genome wide scan and replication study. Our estimator has smaller variance than that based on the replication data alone, but no closed form expression for its actual variance is known. We now describe an efficient algorithm for obtaining a bootstrap variance estimate, giving properly calibrated confidence intervals that use all the data from both stages, hence providing increased power for hypothesis testing. We show that unadjusted *P*-values can be severely biased, whereas combining replication *P*-values with Bonferroni adjustments from a genome scan leads to conservative tests. Our approach gives consistency between estimation and inference following a genome scan, with a natural lead into sequential- and meta-analyses. We highlight further aspects including the perspective that the strongest associations are random effects, which affects the definition of variance, and the implications of the replication sample size being determined by the results of the genome scan.

123
Functional annotation of GWAS hits
Jo Knight (1), Mike R. Barnes (2), Gerome Breen (1), Mike E. Weale (1)
(1) King's College London School of Medicine
(2) GlaxoSmithKline

Genome wide association studies (GWAS) have provided new clues about the aetiology of complex genetic diseases. In addition to a small set of highly significant results, these studies have produced thousands of moderately suggestive results, requiring new systematic approaches to prioritise their follow up. A number of databases and weighting schemes have been developed to allow researchers to use annotation information, but none are empirically based.
In order to calibrate a prioritization scheme, we have compared the functional annotations of a comprehensive list of GWAS hits against random expectation. We have explored three important annotation categories: non-synonomous SNPs, promoter SNPs and eQTLs. We investigated annotated SNPs plus their linkage disequilibrium proxies.

We found an increased proportion of annotation in the GWAS hits compared to the random SNPs in all three categories; as shown below for one of our analyses.

	eQTL (%)	Non-synonomous (%)	Promoter (%)
GWAS hits	13.7	7.9	3.6
Random	7.7	2.7	1.5

This pattern remained even when the SNPs were stratified by minor allele frequency; when the MHC region was excluded; or when SNPs belonging to more than one annotation category were removed.
Our study demonstrates that GWAS hits are enriched for these three functional categories, and hence it would be appropriate to provide a higher weighting for such SNPs when planning follow up studies for GWAS.

124

Novel Genetic Loci Implicated in Fasting Glucose Homeostasis and Their Impact on Related Metabolic Traits

Josee Dupuis (1), Claudia Langenberg (2), Inga Prokopenko (3), Richa Saxena (4), Nicole Soranzo (5), Anne U. Jackson (6), Eleanor Wheeler (5), Nicole L. Glazer (7), Nabila Bouatia-Naji (8), Laura McCullogh (9), Anna Gloyn (9), Robert Sladek (10), Philippe Froguel (8), Richard M. Watanabe (11), James B. Meigs (12), Leif Groop (13), Michael Boehnke (6), Mark I. McCarthy (3), Jose C. Florez (4), Inês Barroso (5), for the MAGIC investigators (14)
(1) Boston University School of Public Health, Boston, MA and National Heart, Lung, and Blood Institute's Framingham, MA
(2) MRC Epidemiology Unit, Cambridge, United Kingdom
(3) WTCHG and OCDEM, Oxford, United Kingdom
(4) Massachusetts General Hospital and Harvard Medical School, Boston, MA and Broad Institute, Cambridge, MA
(5) Wellcome Trust Sanger Institute, Cambridge, United Kingdom
(6) University of Michigan School of Public Health, Ann Arbor, MI
(7) University of Washington, Seattle, WA
(8) CNRS-UMR8090, Lille, France
(9) OCDEM, Oxford, United Kingdom
(10) McGill University, and Genome Quebec Innovation Centre, Montreal, Canada
(11) University of Southern California, Los Angeles, CA
(12) Harvard Medical School, and Massachusetts General Hospital, Boston, MA
(13) Lund University, University Hospital Malmo, Malmo, Sweden
(14) Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)

Impaired β-cell function and insulin resistance are key determinants of type 2 diabetes (T2D). Genetic analyses of these continuous diabetes-associated traits led to the identification of a novel locus, *MTNR1B*, whose glucose-raising allele also increases T2D risk. To identify additional diabetes-related trait loci and investigate their metabolic impact, we performed meta-analyses of 21 genome-wide associations studies informative for fasting glucose (FG, $N = 46{,}263$), fasting insulin, and indices of β-cell function (HOMA-B) and insulin resistance (HOMA-IR) ($N = 38{,}413$). Follow-up of 25 loci in 61,219 independent samples discovered nine new genome-wide significant ($P < 5 \times 10^{-8}$) associations with FG (in or near *ADCY5*, *MADD*, *ADRA2A*, *CRY2*, *FADS1*, *GLIS3*, *SLC2A2*, *PROX1* and *FAM148B*) and a novel locus for fasting insulin and HOMA-IR (*IGF1*). Also associated with FG were established T2D loci *TCF7L2* and *SLC30A8*, and previously reported loci *GCK*, *GCKR*, *G6PC2*, *MTNR1B* and *DGKB/TMEM195*. A variant in *GCKR* also achieved genome-wide significant association with fasting insulin and HOMA-IR. The impact of FG loci on T2D and related metabolic traits suggests shared genetic determinants: *DGKB/TMEM195*, *ADCY5* and *PROX1* with T2D and *FADS1/FADS2* and *MADD* with lipid levels. The wealth of novel FG loci contrasts with the sole fasting insulin/HOMA-IR novel finding and suggests a different genetic architecture for β-cell function and insulin resistance.

125

Finding Unique Filter Sets in PLATO: Preparation for Efficient Interaction Analysis

Benjamin J. Grady (1), Scott Dudek (1), Marylyn D. Ritchie (1)
(1) Vanderbilt University

The methods to detect interactions between variants in GWAS datasets have not been well developed to this point. PLATO is a filter-based method bringing together many analytical methods in an effort to solve this problem. It filters a dataset down to a subset of genetic variants, which may be useful for interaction analysis. To streamline PLATO for efficient data analysis, we determined which of 24 analytical filters produced redundant results. Using a kappa score to identify agreement between filters, we grouped the methods into 4 meta-groups. We then tested the MAX statistic put forth by Sladek et'al. [Sladek et'al., Nature 445:881–885 (2007)]. One filter from each meta-group was run on 100 simulated datasets containing 1.5 OR additive, dominant and recessive single-locus effects. They were also run on 1000 datasets containing no effect to determine the Type 1 error rate. To find the MAX statistic, the four filters were run on each SNP in each dataset and the smallest p-value among the four results was taken. Permutation testing was performed to empirically determine the p-value. The power of the MAX statistic to detect each of the three effects was determined as were the Type 1 error and false positive rates. The results show that PLATO using the MAX statistic has higher power to find all types of effects and a lower false positive rate than any of the individual filters alone. In the future we will extend the concept of PLATO with the MAX statistic to interaction analyses.

126

A Comprehensive Look at the Likelihood and Bootstrap Approaches to Overcome the Winner's Curse in GWAS

Laura L. Faye (1), Lei Sun (2), Apostolos Dimitromanolakis (1), Shelley B. Bull (1)
(1) Dalla Lana School of Public Health, University of Toronto & Samuel Lunenfeld Research Inst, Mount Sinai Hospital, Toronto Canada

(2) Dalla Lana School of Public Health & Department of Statistics, University of Toronto, Canada

The phenomenon known as the Winner's Curse (WC) is a form of selection bias affecting genetic effect estimates that has recently received much attention. The proposed methods to overcome its effect mainly fall into two categories: bootstrap re-sampling techniques and MLE via conditional likelihood. In Genome-Wide Association Studies (GWAS), the source of the WC is two-fold: the association statistic at the top SNP(s) must reach genome-wide significance (threshold effect) and it must be the largest among all competing SNPs (maximization effect). The maximization effect however is largely overlooked by the likelihood approach in part due to the difficulty in specifying a correct joint likelihood for selected SNPs. Our purpose is to (1) evaluate the properties of the bootstrap and MLE approaches when each SNP is considered independently, i.e. not accounting for the maximization effect, and (2) compare bootstrap bias-reduction results obtained with and without maximization. Simulation studies show that under low power common to GWAS, the single-SNP bootstrap method produces estimates that are on average closer to the true genetic effect (lower MSE) than the likelihood methods. Application results indicate that accounting for the maximization effect by considering all SNPs jointly can further reduce the bias. Although computationally expensive, the bootstrap approach has the advantage of flexibility and can be readily modified to reflect different GWAS analysis strategies.

127

Mining Gold Dust under the Genome Wide Significance Level: A Two-Stage Approach

Gang Shi (1), Eric Boerwinkle (2), Alanna C. Morrison (3), Chi C. Gu (1), Aravinda Chakravarti (4), D.C. Rao (1)
(1) Washington University in St Louis
(2) The University of Texas Health Science Center at Houston
(3) The University of Texas School of Public Health at Houston
(4) The Johns Hopkins University

With dense single-nucleotide polymorphisms (SNPs) used in current genome-wide association (GWA) studies, large number of loci with small main effects may fail to pass stringent significance level that is set for controlling false positive rate (FPR). We propose a two-stage GWA approach: stage one controls false discovery rate (FDR); and stage two reduces FPR using least absolute shrinkage and selection operator (LASSO) regression. We simulated quantitative traits based on genome-wide SNP data in 8,861 individuals from Atherosclerosis Risk in Communities (ARIC) study. We found that simply lowering the FPR threshold could improve power but may inflate FDR significantly. The first stage of our proposed approach is targeted directly at controlling FDR and yields better power than using Bonferroni corrected genome-wide significance level. The trade-off between true discovery rate and FPR were also evaluated. In the second stage, selected SNPs were analyzed with LASSO regression, which shows two effects: it reduces redundant significant SNPs at causal loci due to linkage disequilibrium, and it reduces false positive SNPs. Interestingly, the LASSO regression preserves the power from stage one, i.e., the number of causal loci detected after LASSO is almost the same as in stage one, while reducing FPR further.

128

Models, Test Statistics, and Designs for Genetic Association Studies with Pooled Genotyping

Soo Yeon Cheong (1), Eleanor Feingold (2)
(1) Dept. of Biostatistics, University of Pittsburgh
(2) Dept. of Human Genetics, University of Pittsburgh

Although most genetic association studies use individual genotyping, it is possible that many studies can be performed more efficiently using pooled DNA, particularly at the initial screening stage. Pooled studies are challenging, however, because there are unresolved issues of how to incorporate pooling error into the designs and test statistics. We develop several models for the bias and variance introduced by pooling and then use those models to consider optimal case-control test statistics and designs. We consider several different case-control study designs with the same chip number for each cohort group, in order to find out which design is most powerful. In addition, we consider designs that incorporate covariate information. We derive optimal test statistics and compare them both theoretically and using simulation studies to the more conventional chi-squared test and *t*-test. We show that under most realistic pooling error models, the chi-squared test of allele frequencies performed as if there is no pooling is surprisingly close to the optimal test.

129

Practical Considerations for Imputation of Untyped Markers in Admixed Populations

Daniel Shriner (1), Adebowale Adeyemo (1), Guanjie Chen (1), Charles N. Rotimi (1)
(1) National Institutes of Health

Imputation of genotypes for markers untyped in a study sample has become a standard approach to increase genome coverage in genome-wide association studies at practically zero cost. Most methods for imputing missing genotypes extend previously described algorithms for inferring haplotype phase. These algorithms generally fall into three classes based on the underlying model for estimating the conditional distribution of haplotype frequencies: a cluster-based model, a multinomial model, or a population genetics-based model. We compared BEAGLE, PLINK, and MACH, representing the three classes of models, respectively, with specific attention to measures of imputation success and selection of the reference panel for an admixed study sample of African Americans. Based on analysis of chromosome 22 and after calibration to a fixed level of 90% concordance, MACH yielded the most imputed markers and the largest gain in coverage. The HapMap YRI reference panel alone outperformed the HapMap reference panels for: (1) ASW (African Americans from Southwest USA), (2) an unweighted combination of the CEU and YRI reference panels, and 3) a combination of the CEU and YRI reference

panels weighted by estimated admixture proportions. For our admixed study sample, we found that the optimal strategy involved imputing with the HapMap CEU and YRI reference panels separately and then merging the two sets.

130

A Generalized Sequential Bonferroni Procedure for Genome-wide Association Studies Incorporating Information on Hardy-Weinberg Disequilibrium Among Cases

Guimin Gao (1), Guolian Kang (1)
(1) University of Alabama at Birmingham

In genetic case-control studies, traditional tests for detecting association between a single marker and a disease (such as the standard allelic test and genotypic test) can be powerful for additive (ADD) and multiplicative (MUL) disease models, but are less efficient for dominant (DOM) and recessive (REC) models; testing for Hardy-Weinberg disequilibrium (HWD), which has also been used for detecting association, can be powerful for DOM and REC models, but has almost no power for ADD and MUL models. When these association tests for a single marker are applied to genome-wide association studies (GWAS) to test many single nucleotide polymorphisms (SNPs), the Bonferroni procedure is often used to control family-wise error rate (FWER), and therefore the power of GWAS can be very low in some situations. In this study, we adapt a generalized sequential Bonferroni (GSB) procedure to GWAS using the standard allelic association test. We calculate a weight for each marker by using the information of HWD among cases, and then assign a different significance threshold for each maker. Simulation studies show that in GWAS our adapted GSB procedure controls FWER well, and that the power of our method is always much higher under DOM and REC models than and is comparable under ADD and MUL models to the power of the traditional association tests with Bonferroni correction. Our method was applied to GWAS of a coronary artery disease dataset with about 340 K SNPs.

131

Family-Based Genome-wide Association Study of Inflammatory Markers: Individual Measures and PCA Phenotypes

Bhoom Suktitipat (1), Dhananjay Vaidya (2), Lisa R. Yanek (2), Taryn F. Moy (2), Lewis C. Becker (2), Diane M. Becker (2), M. Daniele Fallin (1)
(1) Johns Hopkins Bloomberg School of Public Health
(2) Johns Hopkins Medical Institutes

Vascular inflammation is a major mechanism leading to atherosclerosis.We performed a genome-wide association study for 3 inflammation markers hs-CRP, IL-6, and MCP-1, using each marker, as well as factor scores from principal component analysis (PCA).Samples consist of healthy members from the GeneStar families with early onset CAD, in Black and White.We performed 5 analyses using the Illumina Human1 M platform, with either log-transformed inflammatory marker levels or PC1 or PC2 as the trait.Effect size estimates and tests of significance were calculated using both GEE and mixed effects models to accommodate familial clustering, with adjustment for age, sex, and ethnicity (using EIGENSTRAT).Signals with P-value $<10^{-7}$ were further verified via the computationally intensive family-based association tests using MERLIN.The strongest GEE-derived signal,rs34868670,was seen for PC1, reflecting a latent trait weighted for CRP and IL-6 levels located in an intergenic region on chromosome 5.This SNP was significant in both white ($P = 2.44 \times 10^{-12}$) and black ($P = 2.11 \times 10^{-15}$) families and when using hs-CRP levels as the trait.However, it was not significant for IL-6 alone.SNPs in the myosin phosphatase target subunit 2 (MYPT2) also show strong association ($P < 1.31 \times 10^{-11}$) with PC1 and with CRP.In this presentation,we compare results from GEE and mixed effects model approaches in these data and results for combined analyses such as PCA traits versus individual serum markers.

132

Incorporation of Linkage Disequilibrium into Whole-Genome Association Studies via a Modification of Fused Lasso Regression

Samuel G. Younkin (1), J. Sunil Rao (1)
(1) Case Western Reserve University, Dept. of Epidemiology and Biostatistics

A central challenge faced by whole-genome association studies of complex disease is the inability to distinguish genetic variants with small effect size from statistical noise inherent in multiple testing. At present, large-scale association studies suffer from unwieldy false discovery rates. Without additional information about the nature of the disease or the genetic mechanisms underlying the disease, studies of this type may remain intractable. We investigate whether the incorporation of correlation due to linkage disequilibrium can decrease the false discovery rate to manageable levels. To do this we modify the fused lasso regression method of Tibshirani et'al. Fused lasso regression incorporates two constraints into standard ordinary least squares regression. The first, known as the lasso constraint, addresses sparsity and collinearity, characteristics clearly present in whole-genome association studies. The second, known as the fusion constraint, bounds total variation. Here, it is modified to constrain the difference in effect size estimates by the strength of linkage disequilibrium present among the two variants. We demonstrate the use of this method with data from a Mayo Clinic case-control association study of Alzheimer's disease using the Illumina 300 K HapMap Bead Array. Regression on this scale poses many computational difficulties, but preliminary results have consistently identified the well-known ApoE locus as well as a novel locus approximately 10 MB away.

133

OPCML Variants are Associated with Type 2 Diabetes and Metabolic Risk Factors in the Family Heart Study

Mary F. Feitosa (1), Ping An (1), Aldi Kraja (1), James S. Pankow (2), Michael A. Province (1), Ingrid B. Borecki (1)
(1) Division of Statistical Genomics, Washington University School of Medicine

(2) Division of Epidemiology and Community Health, University of Minnesota

We carried out genomewide association analysis on diabetes, defined as fasting glucose >125 mg/dL and/or taking hyperglycemic medications. We genotyped 962 Caucasians with the Illumina HumMap550 K chip and imputed up to ~2.5 million SNPs with reference to HapMap Release 22 CEU. Logistic regression with generalized estimating equations was used for associations testing for additive effects. Seven intronic-SNPs in complete linkage disequilibrium in OPCML (opioid binding protein/cell adhesion molecule-like, 11q25) were significantly associated with diabetes ($P < 1.75E-08$, MAF = 4%, OR = 0.2, CI:0.12?0.34). This is consistent with murine studies indicating that the protein binds opioid alkaloids in the presence of acidic lipids, and opioid alkaloids are involved in body weight change and glucose metabolism. We also observed a suggestive association between the SNP and HOMA ($P = 0.0005$), an indicator of insulin sensitivity. Furthermore, we tested whether OPCML was associated with obesity, dyslipidemia, and hypertension. We identified suggestive associations with total cholesterol ($P = 0.0007$), HDL ($P = 0.0035$), obesity (BMI ≥ 30 kg/m2, $P = 0.0093$) and MetS ($P = 0.0035$). Using a meta-analysis technique with a correction for correlated data, we combined these signals and obtained $P = 1.02E-13$ implicating this region. OPCML encodes a member of the immunoglobulin protein superfamily. Our findings indicate that OPCML is associated with diabetes and may affect other metabolic syndrome risk factors.

134

Investigating Aspects of Statistical Power in Meta-Analysis of Complex Traits

Gemma Cadby (1), Kim W Carter (2), Steven Wiltshire (3), Lyle J Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics
(2) Telethon Institute for Child Health Research
(3) Western Australian Institute for Medical Research

Many reported associations between genetic variants and complex traits have not been consistently replicated. This may be due to false negative studies; that is, underpowered studies unable to detect associations, even when the variant being tested is causative.
Meta-analysis may assist in solving the problem of underpowered studies. The sample size required for 80% power increases steeply as between-study heterogeneity increases when minor allele frequencies (MAF) are held constant. However, studies from different populations combined in meta-analyses often have different MAF which may affect power.
We simulated datasets to investigate the statistical power associated with combining studies in meta-analyses with different MAF. We also estimated the power associated with combining different numbers of studies with different sample sizes, MAF, effect sizes and disease models.
We found that, when combining studies with different MAF, there was no difference in power compared to combining studies with the same MAF. We also found that for an additive model with Odds Ratios less than 1.5, 40% of parameter combinations yielded meta-analyses which did not reach 80% power.
There may be some situations where meta-analysis may not replicate true associations. Realistic statistical power should be considered when planning studies and adequate numbers of subjects should be collected, particularly when investigating variants with modest effects.

135

Does it Matter Which Genotype Calling Algorithm for Affy SNP 6.0 is Used for Genome Wide Association Studies?

Mariza de Andrade (1), ELizabeth Atkinson (1), William Bamlet (1), Sooraj Maharjan (1), Martha Matsumoto (1), Sharon Kardia (2)
(1) Mayo Clinic College of Medicine
(2) University of Michigan

In this era of genome wide association studies (GWAs) hundred of thousands of single nucleotide polymorphisms are having their genotype called using different platforms and algorithms. However, there has been no investigation focusing on the impact these algorithms might have on the results of a genome wide association study. We will focus our work on the Affymetrix platform. Currently there are two genotype calling algorithms (GCA) available for the Affymetrix SNP Array 6.0: Birdseed and CRLMM. These two algorithms have different recommended workflows including SNP and sample quality control selection criteria. We have evaluated these two GCAs and come up with some best practice guidelines. For instance, Affymetrix suggests filtering samples based on contrast quality control (CQC) values prior to run Birdseed, using only samples with CQC above 0.4. On the other hand, CRLMM uses the sample quality score (SNR) and recommends that samples should be flagged with SNR values less than 5. Based on our own experience we conclude that it is important to use one of these screening algorithms and that the concordance of the resulting genotypes from the two GCA was similar. Additionally, we used results from both GCA in our GWA and found that the GCA does not have a major impact on the study results. We will present our results using 854 samples of hypertensive sibships from Rochester, MN.

136

Why So Often We Find Disease Associations in MHC Region?

Natsuhiko Kumasaka (1), Naoyuki Kamatani (1), Ryo Yamada (2)
(1) CGM, RIKEN
(2) IMS, The University of Tokyo

The major histocompatibility complex (MHC) is the most gene-dense region in human genome with highly strong genetic heterogeneity and linkage disequilibrium (LD). Recent genome-wide association studies (GWAS) have identified signals for various diseases in this region more frequently than the other regions. Although the density of genes and their polymorphisms and their non-neutral origins would explain the accumulation of positive reports in MHC region, we suspected the specific structure of LD

in the region also statistically affected on the frequent identifications. We parameterized LD blocks with the number of markers and the strength of LD, and estimated the power to detect a disease association in each block accoding to the typical case-control GWAS settings. We observed that the strength of LD as well as the number of markers in the block essentially increases the power of association, but it was not in monotonic fashion. We also evaluated the SNP list in the current GWAS platforms for this phenomenon and identified the power of a particular risk locus in MHC region could be 1.5 times more than the other regions in the genome.

137

A Comparison of Reference Panels for Imputation of Genotype Data in Genome-wide Association Studies

Andrew Morris (1), Denise Brocklebank (1), Carl Anderson (1)
(1) Wellcome Trust Centre for Human Genetics

Genome-wide association (GWA) studies have been successful in identifying common variants with modest effects on complex human traits. Nevertheless, the power of such studies can be increased by imputing ''unobserved'' SNPs which are not present on GWA genotyping products, but have been typed on high-density reference panels of phased haplotype data. For GWA studies in European populations, the most commonly utilised reference panels are those generated by the international HapMap project. Phase II of the project (HMP2) typed more than 3 million SNPs in 30 trios with Northern European ancestry, whilst phase III (HMP3) incorporated an additional 30 trios, but were typed at only 1.6 million SNPs. However, we soon expect whole-genome re-sequencing data to be released for the same 60 trios as part of the 1000 Genomes project (1000G).
We have performed a simulation study to: (i) compare the performance, in terms of power and localisation, of imputation in GWA studies from HMP2 and HMP3; and (ii) evaluate the improvement we might expect through imputation from 1000G. Our results highlight that: (i) HMP3 is more powerful than HMP2 for detecting causal variants with MAF of at least 5%; (ii) HMP2 may offer some advantages over HMP3 for rarer causal variants, although power is low for both; and (iii) both HapMap reference panels fall short of 1000G, except for common causal variants which are already well covered by HMP2 and HMP3.

138

Model Selection Strategies in Genome-Wide Association Studies

Sarah L. Keildson (1), Martin Farrall (2), Andrew P. Morris (1)
(1) Wellcome Trust Centre for Human Genetics
(2) Department of Cardiovascular Medicine, University of Oxford

Unravelling the complex genetic architecture of common diseases is a continual challenge in human genetics. While genome-wide association studies have been successful in identifying novel disease-susceptibility loci, the extension of these studies beyond the scope of single SNP analyses have been limited. Multi-locus methods of analysis may, however, have the potential to increase the power of these studies to detect genes of smaller effect size as well as genes that interact with each other and the environment. We have carried out large scale simulations of four multi-locus model selection techniques, namely forward selection, backward selection, Lasso and Bayesian model averaging, in order to compare the type I error rate and power of each method across a range of genetic models. To decrease the type I error rate to approximately 5%, adjustments were made to each method and the power at these error rates was calculated over 1000 simulated data sets. At a type I error of 0.053, Lasso generally showed the highest power over various genetic models, followed by forward and backward selection, both of which had estimated type I error rates of 0.058. While Bayesian model averaging proved to be the least powerful technique, it also had the lowest frequency of false positive results (0.038). Forward selection and a lasso/bayesian model averaging combination was then applied to real lipoprotein(a) data and yielded results consistent with those published in the literature.

139

Genome-Wide Association Analysis Reveals That PTPRD (Protein Tyrosine Phosphatase Receptor Type Delta) Is Associated with Smoking in Non-Drinkers

Chuanhui Dong (1), Ma-Li Wong (1), Julio Licinio (1)
(1) University of Miami Miller School of Medicine

Smoking poses a major problem for public health and constitutes the second leading cause of death in the world. Studies have suggested that the smoking behaviors have a substantial hereditary component with a heritability in the range of 37–59%. To identify genetic loci for smoking and control the potential bias due to alcohol drinking behavior, we performed a secondary analysis of genome-wide association data from the Genetic Association Information Network (GAIN) major depression sample using smoking status as phenotype and limiting the subjects to those who had no history of alcohol drinking ($N = 959$). Allelic association analysis identifies 4 single nucleotide polymorphisms (SNP) within a *PTPRD* (Protein Tyrosine Phosphatase Receptor Type Delta) region from 10363600 to 10417955 bps that are associated with smoking status with an allelic OR ≥ 2.3 and nominal $P < 5.0 \times 10^{-7}$, including rs10959102 (OR = 2.5, 95% CI = 1.8–3.5, $P = 5.8 \times 10^{-8}$), rs12344987 (OR = 2.4, 95% CI = 1.7–3.4, $P = 8.9 \times 10^{-8}$), rs45454697 (OR = 2.4, 95% CI = 1.7–3.3, $P = 2.4 \times 10^{-7}$) and rs10363600 (OR = 2.3, 95% CI = 1.6–3.1, $P = 4.6 \times 10^{-7}$). All four SNPs are in Hardy-Weinberg Equilibrium with a $P > 0.7$ in non-smokers. Our results suggest that *PTPRD* genetic variants may be associated with susceptibility to smoking behaviors among non-drinkers. More independent samples are needed to validate the observed association.

140

Genome–wide Scan of Genetic Variants Associated with DNA Copy Number Aberrations in Lung Cancer

Qunyuan Zhang (1), Li Ding (1), Ling Lin (1), Ingrid Borecki (1), Michael A Province (1)
(1) Washington University School of Medicine

Although DNA polymorphisms in normal cells and genomic abnormalities in tumor cells have been recognized as two major classes of genetic factors contributing to cancers, studies on the connection between the two phenomena are very limited. In this study, we investigate the association between genotypes in normal cells and DNA Copy Number Aberrations (CNA) in tumor cells using the Affymetrix Human Mapping 250 K array data of normal and tumor tissue matched samples from 357 Lung Cancer (Adenocarcinoma) patients. First, we identified 6 recurrent CNA regions, 7p11.2, 8q24, 11q13.3, 12q15, 12p12 and 14q13.3, using a novel approach, correlation matrix diagonal segmentation (CMDS). Then we conducted a genome-wide association scan, identifying a total of 121 SNPs (FDR < 0.05 and MAF > 0.03) associated with the CNA status of these regions. These SNPs anchor 152 genes, most of them from the KEGG pathways of Signaling Molecules and Interaction, Signal Transduction and Cell Communication. Important candidate genes are MUM1, KIAA1505, IL1A, IL1B, PXMP3, UTX, FGFR2, HIVEP1, PDGFD, PPP1R3F, and RREB1 from pathways of Nucleotide Metabolism, Replication and Repair, Cell Growth and Death, and Cancers.

141

A Genome-Wide Association Study Identifies Variation near SORCS1 as a Major Locus for Glycemic Control in Type 1 Diabetes, as Measured by Both HbA1c and Glucose

Andrew Paterson (1), Daryl Waggott (2), Andrew Boright (3), Mohsen Hosseini (1), Enqing Shen (2), Marie-Pierre Sylvestre (2), Isidro Wong (1), Bhupinder Bharaj (1), Patricia Cleary (4), John Lachin (4), Angelo Canty (5), Lei Sun (6), Shelley Bull (2), DCCT/EDIC Research Group (4)
(1) Hosp for Sick Children
(2) SLRI
(3) UHN
(4) GWU
(5) McMaster U
(6) UofT

Glycemia is a major risk factor for long-term complications in type 1 diabetes (T1D), and is heritable, but no specific genetic loci have been identified in T1D. To identify genetic loci influencing glycemic control in T1D, we performed a GWAS with longitudinal repeated measures of HbA1c from the 1304 white individuals with T1D from the Diabetes Control and Complications Trial (DCCT) separately by treatment group. We identified a major locus associated with HbA1c levels near the SORCS1 gene (rs1358030, $P < 10^{-9}$). Evidence confirming the association at this SNP was obtained from the intensive treatment group ($P = 0.01$) in the same direction. This SNP was also associated with the mean of a one-day seven-point glucose measures obtained quarterly ($P < 10^{-4}$). An additional locus was close to genome-wide significance rs10810632 ($P < 10^{-7}$). rs1358030 and rs10810632 were also associated with the risk of hypoglycemia ($P < 10-^{3}$), an important complication of diabetes. We identify that a major locus for HbA1c and glucose is near the SORCS1 gene. The influence of genetic factors on an individual's ability to control their HbA1c levels needs to be taken into account in the design and analysis of genetic studies attempting to identify risk factors for long-term diabetic complications.

142

Tiled regression: the use of regression methods in hotspot defined genomic segments to identify independent genetic variants responsible for variation in quantitative traits

Alexander F. Wilson (1), Yoonhee Kim (1), Heejong Sung (1), Juanliang Cai (1), Francis J. McMahon (2), Alexa J.M. Sorant (1)
(1) Genometrics Section, Inherited Disease Research Branch, NIH/NHGRI
(2) Genetic Basis of Mood and Anxiety Disorders, NIH/NIMH

Rather than focus on identifying markers that best characterize Linkage Disequilibrium (LD) blocks, stepwise regression is used to identify independent markers that are responsible for additive effects on a quantitative trait. The genome is first divided into segments based on ''hotspot'' regions. The term ''tile'' denotes both the sequence of DNA between two hotspot regions and the hotspot region itself. A tile may include multiple markers in one or more LD blocks. A test of the overall multiple regression model is performed on the markers in each tile in order to determine if any marker in the tile makes a significant contribution to the overall regression. Stepwise regression is then used to select the independent markers in each tile. If the multi-colinearity between markers is high, stepwise regression can be used from the outset. Thereafter, the significant markers are combined in higher order stepwise regressions. The tiled regression framework can be extended to include qualitative traits with stepwise logistic regression, and to family data that can be represented in a linear regression context (e.g. ROMP or GEE). The tiled regression method for quantitative traits is illustrated with data from the STAR*D study [Rush et'al., 2004]. The method classified 705 SNPs from 68 genes into 197 tiles. Results were similar to those reported by McMahon et'al. [2006], with the most significant marker being rs7997012.

143

Analyses Conditional on Established Type 2 Diabetes (T2D) Loci Reveal Putative Novel Associations

Teresa Ferreira (1), Eleftheria Zeggini (2), Andrew P. Morris (3)
(1) Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom
(2) Wellcome Trust Sanger Institute, Hinxton, United Kingdom
(3) Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom

Despite recent successes of genome-wide association studies, much of the genetic contribution to the variation in T2D risk remains unexplained. One way to potentially increase the power to identify further novel susceptibility variants is to perform conditional analyses, accounting for genotypes at leading SNPs at established risk loci as covariates in a logistic regression framework. We consider

two specific paradigms: (i) model the linear trend in the main effect of a SNP, conditional on the linear trend of the main effect of the leading SNP at a known risk locus; and (ii) the same model, but incorporating an interaction between the two SNPs.
These tests were applied, genome-wide, to ~2000 cases of T2D and ~3000 controls from the Wellcome Trust Case Control Consortium, conditioning on genotypes at each known risk locus. When searching for main effects only, conditional on all 19 established loci, the strongest novel signals of association were observed on chromosomes 8p21 ($P = 3.4 \times 10^{-6}$; unadjusted $P = 8.9 \times 10^{-5}$) and 4q31 ($P = 5.7 \times 10^{-6}$; unadjusted $P = 6.9 \times 10^{-5}$). When allowing for interaction effects, considering each established locus in turn, the strongest signals of association were observed on chromosomes 5p15 by conditioning on the lead SNP at the *FTO* locus ($P = 2.3 \times 10^{-7}$; unadjusted $P = 2.5 \times 10^{-2}$), and 11q14 by conditioning on the lead SNP at the *IGF2BP2* locus ($P = 4.0 \times 10^{-6}$; unadjusted $P = 2.5 \times 10^{-3}$). These signals are currently under investigation in samples from independent T2D cohorts.

144
Power of Genotype Similarity Analyses Under Scenarios of Allelic Heterogeneity
Kathryn E. McDougal (1), M. Daniele Fallin (1)
(1) Johns Hopkins/Bloomberg School of Public Health

One of the limits to genetic association studies often cited is the potential for allelic heterogeneity, such that multiple variants in the same gene or region are separately responsible for some proportion of cases in any particular study. This makes detection of the association to any one variant difficult. Genetic similarity analysis methods exploit the similarity of allelic profiles in cases versus controls. Thus, they may be better able to address allelic heterogeneity than conventional association methods. However, the effect of allelic heterogeneity on the power of these methods has not been fully characterized. We have performed simulation studies assuming either 1, 2 or 3 separate susceptibility variants in the same gene region, under scenarios of tight, moderate, and weak LD between the variants. We explore varying genotype models, effect sizes, and disease frequencies. We show power and type 1 error comparisons between genotype similarity analyses using the method described previously in *Tzeng et al.*1 and conventional Armitage trend tests for each SNP in the region to evaluate the performance of similarity statistics in situations of allelic heterogeneity.
1 *Am. J. Hum. Genet.* **72** (2003): 891–902.

145
Genome-wide Significant Confirmation of SNPs in SNCA and The MAPT Region as Common Risk Factors for Parkinson Disease
Todd L. Edwards (1), William K. Scott (1), Cherylyn Almonte (1), Amber Burt (1), Eric H. Powell (1), Gary Beecham (1), Liyong Wang (1), Stephan Zuchner (1), Ioanna Konidari (1), Gaofeng Wang (1), Margaret A. Pericak-Vance (1), Jonathan Haines (2), Jeffery Vance (1), Eden R. Martin (1)
(1) Miami Institute for Human Genomics, Miller School of Medicine, University of Miami
(2) Center for Human Genetics Research, Vanderbilt University Medical Center, Vanderbilt University

Parkinson disease (PD) is a chronic neurodegenerative disorder with a cumulative prevalence of greater than one per thousand. While several rare Mendelian forms of PD have been described, the genetic factors for idiopathic PD have been elusive. Three independent genome-wide association studies (GWAS) have investigated the genetic susceptibility to PD using various study designs and genotyping platforms. These studies implicated several genes as PD risk loci with strong, but not genome-wide significant associations. In the current study, we imputed data for joint analysis from two previously published GWAS from dbGAP with our new GWAS with 605 cases and 621 controls. Genotyped SNPs in *SNCA* (rs2736990 genome-wide $P = 0.0109$, OR = 1.29 95%CI 1.17–1.42 PAR% = 12%) and the *MAPT* region (rs11012 genome-wide $P = 0.008$, OR = 1.42 95%CI 1.25–1.61 PAR% = 8%) were statistically significant at the genome-wide level. No other SNPs were genome-wide significant in this analysis, though several biologically relevant genes (*RORA*, *NPAS3*, *WIPF1*, *DBC1*, *GFPT2*) replicated in at least two of three data sets at the 0.05 level with consistent effects across samples. These genes are being further investigated in a fourth PD dataset. This study confirms that *SNCA* and the *MAPT* region are major genes influencing risk for most PD patients as they have been consistently observed to significantly associate with PD here and in other studies, and these exposures combined explain at most 20% of PD cases.

146
PARK2 and SVOPL Loci are Associated with Successful Aging in the Amish
Digna R. Velez (1), John R. Gilbert (1), Jamie L. Myers (1), Lan Jiang (2), Anna C. Davis (2), Paul J. Gallins (1), Ioanna Konidari (1), Laura Caywood (1), Marilyn Creason (1), Denise Fuzzel (1), Clara Knebusch (2), Renee Laux (2), Michael L. Slifer (1), Charles E. Jackson (3), Margaret A. Pericak-Vance (1), Jonathan L. Haines (2), William K. Scott (1)
(1) Miami Institute for Human Genomics, University of Miami
(2) Center for Human Genetic Research, Vanderbilt University
(3) Scott & White, Temple, TX

Successful aging (SA) is defined as living to older age with high physical function, preserved cognition, and continued social engagement. Several domains underlying SA have evidence of heritability: longevity, grip strength (GS), lower extremity function, and retention of cognitive ability. We examined 263 Amish individuals age ≥ 80 (74 SA cases and 189 controls), part of a single 11-generation pedigree, using 630,439 SNPs from an Affymetrix Genome-Wide Human SNP Array 6.0. GS was examined as marker of musculoskeletal function. MQLS was used to analyze SA and for GS the GRAMMAR-GC method implemented in GenABEL was used; both adjust for pairwise kinships among related individuals. Chromosome 6q25–q27 (including *FRA6E* fragile site and *PARK2*) contained several SNPs

associated with SA (min $P = 2 \times 10^{-6}$) and analyses of GS produced a genome-wide significant result ($P = 1 \times 10^{-7}$) in the *SVOPL* gene on chromosome 7q34. The associations of SNPs of in *PARK2* are of note because *PARK2* has been associated with several aging-related phenotypes including Parkinson's disease, general neuron degeneration, and several forms of cancer. Although the role of *SVOPL* in aging is unknown, the adjacent gene *TRIM24* has been associated with several cancers. Both *PARK2* and *TRIM24* have been described as tumor suppressor genes, consistent with the theory that preserved DNA repair and tumor suppression activity are essential mechanisms for cancer-free longevity and SA.

147

Evaluation of Imputation Strategies for Family Data

Yun J. Sung (1), Yanan Duan (1), Treva K. Rice (1), Tuomo Rankinen (2), Claude Bourchard (2), C. C. Gu (1)
(1) Washington University
(2) Pennington Biomedical Research Center

Imputation methods to infer missing or untyped SNPs using HapMap data as a reference have been used successfully to improve power in association studies, to facilitate meta analyses, and to replicate significant findings in follow-up studies. However, currently available imputation methods are designed for population data and not suited for family data. In the current study, we utilize three well-known programs for genotype imputation (Mach, BimBam and IMPUTE) to infer missing and non-typed markers for the HERITAGE family study that consisted of 99 Caucasian nuclear families (479 individuals) using HapMap CEU phased data. We evaluate several imputation strategies for family data. Association analyses are performed using the original data without imputations and also using the expanded data under each of the three methods of imputation. A measure of comparison between three imputation approaches is developed, the consistency and robustness of the results among the three methods is evaluated, and the consistency is compared with the original analyses of non-imputed data.

148

A Whole-genome Simulator Capable of Modeling High-order Epistasis for GWAS Studies of Complex Disease

Wei Yang (1), Chi C. Gu (2)
(1) Division of Biostatistics, Washington University in St Louis
(2) Division of Biostatistics, Department of Genetics, Washington University in St Louis

For genome-wide association (GWA) studies of complex diseases, synergetic effect of multiple risk loci is an important factor to consider in statistical analysis. However, published GWAS analyses rely almost exclusively on single-marker scans, because high-order epistasis (gene-gene interactions) is poorly understood. Therefore, a GWAS data simulator becomes essential to model complex interactions, and to simulate realistic linkage disequilibrium (LD) structure and huge amount of data in practical time. We have developed previously a novel method to characterize multilocus penetrance that allows for specifying high-order epistasis conformal to marginal penetrance constraints. In the current study, we combine it with an existing program (GWAsimulator) to achieve rapid whole-genome simulation with accurate interaction modeling. We considered various approaches to specify interaction models, including (1) departure from product marginal effects for pair-wise interactions, (2) logistic regression models for low-order interactions, and (3) penetrance tables generated conforming to marginal effect constraints for high-order interactions. The new program, called simGWA, is capable to generate large GWAS data efficiently and with high precision. We will present performance assessment of simGWA to verify that the simulated data are faithful to assigned genome-wide LD structure, and conform to pre-specified diseases models with (or without) interactions.

149

Whole-Genome Detection of Disease-Associated Deletions in Case-Control Studies of Rheumatoid Arthritis

Chih-Chieh Wu (1), Sanjay Shete (1), Eun-Ji Jo (1), Wei V. Chen (1), Annette T. Lee (2), Peter K. Gregersen (2), Christopher I. Amos (1)
(1) M. D. Anderson Cancer Center
(2) North Shore-Feinstein Medical Research Institute

Recent genetic studies have increasingly shown that interstitial deletions are common in patients with cancers and psychiatric disorders, suggesting genomic deletions play an important role in the genetic basis of complex traits. Whole-genome studies of deletions have been performed extensively over the past few years. Many focus on investigating genetic variations in non-diseased individuals. However, the association between deletions and complex diseases has not yet been systematically investigated in genetic mapping studies. Here, we propose to perform whole-genome detection in rheumatoid arthritis (RA) using 2 recently developed statistical methods. Our methods have been shown to be useful and robust in the presence of linkage disequilibrium. The 1st method was designed for SNP-by-SNP analyses and the 2nd for cluster analyses based on combined evidence from multiple SNPs. We performed whole-genome detection of deletions on 550 K SNP data in 868 RA patients and 1197 controls from the North American Rheumatoid Arthritis Consortium. In SNP-by-SNP analyses, 65 significant SNPs overly aggregated within 31–34 Mb on chromosome 6p in the HLA region at $\alpha = 10(-8)$. Most notably, we identified deletions that encompassed HLA-DRB1 with P-value $< 10(-15)$ in which deletions were previously discovered in RA patients. PTPN22 and TRAF1-C5 are another 2 known susceptibility variants of RA. Our analysis detected neither deletions over the regions that encompass PTPN22 and TRAF1-C5.

150

Analysis of Two Gene-centric Approaches for Genome-Wide Association Studies

Guolian Kang (1), Bo Jiang (2)
(1) Section on Statistical Genetics, Department of Biostatistics, The University of Alabama at Birmingham
(2) Department of Biostatistics, The University of Alabama at Birmingham

To test whether a gene with multiple single nucleotide polymorphisms (SNPs) is associated with a disease, the simplest method is to test each single SNP and then to use Bonferroni correction, and however this method is often conservative. Joint analysis of multiple SNPs within one gene can have a high power because correlation or linkage disequilibrium (LD) among SNPs within one gene is considered. Several joint analysis methods have been developed and showed improved power compared to the simplest method. These methods include a Monte Carlo (MC) method and an entropy-based method. The objective of this study is to compare the performance of these two methods as well as the simplest method and to discuss the strengths and weaknesses of these methods. Simulation studies show that (1) Both the simplest method and the MC method can control type one error rate quite well; (2) the entropy-based method can also control type one error rate quite well when there is moderate and large LD among SNPs within one gene and it has an inflated type one error rate when there is no LD among SNPs within the gene; (3) the MC method has higher power than the entropy-based method when there is one disease SNP within the gene; (4) the entropy-based method has higher power than the MC method when there is two, three or more disease SNPs within the gene.

151
Consequences of Correlation Due to Shared Control Design in GWAS.
Dmitri V. Zaykin (1), Damian O. Kozbur (1)
(1) National Institute of Environmental Health Sciences

An appealing GWAS design is where a large control group is tested against several disease samples. Such design is efficient, since the control group can be reused multiple times. Here I focus on a statistical issue of correlation between association test results due to the usage of shared controls. For the usual allelic trend test at a SNP with sample sizes of N0, N1, N2 for the control and two disease samples, the correlation between association test statistics for the two diseases is found to be $1/(1+N0/N1)*1/(1+N0/N2)$. This correlation holds even in the absence of association with either of the two diseases. Further, the distribution of *P*-values for an association test with disease 1 is obtained given that a *P*-value for the disease 2 is smaller than a particular threshold value. This distribution is used to correct association *P*-values for the correlation. The findings are illustrated using shared control data from the Wellcome Trust Case Control Consortium GWAS [Nature, 2007].

152
Single-Marker and Two-Marker Association Tests for Unphased Case-Control Genotype Data, with a Power Comparison
Sulgi Kim (1), Nathan Morris (1), Sungho Won (2), Robert Elston (1)
(1) Case Western Reserve University
(2) Havard University

In case-control SNP data, the Allele frequency, Hardy Weinberg Disequilibrium (HWD) and Linkage Disequilibrium (LD) contrast tests are three distinct sources of information about genetic association. While all three tests are typically developed in a retrospective context, we show that prospective logistic regression models may be developed that correspond conceptually to the retrospective tests. This approach provides a flexible framework for conducting a systematic series of association analyses using unphased genotype data and any number of covariates. Two single-marker tests and four two-marker tests are discussed. The true association models are derived and they allow us to understand why a model with only a linear term will generally fit well for a SNP in weak LD with a causal SNP, whatever the disease model, but not for a SNP in high LD with a non-additive disease SNP. We investigate the power of the association tests using real LD parameters from chromosome 11 in the HapMap CEU data. Among the single-marker tests, the allelic test has on average the most power in the case of an additive disease; but, for non-additive diseases, the genotypic test has the most power. Among the two-marker tests, the Allelic-LD contrast test, which incorporates linear terms for two markers and their interaction term, provides the most reliable power overall. Therefore, our result supports incorporating an interaction term as well as linear terms in multi-marker tests.

153
Identification of Loci Influencing Age-At-Onset in Late-Onset Alzheimer Disease Implicates Variation on Chromosome 12
Adam C. Naj (1), Gary W. Beecham (1), Eden R. Martin (1), Michael A. Slifer (1), Eric H. Powell (1), Paul J. Gallins (1), Ioanna Konidari (1), Patrice Whitehead (1), John R. Gilbert (1), Jonathan L. Haines (2), Margaret A. Pericak-Vance (1)
(1) Institute for Human Genomics, University of Miami
(2) Center for Human Genetics Research, Vanderbilt University

Alzheimer Disease (AD) is the leading cause of dementia among the elderly and is highly genetic ($H2\sim70\%$). AD susceptibility loci may contribute to earlier age-at-onset (AAO). To identify risk loci for late-onset AD (LOAD), we performed a genome-wide association study (GWAS) of AAO of AD and risk for LOAD. We analyzed data on 1,169,331 SNPs on 1,474 cases combining three GWAS datasets, imputing non-overlapping SNPs. We tested SNP associations with AAO among AD cases, following up on associations of $P<10^{-5}$ by testing association with LOAD risk among cases and 1,331 controls. 26 SNPs on chromosome 12 near 33.2 Mb were associated with AAO at $P<10^{-5}$, the strongest rs10047666 ($P=1.40\times10^{-6}$), though none demonstrated statistically significant association ($P<0.05$) with LOAD risk. These signals fell ~200 kb upstream of synaptotagmin X (*SYT10*), encoding a presynaptic protein that may contribute to spatial memory. Additionally, two independent SNPs associated with AAO (rs7970175, 12q14.1, $P=7.65\times10^{-6}$; rs10944728, 6q16.1, $P=7.73\times10^{-6}$) also showed significant or strong associations with LOAD risk (rs7970175, $P=1.01\times10^{-9}$; rs10944728, $P=1.23\times10^{-5}$). Combining three LOAD GWASs, we observed multiple novel associations on chromosome 12 with AAO

of AD, and associations of chromosome 6 and 12 SNPs with both AAO and LOAD risk, however potential roles of these SNPs remain uncertain and merit further investigation.

154
Simulation of large-scale SNP data for complex pedigrees using GenomeSIMLA
William S. Bush (1), Anna C. Davis (1), Mary F. Davis (1), Eric S. Torstenson (1), Jonathan L. Haines (1)
(1) Vanderbilt University

The cost-efficiency of large-scale genotyping has prompted its use in population isolates with complex pedigrees. Genome-wide association studies (GWAS) are now a popular approach for case/control and family-based designs, however even in these data sets, the statistical power and false-positive rate of even basic statistics are often unclear. The large number of correlated SNPs complicates multiple testing issues, and further complicates the challenge of analyzing complex pedigrees. As such, simulation studies can be invaluable tools for evaluating statistical methods, but efficiently simulating novel data of GWAS scale is problematic. GenomeSIMLA is a forward-time population-based simulation package for generating large-scale SNP data with realistic patterns of linkage disequilibrium in both case/control and family-based designs. Complex disease models can be applied using multiple approaches, and extensive descriptive statistics and plots are provided to describe the simulation output. To aid the development and evaluation of methods for analyzing complex pedigrees, we have adapted GenomeSIMLA to efficiently produce GWAS data for pedigree structures. In this work, we highlight the new features of genomeSIMLA software and present preliminary simulations of an Old Order Amish community including empirical distributions of allele sharing and homozygosity, and an exploration of feasible genetic models based on the pedigree structure.

155
Latent Class Analysis and GWA in the GAIN BP Data
Berit Kerner (1), Judy Y. Kong (1), Bengt O. Muthen (1)
(1) University of California, Los Angeles

In order to explore phenotype heterogeneity in the Genetic Association Information Network (GAIN) BP sample (1058 individuals) we used a latent class approach, in which we included co-morbid substance abuse/dependence (SUB), obsessive compulsive disorder (OCD), panic disorder (PD), social phobia (SP), eating disorder (ED), attention deficit hyperactivity disorder (ADHD), alcohol abuse (ALCA), alcohol dependence (ALCD), nicotine dependence (NIC), and psychotic symptoms (PSYC) as variables. The analysis was performed in the statistical software program Mplus. A three class solution fit the data best based on Bayesian Information Criterion (BIC). Class 1 (28.3%) was characterized by a 40% probability of endorsing SUB, PD, ALCA and PSYC. Class 2 (24.4%) was characterized by a 100% probability of endorsing ALCD and an 80% probability of endorsing SUB. Class 3 (47.2%) had a very low probability of endorsing co-morbid conditions overall. We then used the latent class membership probability as phenotype in genome-wide genetic analyses using the software program GOLDEN HELIX. 1000 individuals with BP and 1034 controls had been genotyped on the Affymetrix 6.0 chip. We found the most significant associations between Class 1 membership and SNP rs13220542 on chromosome 6q15 ($P = 5.3 \times 10^{-8}$) and between Class 2 membership and SNP rs7528071 on chromosome 1p12 ($P = 1.2 \times 10^{-9}$). Latent class analysis might be a useful approach to phenotypic heterogeneity in genetic analyses.

156
A Common Variant on Chromosome 11q13 is Associated with Atopic Dermatitis
Jorge Esparza Gordillo (1), Stephan Weidinger (2), Regina Regina Fölster-Holst (3), Anja Bauerfeind (1), Franz Ruschendorf (1), Klaus Rohde (1), Ingo Marenholz (1), Florian Schulz (1), Tamara Kerscher (1), Hubner Hubner (1), Stefan Schreiber (4), Andre Franke (4), Simon Heath (5), Natalija Novak (6), Elke Rodriguez (2), Thomas Illig (7), Min-Ae Lee-Kirsch (8), Milan Macek (9), Andreas Ruether (4), Young-Ae Lee (1)
(1) MAX-DELBRÜCK-CENTER FOR MOLECULAR MEDICINE (MDC)
(2) Technische Universität München, Germany
(3) University Hospital Schleswig-Holstein, Kiel, Germany
(4) Christian-Albrechts-University, Kiel, Germany
(5) Centre National de Génotypage, Evry, France
(6) University of Bonn, Bonn, Germany
(7) Helmholtz Zentrum Munich-German
(8) Technical University Dresden, Dresden, Germany
(9) Charles University Prague

Atopic dermatitis (AD) is a chronic inflammatory skin disorder with complex etiology. We conducted a genome-wide association study in 939 individuals with AD and 975 controls as well as 270 complete nuclear families with AD. SNPs consistently associated with AD in both discovery sets were investigated in two additional replication sets totalling 2637 cases and 3957 controls. Highly significant association was found with a common sequence variant on chromosome 11q13.5 ($P = 7.6 \times 10^{-10}$). 13% of individuals of European origin are homozygous for the risk allele, and their risk of developing AD is 1.47 times that of noncarriers. Notably, the same AD risk allele reported here has recently been identified as a susceptibility factor for Crohns disease, a complex chronic inflammatory bowel disorder sharing many pathophysiological characteristics with AD. Our data suggest that rs7927894[A] confers susceptibility to AD and Crohns disease jointly which may contribute to the higher incidence of AD observed among Crohns patients. rs7927894 is located in a 200 kb LD block containing the C11orf30 gene, which encodes the nuclear protein EMSY implicated in chromatin modification, DNA repair, and transcriptional regulation. Finally, we provide a list of additional candidate genes. Further replication in independent cohorts, fine mapping and functional studies will be required to gain a better understanding of the physiological mechanisms underlying this common allergic disorder.

157

Using Prior Information Attained from the Literature to Improve Ranking in Genome-wide Association Studies

Mattias Johansson (1), Yaoyong Li (2), Jon Wakefield (3), Mark Greenwood (2), Thomas Heitz (2), Ian Roberts (2), Hamish Cunningham (2), Paul Brennan (1), Agnus Roberts (2), James Mckay (1)
(1) International Agency for Research on Cancer (IARC)
(2) Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
(3) Departments of Statistics and Biostatistics, University of Washington, Seattle

Advances in high-throughput genotyping have made it feasible to conduct genome-wide association studies (GWAS) aiming to investigate the majority of common genetic variation and relate it to some phenotypic differences, often to risk of some disease. Whilst the price of GWAS assays are decreasing rapidly, conducting a GWAS is still a very expensive exercise, typically requiring genotyping several thousands of subjects to gain sufficient statistical power to distinguish the true association signals from the background noise.

Recognizing that a large proportion of GWAS findings reside near potential candidate genes for many of the investigated phenotypes, we here explore means to incorporate prior information attained from the literature to improve ranking in GWAS. We use this information to assign a crude prior probability of association for each SNP. The prior probabilities are thereafter integrated with the association result from the GWAS and the SNPs are re-ranked according to Baysian false-discovery probability (BFDP). We show that this methodology improves the ranking for many known susceptibility loci with examples from studies on lung cancer and cancer of the upper aero digestive tract (UADT). We have implemented this methodology in a web application where a user can specify a list of keywords and receive priors for all SNPs of interest. These priors can thereafter be used to rank the SNPs according to the BFDP.

158

Imputation Quality for Hispanics Using Different Reference Populations

Jinghua Liu (1), James Baurley (1), Frank Gilliland (1), Jim Gauderman (1), David Conti (1)
(1) Department of Preventive Medicine, University of Southern California

HapMap is frequently used as the reference population for imputation, and there have been discussions about which HapMap populations should be used for imputation in Hispanics, an admixed population with mainly Native American and European ancestries. In order to compare the imputation quality from different approaches, we use the USC Children's Health Study (CHS) to perform the imputation through MACH: (1) using four HapMap populations (CEU, MEX, Asian, and YRI) respectively as the reference population; (2) combining four HapMap populations as a single reference population; and (3) assigning individuals into groups according to their ancestry estimates (using STRUCTURE program) and reconstructing a reference panel according to the average ancestry estimates within each group. In addition, (4) we calculate an ancestry-adjusted genotype by weighting the imputed genotypes from approach (1) according to the individual ancestry estimates. Concordance is calculated between the imputed markers from the CHS GWAS data (Illumina 550 K and 610 K) and the genotypes obtained as part of the CHS candidate gene study(Illumina GoldenGate). We find that using HapMap Mexican (MEX) results in higher concordance for individuals with higher Native American ancestry; while using HapMap CEU results in higher concordance for individuals with higher European ancestry. By using all combined/weighted reference populations, we gain a little better concordance at the cost of longer time for imputation.

159

Interpreting the Mod Score Statistic in a Genomewide Scan for Asthma

Craig Teerlink (1), Martin Cryer (1), Alun Thomas (1)
(1) University of Utah

The mod score has been proposed as a possible solution to the sensitivity of the multipoint linkage statistic to model parameter misspecification. We conducted a genomewide scan for asthma using the multipoint mod score statistic in an extended pedigree resource ascertained for asthma. In our implementation of the mod score statistic, the LOD score was optimized over all model parameters including parameters to model parent of origin effects and an additional parameter to account for the presence of interfamilial heterogeneity. The genomewide scan for asthma was conducted in 81 extended pedigrees containing 1,314 people genotyped at 505 autosomal microsatellite markers. We observed a maximum mod score of 5.1 at chromosome 4p under a maternal imprinting model. This result appears statistically significant when evaluated using previously published thresholds for interpreting a multipoint mod score statistic in a genomewide scan. Despite the high magnitude of the score, the finding failed to achieve significance under empirical assessment using either a conventional *P*-value calculation ($P = 0.30$) or the recently proposed latent *P*-value method (latent P-value = 0.29). This result stresses the importance of empirical evaluation for establishing significance when using the mod score statistic. Furthermore, our analysis demonstrates the computational efficiency of the latent *p*-value method to empirically assess findings relative to a conventional *P*-value calculation.

160

Genome-wide Studies of Complementary Designs Identify Coagulation and Fibrinolytic Loci and Variants Potentially Implicated in Venous Thromboembolism

France Gagnon (1), Guillemette Antoni (2), Noemie Saut (3), Yiqiang Luo (1), Ashleigh Tuite (1), Philip S. Wells (4), Joseph Emmerich (5), Pierre E. Morange (6), David A. Tregouet (7)
(1) University of Toronto Dalla Lana School of Public Health
(2) INSERM, UMR_S 937, F-75013, Paris, France & University of Toronto Dalla Lana School of Public Health
(3) INSERM, URM_S 626, F-13385, Marseille, France

(4) Ottawa Health Research Institute, Ottawa, ON, Canada
(5) INSERM U765, Universite Paris-Descartes, France
(6) INSERM, URM_S 626, F-13385 & Universite de la Mediterranee, F-13385, Marseille, France
(7) INSERM, UMR_S 937, F-75013 & UPMC Univ Paris 06, UMR_S 937, F-75013, Paris, France

A multi-stage/design strategy is being used to identify loci that would contribute to venous thromboembolism (VTE) susceptibility by modulating coagulation & fibrinolytic factors known to be implicated in thrombosis & hemostasis. Analyses of 26 quantitative traits (QT) were first performed on 5 extended French-Canadian families including 261 individuals genotyped for 1079 microsatellites. Using Bayesian linkage methods, we identified 25 loci for 14 QT with Bayes Factors (logBF) >1.5 (strong linkage). Treatment effects were taken into account in follow-up analyses. These regions were then validated by *in silico* association analysis of published GWAS data on VTE with the aim of narrowing the linkage signals by focusing on candidate loci for VTE. Among the results, the replication of the FXII chromosome (ch) 5 locus previously identified in the GAIT study, and a novel locus on ch 8, with logBF of 2.9 and 1.9, respectively. One and two SNPs for ch 5 and 8, respectively, were associated with VTE at $P<10^{-4}$ in the GWAS. Prioritized analyses investigating their association with FXII in the Stanislas cohort, a sample of 123 French nuclear families, and with VTE in two French case-control studies, are underway. In conclusion, using a multi-stage approach based on pedigree analysis of QT, followed by an independent GWAS of VTE, was critical in identifying putative variants involved in the susceptibility to VTE, as well as providing clues into the underlying mechanisms.

161
Weighting of the Test Statistic by Family History Improves the Power to Detect Linkage.
Jeanine J. Houwing-Duistermaat (1), Andrea Callegaro (1), Ingrid Meulenbelt (1)
(1) Leiden University Medical Center

Recently a new susceptibility locus for symptomatic osteoarthritis (OA) was identified. Affected sibling pair linkage analysis was performed, followed by association analysis of candidate genes under the peak. For linkage analysis 179 affected sibling pairs and four affected trios were available. The score statistic to test for excess identical by descent sharing among the affected siblings yielded a maximum lod score of 3.03. Testing for association in the presence of linkage gave a significant *p*-value of 0.006 for rs225014 (*DIO2*). In this paper we consider additional information on family history, i.e. the number of first degree relatives with similar symptoms. 75% of the families had at least one additional affected first degree relative. One family had eight affected ungenotyped relatives.
We propose a weighted non parametric linkage test which uses the average of the scoring function (for example Sall) over the unobserved genotypes as weights. Simulations showed that the type I error of the statistic was correct and that except for high heritable traits (heritability above 80%), the power was increased when information on family history was used. By applying this new method to the OA data set, the lod score increased to 3.5. Moreover the location of the maximum lod score was now closer to the *DIO2* locus. We conclude that for complex genetic traits weighting of the statistic by family history increases the power to detect linkage.

162
Two Loci Sequentially Control Tuberculin Skin Test Reactivity in an Area Hyperendemic for Tuberculosis
Aurelie Cobat (1), Caroline J. Gallant (2), Gillian F. Black (3), Anne Boland-Auge (4), Jean-Laurent Casanova (5), Laurent Abel (1), Eileen G. Hoal (3), Erwin Schurr (2), Alexandre Alcaïs (1)
(1) Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U550, Paris, France
(2) McGill centre for the Study of Host Resistance & Departments of Human Genetics and Medicine, McGill University, Canada
(3) Molecular Biology and Human Genetics, Stellenbosch University, Tygerberg, South Africa
(4) Centre National de Génotypage, 2 rue Gaston Crémieux, 91057 Cedex Evry, France, EU
(5) Laboratory of Human Genetics of Infectious Diseases, Rockfeller Branch, The Rockfeller University, New York

Approximately 20% of persons living in areas hyperendemic for tuberculosis display persistent lack of tuberculin skin test (TST) reactivity and appear to be naturally resistant to infection by Mycobacterium tuberculosis. Among those with a positive response, the intensity of TST reactivity varies greatly. The molecular basis of TST reactivity is not known. We report here on a genome-wide linkage search for loci impacting on TST reactivity, defined either as a binary (i.e. zero vs. non zero, TST-BIN) or a quantitative (i.e. TST in mm, TST-QTL) trait in a panel of 128 families including 350 siblings from an area hyperendemic for tuberculosis (TB). We detected a major locus (TST1) on chromosome region 11 ($P=1.5\times10^{-5}$) that controlled TST-BIN, i.e. T-cell-independent resistance to M. tuberculosis. We also detected a second major locus (TST2), on chromosome 5 ($P<10^{-5}$) that controls TST-QTL, i.e. the intensity of T-cell-mediated delayed type hypersensitivity (DTH) to tuberculin. Refined genetic analysis demonstrated that these two loci control TST reactivity in a sequential manner. Our results pave the way for the understanding of the molecular mechanisms involved in resistance to M. tuberculosis infection in endemic areas (TST1), and for the identification of critical regulators of T-cell dependent DTH to tuberculin (TST2).

163
Linkage and Association Analysis for Genetic Modifiers in a Large Family with Cardiac Sodium Channel Disease.
Iris C. Kolder (1), Pieter G. Postema (2), Maarten P. Van den Berg (3), Marcel M. Mannens (4), Peter Van Tintelen (5), Freek van den Heuvel (6), Arthur A. Wilde (7), Connie R. Bezzina (8), Michael W. Tanck (9)

(1) Dep. Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam
(2) Dep. of Cardiology, Academic Medical Center, Amsterdam, the Netherlands
(3) Dep. of Cardiology, Thorax Center, University Medical Center Groningen, Groningen, the Netherlands
(4) Dep. of Clinical Genetics, Academic Medical Center, Amsterdam, the Netherlands
(5) Dep. of Genetics, University Medical Center Groningen, Groningen, the Netherlands
(6) Dep. of Pediatric Cardiology, University Medical Center Groningen, Groningen, the Netherlands
(7) Dep. of Cardiology, Academic Medical Center, Amsterdam, the Netherlands
(8) Heart Failure Research Center, Department of Experimental Cardiology, Academic Medical Center, Amsterdam
(9) Dep.Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam

To identify loci modulating electrocardiographic (ECG) parameters, we carried out linkage and association analyses at loci harboring 19 candidate genes involved in cardiac electrical activity in two large genealogically-linked kindreds with the sodium channel mutation *SCN5A*-1795insD (chromosome(Chr) 3). The mutation leads to the prolongation of several ECG parameters, including P-wave and QRS duration, PQ and QTc interval, as well as slower heart rate. ECGs and DNA were available for 327 individuals (127 carriers). Individuals were genotyped for 1433 tagging SNPs. Linkage and association analyses were performed using SOLAR and Pbat.
As expected, linkage peaks were found for all ECG parameters studied on Chr 3 in the region of the mutation (LOD 4.55–19.48). Additional peaks were found after adjusting for carriership on other loci, including PR interval (*KCNE1/2*; LOD 4.05). In contrast, association did not reveal any significant effects with or without adjusting for carriership. The lowest *P*-values ($P < 0.01$, uncorrected for multiple testing) on Chr 3 were observed for SNPs in LD with *SCN5A*. Based on the present study, the association test used, picked up the major known effect of the mutation only in *P* wave duration, not in the other parameters. This main effect was easily identified by linkage analysis in all ECG parameters known to be affected by the mutation. Therefore, linkage analysis still proved to be the most effective choice for the analysis of our family data.

164
Maximum-Likelihood-Binomial Method Revisited
Aurelie Cobat (1), Laurent Abel (1), Alexandre Alcaïs (1)
(1) Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U550, Paris, France

In complex traits, model-free linkage analysis methods based on identity-by-descent allele sharing are commonly used. In this context, the Maximum-Likelihood-Binomial (MLB) approach is a popular approach that relies on the idea of binomial distributions of parental alleles among affected sibs. Extension to quantitative traits (QT) has been proposed (MLB-QTL) based on the introduction of a latent binary variable which captures the linkage information between the QT and the marker. Interestingly, the MLB-QTL does not need to decompose the sibships into their constitutive sibpairs and makes no assumption on the distribution of the QT. We propose a new formulation of the MLB method for quantitative traits (nMLB-QTL), that explicitly takes advantage of the independence of the paternal and the maternal allele transmission under the null hypothesis of no linkage. Simulation studies under H0 showed that the nMLB-QTL method provides very consistent type I errors. In addition, simulations under the alternative hypothesis showed that the nMLB-QTL was slightly but systematically more powerful than the MLB-QTL, whatever the genetic model, the residual correlation, the ascertainment strategy and the sibship size. Consistent with the simulation study, linkage analysis of the Mitsuda reaction (a quantitative endophenotype of leprosy infection) and chromosome 17, as already reported in Vietnamese families, with the nMLB-QTL provided a more significant linkage signal than the original MLB-QTL.

165
Localization of a Recessive IPF1 P63fsX60 Mutation Causing Pancreatic Agenesis in Two Ostensibly Unrelated Probands Using Linkage Analysis and an Artificial Family Structure
Jennifer E. Below (1), Stefan S. Fajans (2), Graeme I. Bell (1), Veronica P. Paz (1), Catherine Martin (2), Inas H. Thomas (2), Ming Chen (2), Nancy J. Cox (1)
(1) The University of Chicago
(2) University of Michigan

MODY4 is a rare form of maturity-onset diabetes of the young due to mutations in *IPF1*, a transcription factor critical to pancreatic and beta-cell development and function. Two unrelated probands from MODY4 families affected with pancreatic agenesis were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0. Standard methods of detecting extended haplotypes shared identically by descent (IBD), such as PLINK, assume that the probability that two ostensibly unrelated individuals share both segments of a chromosome IBD is negligible. Since the inheritance of pancreatic agenesis is consistent with a rare recessive genetic model, another method was developed to map shared regions. An artificial family structure including two consanguineous loops was created to relate the probands. We conducted a linkage analysis with the software program ALLEGRO assuming a fully penetrant rare recessive model, using the imposed pedigree structure and a reduced SNP map to test for the possibility of homozygosity by descent. A single 2.5 MB region (chr 13: 27.3–29.8 Mb, NCBI) was shared homozygous IBD in these patients; overlapping the known mutation within the IPF1 gene presumed to cause disease. The size of the shared region and the geographic origin of the families suggest that the P63fsX60 mutation emerged in a recent ancestor common to both patients, and that a complex pedigree structure connects them.

166
Genome-Wide Linkage Scan for Prostate Cancer Susceptibility in Combined Finnish Populations Identifies Evidence of Linkage on 17q21–22

Cheryl D. Cropp (1), Claire L. Simpson (1), Tiina Wahlfors (2), Asha George (3), Ha Nati (2), Teuvo Tammela (4), Johanna Schleutker (2), Joan E. Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health
(2) Institute of Medical Technology, University of Tampere and Tampere University Hospital, Tampere, Finland
(3) National Human Genome Research Institute, National Institutes of Health and Fox Chase Cancer Center, Philadelphia, Pennsylvania
(4) Department of Urology, Tampere University Hospital, University of Tampere, Tampere, Finland

Prostate cancer (PRCA) genome-wide-linkage (GWL) studies have been used to localize rare, highly penetrant susceptibility loci, resulting in multiple signals. Our first GWL (GWL-I) study identified two novel loci in 10 families. Our second GWL (GWL-II) using 467 microsatellite markers in 44 new Finnish PRCA families gave some evidence of linkage to 15q26, 4q13, 17q21–22, and 14q32. An Ordered Subset Analysis (OSA) of GWL-II gave increased evidence of linkage to 17q21–22, 15q26, and 4q24. We then combined GWL-I and GWL-II families and conducted multipoint analyses using GENEHUNTER-PLUS. We also performed a subset analysis on 27 families from GWL-I/GWL-II with at least two aggressive PRCA patients. The combined GWL-I/GWL-II gave evidence of linkage to nine regions: 17q21-22, 10q22, 14q32, 4q22–23, 4q25, 3q25–26.3, 15q26, 13q34, 6q12–16 with the highest multipoint HLOD = 3.62 ($\alpha = 0.91$) at 17q21–22 with a corresponding maximum NPL score of 2.92 ($P = 0.002$) and homogeneity multipoint LOD = 3.61. Linkage to this region has been observed in other studies. However, unlike several studies, the results of our subset analysis of aggressive cases in 27 pedigrees gave lower evidence of linkage to 17q21–22 than when all families were analyzed together. Our genome-wide significant evidence of linkage to this region, combined with other published evidence of linkage to 17q21–22 suggests that this region may harbor an important susceptibility gene for PRCA.

167
Linkage and Association Analysis for Ocular Cup/Disc Ratio

Priya Duggal (1), Robert Wojciechowski (2), Alison P. Klein (3), Ching-Yu Cheng (2), Kristine Lee (4), Ronald Klein (4), Barbara E.K. Klein (4), Joan E. Bailey-Wilson (2)
(1) Johns Hopkins Bloomberg School of Public Health
(2) NHGRI/NIH
(3) Johns Hopkins Medical Institution
(4) University of Wisconsin

Primary open-angle glaucoma (POAG) is a leading cause of blindness in the world. POAG is characterized by abnormal retinal ganglion cell death that leads to optic nerve damage. However, glaucomatous nerve damage is often difficult to assess because of a lack of uniform diagnostic criteria. The cup-disc ratio of the optic disk (CDR) is a highly heritable trait and is strongly associated with glaucoma development. We evaluated the quantitative trait CDR to identify genes that may increase susceptibility to POAG. We performed a genome wide linkage scan of CDR on 895 sibling pairs 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 LD. In addition, we followed-up potential linkage regions identified in this linkage study with an association analysis of markers from a 550 K Affymetrix SNP panel in the Framingham Eye Study. Results from the linkage and association analysis will be presented.

168
Ascertainment Bias for Markov Chain Monte Carlo Segregation and Linkage Analysis of Age-at-onset data

Jianzhong Ma (1), Christopher I. Amos (1)
(1) UT MD Anderson Cancer Center, Department of Epidemiology

We have previously investigated the ascertainment problem for the Markov chain Monte Carlo (MCMC) method (Loki) for usual quantitative traits. In this study, we evaluated the performance of the age-at-onset version of Loki, with special attention on ascertainment bias. We first simulated the ascertainment of an extended pedigree with variable age-at-onset of a disease determined by two loci. For our simulated data, we detected no effect from ascertainment on the estimates of the trait locus location and found a strong linkage signal even when the allele frequencies and penetrances could not be well estimated due to small sample size. Ascertainment bias affected estimation of allele frequencies, as expected. Estimation of other segregation parameters was affected by both the ascertainment scheme and the residual variance. When pedigrees were ascertained through young to middle-aged probands, a gene was found to be overdominant if the difference between its genotypic means was large compared to the age variation in the sample, a phenomenon that is often encountered when analyzing real data. Our results show that estimation of trait-affecting gene location is not affected by ascertainment but the estimated mean genotypic ages at onset can be more difficult to estimate from ascertained samples.

169
Suggestive Linkage for an Electrophysiological Trait Indexing a Schizophrenia Endophenotype in a Nepalese Population Genetic Isolate

Susan L. Santangelo (1), Tatiana Sitnikova (1), Payas Shrestha (2), Mei Hua Hall (3), Dongmei Yu (1), Saroj Prasad Ojha (4), Janardan Subedi (5), Ram Hari Chapagain (2), Sher Bahadur Kamar (2), John L. VandeBerg (6), John Blangero (6), Sarah Williams-Blangero (6)
(1) Harvard Medical School, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA
(2) Nepal Biomedical Research Center, Kathmandu, Nepal
(3) Harvard Medical School, Psychology Research Laboratory, McLean Hospital, Belmont, MA
(4) Tribhuvan University Institute of Medicine, Dept. Psychiatry, Kathmandu, Nepal

(5) Miami University, Dept. Sociology, Gerontology, Oxford, OH
(6) Southwest Foundation for Biomedical Research, San Antonio, TX

Electrophysiological traits, indexing schizophrenia endophenotypes, were measured in over 600 members of a genetic isolate in Jiri, Nepal, who are members of a single pedigree. Measured traits include P50 and P300 amplitude and latency, and oscillatory brain activity in the gamma-band range (35–45 Hz). The power of induced gamma-band activity, as measured in the classic auditory oddball paradigm, was calculated at the PZ, P3 and P4 scalp locations in 614 individuals and evaluated in a quantitative trait locus (QTL) genome-wide linkage analysis. Heritabilities for gamma response at all three scalp locations were significant. QTL linkage analysis for gamma response yielded peak LOD scores of 1.63 on chromosome 6 (182 cM) for PZ; 1.42 on chromosome 2 (98 cM) for P3, and 1.84 on chromosome 2 (98 cM) for P4, which qualifies as suggestive linkage in this pedigree and marker set. There was also a LOD of 1.31 on chromosome 2 (98 cM) for PZ. Therefore, suggestive linkage was obtained at chromosome 2p13 at the P4 electrode, with similar signals at the same chromosomal location at the PZ and P3 electrodes. Linkage to this same region was previously found for schizophrenia in another population isolate in Palau, Micronesia. Genes in this region include sepiapterin reductase (SPR), an enzyme that catalyzes the final step in the synthesis of tetrahydrobiopterin (BH4), an essential cofactor for synthesis of many neurotransmitters, including serotonin.

170

Linkage Analysis Conditional on Measured Genotype Identifies a Candidate Region for apoB Level on Chromosome 4q32.3.

Ellen M. Wijsman (1), Joseph H. Rothstein (1), Ron M. Krauss (2), John D. Brunzell (1), Arno G. Motulsky (1), Gail P. Jarvik (1)
(1) University of Washington
(2) University of California

Familial combined hyperlipidemia (FCHL) is a complex trait leading to cardiovascular disease risk. Elevated apolipoprotein B (APOB) levels, and low density lipoprotein (LDL) size and density, are associated with the trait. FCHL is genetically heterogeneous and likely caused by rare variants. Identification of rare variants may be easier in large pedigrees than in small pedigrees or population samples, because of the opportunity to use linkage detection and linkage conditional on genotype within pedigrees to identify causal variants. We have carried out an STR-based genome scan of 4 large FCHL pedigrees ($N = 255$ individuals) for APOB adjusted for LDL (APOB-a). Analysis used MCMC methods: Bayesian joint oligogenic linkage and segregation analysis and parametric Lod scores. Followup included SNP genotypes from the Illumina 48 K HumanCVD panel. Several regions with evidence for linkage of APOB-a in individual pedigrees supports the rare-variant model. Evidence for linkage was strongest on chr 4q, with analysis in one pedigree giving Lod = 3.1 and log Bayes Factor (BF) = 1.5. Evaluation of 293 SNPs spanning the region, using each as a major gene covariate to explain the linkage signal, identified one SNP that completely explained the evidence for linkage ($\log BF = 0$). This SNP accounted for 23% of the phenotypic variance, with heterozygotes for the rare allele having a trait value that was ~30% higher than in the common homozygote, identifying a strong candidate region.

171

A Genome-wide Scan for Quantitative Trait Loci Influencing Neurocognitive Phenotypes for Schizophrenia

Wei J. Chen (1), Yin-Ju Lien (1), Chih-Min Liu (2), Po-Chang Hsiao (3), Stephen V. Faraone (4), Ming T. Tsuang (5), Hai-Gwo Hwu (2)
(1) Institute of Epidemiology, College of Public Health, National Taiwan University
(2) Department of Psychiatry, College of Medicine, National Taiwan University
(3) Genetic Epidemiology Core Laboratory, Research Center for Medical Excellence, National Taiwan University
(4) Departments of Psychiatry and of Neuroscience and Physiology, SUNY Upstate Medical University
(5) Department of Psychiatry and the Center for Behavioral Genomics, University of California

Neurocognitive impairment is one of the core symptoms of schizophrenia. This study aimed to identify regions containing susceptibility loci for the neurocognitive phenotypes as well as the factor scores of these schizophrenia-related traits. The sample comprised 1,207 affected individuals and 1,035 unaffected individuals of Han Chinese ethnicity from 557 sib-pair families co-affected with DSM-IV schizophrenia. Subjects completed a face-to-face semi-structured interview, the Continuous Performance Test (CPT), the Wisconsin Card Sorting Test, and were genotyped with 386 microsatellite markers across the genome. A series of autosomal genome-wide multipoint non-parametric quantitative trait locus linkage analysis were performed in affected individuals only. Determination of genome-wide empirical significance was implemented using 1,000 simulated genome scans. Evidence for nonparametric linkage z (NPL-Z) scores greater than 3.0 were found on 5q, 6q, and 12q for CPT indexes, respectively. The highest linkage peak was 3.32 for CPT hit rate on 12q24.32 at marker D12S2078 with a significance genome-wide significance (NPL-Z scores = 3.32, genome-wide empirical $P = 0.03$). The region 12q24.32 has seldom been implicated in previous linkage studies of schizophrenia. Therefore, the identification of this chromosomal region as a potential quantitative trait locus for schizophrenia is a novel finding. This result may inform functional hypotheses in further genetic analyses for schizophrenia.

172

Statistical Analysis of Genetic Data in South Africa

Lize van der Merwe (1)
(1) Medical Research Council, South Africa

In South Africa, there are many active studies in human genetics. Some are based on extended families, from which

we have been recruiting individuals and collecting data, for several years, There are also many case-control studies, case-only studies and cross sectional studies. We are interested in genotype-phenotype association and gene-gene as well as gene-environment interactions on phenoypes. Some of our cohorts consist of families bearing known disease-causing mutations. In those studies, we are investigating genes and environmental factors that are confounding or modifying the the genetic founder mutation effects. Before any of these analyses can be started, exploratory analyses must be done to determine whether assumptions underlying the validity of statistical tests are met, including Mendelian inheritance and Hardy-Weinberg equilibrium are met. Otherwise corrective action needs to be taken prior to analysis. The phenotypes being investigated take many forms, most often continuous, often dichotomous, counts or censored time-to-event data. I will introduce some of the studies, which I am currently analysing and the statistical methods used in the analyses. I will also specify the software I use and ellaborate on my reasons.

173

The Longitudinal Association of Common Susceptibility Variants for Type 2 Diabetes and Obesity with Fasting Plasma Glucose and BMI

Rebecca J. Webster (1), Nicole M. Warrington (1), Michael N. Weedon (2), Andrew T. Hattersley (2), John P. Beilby (3), Timothy M. Frayling (2), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, University of Western Australia
(2) Peninsula Medical School, University of Exeter
(3) Pathology and Laboratory Medicine, University of Western Australia

Variation in the effects of genetic variants on physiological traits with age or over time may alter the trajectories of these traits. However, few studies have investigated this possibility for variants associated with diabetes or obesity, and these show little consensus. We investigated the longitudinal associations of common diabetes susceptibility variants in the *KCNJ11*, *PPARG*, *TCF7L2*, *IGF2BP2*, *CDKAL1*, *SLC30A8* and *HHEX* genes, with fasting plasma glucose; and of an obesity-associated variant in the *FTO* gene, with BMI. The study analysed data from the Busselton Health Study ($n = 4{,}554$). Cross-sectional association analyses included family data. Longitudinal association analyses of unrelated participant data ($n = 2{,}864$) used linear mixed-effects models. Cross-sectional analyses showed associations of the T allele at the *IGF2BP2* single nucleotide polymorphism (SNP) rs44022960 with raised fasting glucose ($P = 0.045$), and the A allele at the *FTO* SNP rs9939609 with raised BMI ($P = 0.003$). Longitudinal analyses showed no significant associations between SNPs and changes in fasting glucose or BMI, either over a mean follow-up time of 17 years or with age from 18–80 years, though there was evidence that the BMI-raising association of rs9939609 decreased with age in the 40–60 year age group ($P = 0.024$). In summary, there was no indication that the effects of common variants on fasting glucose varied longitudinally. However, the BMI-raising effect of rs9939609 may decline during middle-age.

174

Modelling Complex Longitudinal Data in Genetic Association analyses

Nicole M. Warrington (1), Laurent Briollais (2), Julie A. Marsh (1), Craig E. Pennell (3), Stephen J. Lye (2), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, The University of Western Australia, Perth, Australia
(2) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto, Canada
(3) School of Women's and Infants' Health, The University of Western Australia, Perth, Australia

The focus of genetic epidemiology is moving beyond simple case-control analyses to gene characterization and more complex analyses involving changes over time in quantitative measures related to common disease. Linear mixed models (LMM) have become the most frequently used analytic tool for longitudinal data analysis with continuous repeated measures. The LMM framework assumes that the random effects and the within-subject measurement errors have a multivariate normal (MVN) distribution. Although this assumption allows for computational and mathematical convenience, it is often too restrictive and lacks robustness to departures from MVN. We have considered more flexible modelling approaches using the skew-normal and skew-*t* distributions for the random effects and error terms. These models assume that the random effects are multivariate skew-normal; a skewness parameter for each component of the random effects and the within subject measurement errors are assumed to follow a multivariate *t* distribution, and a parameter controlling the kurtosis of the distribution is estimated. We investigated the performance of these models using simulated data. The models were also illustrated using complex BMI trajectory data over the first 14 years of life (8 time points) and the FTO gene in the Raine Birth Cohort Study. In both the simulated and Raine datasets, we demonstrated a better fit to the positively skewed data using the skew-normal and skew-*t* distributions than the LMM.

175

Weighing up the Evidence: a Comparison of Antenatal Growth Trajectories and Birth Weight in Genetic Analyses

Julie A. Marsh (1), Nicole M. Warrington (1), Craig E. Pennell (2), John P. Newnham (2), Adrian J. Baddeley (3), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, The University of Western Australia
(2) School of Women's and Infants' Health, University of Western Australia
(3) School of Mathematics and Statistics, University of Western Australia

Birth weight is commonly used as a surrogate for growth *in utero*, despite low correlations between ultrasound biometry and birth size. Many pregnancy cohorts include only a single ultrasound and birth measurement. We compared the sensitivity and power of anthropometric measurements at birth to those derived from antenatal ultrasound, in the context of genetic association studies of early life growth. The Raine Pregnancy Study recruited

$n = 2900$ women at 16-18 weeks gestation and randomised them equally to receive a single 18 week ultrasound scan or multiple scans at 18, 24, 28, 34, 38 weeks gestation. Analyses focused on 2,065 full-term, singleton-births. Linear mixed models (LMM) were used to predict individual growth trajectories for head circumference (HC) from week 18 to birth using best linear unbiased predictors (BLUPs). Finally, a method for imputing individual growth trajectories is proposed based on a single ultrasound and birth measurement. The performance of BLUPs under a range of conditions was investigated using simulated data. Our results suggest that antenatal growth trajectories have greater sensitivity and power to detect genetic associations compared to measurements at birth, are broadly comparable to LMMs in complete data, and that it is feasible to impute BLUPs for growth trajectory in pregnancy cohorts with only two timepoints measured. This work suggests new avenues to investigate the developmental origins of health and disease (DOHaD) hypothesis.

176

Multiple Testing Correction Methods for Genetic Association Studies

Changchun Xie (1)
(1) McMaster University

Gao et'al. [2008] considered a multiple testing correction for genetic association studies and claimed that their method provides a highly accurate approximation to the permutation based method. Their simulation showed Li and Ji's [2005] method is too liberal and Bonferroni correction is too conservative. They did not include Cheverud's [2001] method in their comparison since it did not offer much improvement over Bonferroni correction. However, for the correlation matrix given in this paper, we show that Gao's method is more conservative than Cheverud's method. In this paper, we illustrates that the performance of all these non-permutation based methods depends on the structure of the correlation matrix of the SNPs used. In order to balance the computing time and accuracy, we propose a new approach, which combines permutation method and Bonferroni correction. This new approach requires much less computing time than permutation method alone and is less conservative than Bonferroni correction alone.

References

Cheverud JM. 2001. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 87:52–58.
Gao XY, Starmer J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlation single nucleotide polymorphisms. *Genetic Epidemiology* 32:361–369.
Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95:221–227.

177

Adjustment for Multiple, Correlated Tests in Pathway Analysis of Sex Steroid Hormone Levels in Pre- and Postmenopausal Women from the Breast and Prostate Cancer Cohort Consortium (BPC3)

Lars Beckmann (1), Anika Huesing (1), Wendy V. Setiawan (2), Regina Ziegler (3), Susan Hankinson (4), Rudolf Kaaks (1)
(1) Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany
(2) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles
(3) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD
(4) Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, and Department of Epidemiology, Harvard School of

We present an approach to control the family-wise error rate in association analysis of traits with SNPs in multiple genes, which requires careful adjustment for the number of tests as well as the correlation between the SNPs due to linkage disequilibrium (LD).
The approach is evaluated in an analysis of the association of 700 SNPs in 36 genes in the sex steroid pathway in a pooled sample from the European Prospective Investigation into Cancer and Nutrition (EPIC) and the Nurses' Health study (NHS). Levels of dehydroepiandrosterone (DHEAS), androstenedione (Ä4), and testosterone (TESTO), as well as estrone (E1) and estradiol (E2), and sex-hormone binding globulin (SHBG) were measured for women who subsequently developed breast cancer and in matched control subjects ($n = 3{,}852$).
We performed linear regression on each SNP for four models: recessive, dominant, codominant and log-additive, and defined pmin = min(prec, pdom, pcod, padd). For all SNPs within a gene, 1,000 permutations were used to adjust pmin for multiple testing and correlation using a step-down-min-p-algorithm. A global p-value was derived by Bonferroni-correction by the number of genes.
Significant signals were found for pre- and postmenopausal women for SHBG in the SHBG gene ($P < 10^{-3}$). In postmenopausal women, we found globally significant results for E1 and E2 with CYP19 ($P < 10^{-3}$) and FSHR ($P < 10^{-3}$); E1 and ESR1 ($P < 10^{-3}$); and DHEAS with FSHR ($P < 10^{-3}$) and AKR1C3 ($P < 10^{-3}$).

178

Rapid Correction of Multiple Testing for Multilocus Analysis

Zhaoxia Y.U. (1)
(1) Department of Statistics, University of California, Irvine

Permutation-based methods have been widely used to control the familywise error rate when a large number of correlated tests are performed. Recently several strategies have been developed to reduce the computational costs brought by permutations, and most of which are based on the assumption that test statistics of individual loci follow a multivariate normal distribution asymptotically. However, it is not clear how these methods can be applied to more complicated and powerful multilocus analysis. We present a new method to efficiently correct for multiple testing when multilocus tests are considered. Our methods are evaluated using real data with thousands of single nucleotide polymorphisms.

179

Correcting For Multiple-Comparisons in Genome-Wide Association Studies

Dalin Li (1), Juan P. Lewinger (1), Duncan C. Thomas (1), David V. Conti (1)
(1) University of Southern California

Due to linkage disequilibrium across the genome, association tests in genome-wide association studies (GWAS) are correlated. Conventional approaches to adjust for multiple comparisons such as Bonferroni correction are overly conservative, while resampling-based methods such as permutation test are computationally infeasible in GWAS. Here we propose a multiple-comparison adjustment approach that can be efficiently applied in GWAS. The approach is based on the following rational: if we break the genome-wide data into blocks of hundreds or thousands of SNPs each, under the null hypothesis SNPs in non-adjacent blocks can be viewed as conditionally independent. Based on this rational, the test statistics under the null hypothesis in each block are sequentially generated from multivariate normal distribution conditional on the null test statistics in previous block. Adjusted *P*-values can be easily obtained by comparing the observed test statistics to the generated null. Our simulations indicate that adjusted *P*-values from the proposed approach are highly consistent with permutation test ($R^2 = 0.99997$) with greatly increased computational efficiency (about 4 hours for 1 M SNPs; permutation test would take years). With the nominal type I error in the proposed approach, a moderate gain in power is observed comparing to the Bonferroni correction. Using this approach, we further calculated the genome-wide threshold of significance across variant platforms for the four HapMap populations.

180

Some Issues on *P*-values in Genetic Association Study

Indranil Mukhopadhyay (1), Arunabha Majumdar (1), Partha P Majumder (1)
(1) Indian Statistical Institute

In a genome-wide association study or a large-scale candidate gene study, the main goal is to identify genes associated with a phenotype. Availability of a large number (say, *K*) of single nucleotide polymorphisms (SNPs) compels us to use corrected probability of Type-I error (α) for the multiple testing problem that can arise due to testing *K* SNPs. However, the usual Bonferroni correction with *K* becomes too conservative. Thus a very low α-value might lead to failure to detect a true signal for association. Lowering the value of *K* by an arbitrary number, say $K/3$, also demands explanation on its effect to detect the true association, sometimes in presence of LD. In this study we have done extensive simulations to analyze different situations. SNPs with and without LD structure have been considered. The SNPs within an LD block varies both in numbers and strength. The decision using one SNP might have an influence on another if the two are in strong LD. On the other hand, if the two SNPs are not in LD, both SNPs should be tested independently at the same level. Many such scenarios should be taken into account while applying any correction to α. We have done an extensive study of the different scenarios in the context of correction applied to the p-value using Bonferroni and also FDR correction. This study might provide us to a better understanding of the situations and lead to a more meaningful solution to this mutiple-testing problem.

181

Comparison of Self-Contained Gene Set Methods for Gene Expression Studies

Brooke L. Fridley (1), Gregory Jenkins (1), Joanna M. Biernacka (1)
(1) Mayo Clinic College of Medicine

Gene set methods incorporate prior biological knowledge into statistical analyses and aid researchers in the interpretation of the results. Over the past few years, multiple approaches for gene set analysis have been proposed for expression and SNP data. The various methods can be divided into two types: competitive and self-contained. Benefits of the self-contained methods are that they can be used for genome-wide, candidate gene, or pathway studies; and these tests are more powerful then the competitive methods. We investigated numerous self-contained methods that can be used for both continuous and discrete phenotypes. To assess the power and type I error rate for the approaches, an extensive simulation study was completed in which the scenarios varied according to: correlation between genes within a pathway, number of genes in a pathway, number of associated genes, effect sizes, and the sample size. The following methods were assessed: tail strength (TS), principal component analysis (PCA) using either 80% of the explained variation threshold, first PC or top 5 PC, a global model, Kolmogorov-Smirnov (KS) test, Fisher's method based on asymptotic and empirical distribution. Results from simulations in which genes were independent showed that the KS and TS methods perform the worst for most scenarios, while the PCA using 80% variation threshold and Fisher's methods preformed the best for the majority of scenarios.

182

Gene Set analysis in Genome-wide Association Studies

Nathan L. Tintle (1)
(1) Hope College

Gene set analysis is now a standard method for analyzing gene expression data. In contrast to traditional gene expression data analysis which examines significance gene-by-gene, gene set analysis is used to establish the significance of biologically relevant sets of genes. Recently, gene set analysis has been proposed for the analysis of single nucleotide polymorphism data in genome-wide association studies (GWAS). First attempts to use gene set analysis in GWAS have proposed the use of Gene Set Enrichment Analysis (GSEA) and methods based on Fisher's exact test (FET), arguably the two most popular gene set analysis methods for gene expression data. However, recent advances in gene set analysis for gene expression data have proven to be more powerful and robust than GSEA and FET. I will present results showing

that these newer methods can be used for gene set analysis of GWAS to increase power compared to GSEA and FET. Results are based on analysis of real and simulated GWAS data. I will also present a number of open, practical questions about best practices for the use of gene set analysis in GWAS.

183
Expert Knowledge from Protein–Protein Interaction Databases to Guide Genome-Wide Genetic Analysis of Common Human Diseases
Kristine A. Pattin (1), Jiang Gui (1), Jason Moore (1)
(1) Dartmouth College

Discovering gene-gene interactions in genome-wide studies is a computational problem resulting from the analysis of all possible combinations of single-nucleotide polymorphisms (SNPs), and we wish to exploit the expert knowledge from protein-protein (PPI) interaction databases to facilitate the analysis of genome-wide studies. We use the confidence score for PPIs in the database STRING to develop metrics by which we can prioritize SNPs in pseudo-artificial bladder cancer datasets. Our data sets were simulated to have two functional SNP interactions where the SNPs represent a diversity of interaction scenarios and a range of confidence scores. We evaluate a total of 5 metrics to see which metric(s) allow us to reduce gene list of the data set significantly while retaining the two functional SNPs. We observed a correlation between confidence score and the gene list size for all metrics. The metric MAX-SUM is a measure of genes prioritized by their maximum confidence score and then subsequently by the sum of the confidence scores of all interactions per gene. We find that MAX -SUM reduced the data sets significantly more than metrics AVE, MAX, and MAX-AVE ($P = 0.03$, 0.05, 0.03). While overall this metric was most effective at reducing the gene list in a majority of scenarios, other metrics were more effective for certain interaction scenarios. We plan to implement these metrics in a bioinformatics tool designed for this purpose.

184
Joint Analysis of Multiple Genes in a Pathway or a Gene Set
Jingyuan Zhao (1), Jianjun Liu (1), Anbupalam Thalamuthu (1)
(1) Genome Institute of Singapore

Multiple testing of individual Single Nucleotide Polymorphism (SNP) only captures a small proportion of associated SNPs due to small marginal effects and multiple corrections. The discovery of susceptible pathways involving multiple genes holds promise to reveal the disease mechanism. We introduce a gene-based approach for the joint analysis of multiple genes in a pathway or a gene set. First, the matrix of SNP genotypes for each gene is represented by the gene-based score using principle component analysis. It is well known that among all the genes within a pathway, only a small subset of genes has contribution to the disease. All gene-based scores are ranked by the penalized likelihood method and assessed by some model selection criterion to select the best subset of genes. The proposed method can be used for testing the joint effect of several genes in a candidate gene study and can easily be extended to identify important genes in specific biological pathways generated from a Genome-wide Association (GWA) studies. Simulation studies show that our proposed method enjoys a higher power than other approaches, even if associated genes have a large number of genotyped SNPs.

185
Evidence of ACOX3, B4GALT6, CHAT, NQO1, and TBXAS1 Variants for LDL/HDL Ratio in the NHLBI Family Heart Study
Ping An (1), Mary Feitosa (1), Michael A. Province (1), Ingrid B. Borecki (1)
(1) Division of Statistical Genomics, Washington University School of Medicine

LDL/HDL ratio is a risk factor of T2D and CHD. Variants of the KEGG lipid metabolism pathway genes were assessed for their association with the LDL/HDL ratio in the NHLBI Family Heart Study (FHS). A total of 831 Caucasian family members had phenotype and genotype data with up to 2.5 million typed and imputed SNPs in the FHS. A total of 2,141 SNPs covering 300 KEGG lipid metabolism pathway genes were assessed using regression under a mixed model with sandwich estimator to account for familial dependencies, assuming additive effects. Covariates in this analysis included age, age2, sex, and field center. $P < 0.0061$ from a modified FDR approach was used to flag significance considering correlations of these genes and SNPs. Significant association was found in 5 genes that included 2 SNPs in ACOX3 (rs735812, $P = 0.0024$), 29 SNPs in B4GALT6 (around rs1449087, $P = 0.0002$), 1 SNP in CHAT (rs4838541, $P = 0.0047$), 6 SNPs in NQO1 (around rs1437135, $P = 0.0013$), and 16 SNPs in TBXAS1 (around rs12532529, $P = 0.0051$). For risk of T2D, risk allele frequencies of these variants were 0.17–0.87, and effect size estimates were $<1\%$. Mechanisms of these gene variants in influencing the LDL/HDL ratio are currently unclear. In conclusion, we screened 300 lipid metabolism pathway genes and found effects of variants in ACOX3, B4GALT6, CHAT, NQO1, and TBXAS1 genes on the LDL/HDL ratio. Survey of SNP*SNP and BMI*SNP interacting effects of these variants for the LDL/HDL ratio is underway in the FHS cohort.

186
Withdrawn

187
SalamboMiner: a Literature Database Mining Tool Based on Bayesian Networks
Alfonso Buil (1), Leonor Rib (1), Jose Manuel Soria (1), Ricard Gavaldà (2)
(1) Unit of Genomics of Complex Diseases. Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau. Barcelona. Spain
(2) Departament de Llenguages i Sistemes Informatics. Universitat Politecnica de Catalunya. Barcelona. Spain

Literature database mining tools are becoming an important help in the task of finding genes related to complex diseases. We present a tool that allows the discovery of relationships among genes, proteins and diseases, based on the degree of co-citation among concepts in the Pubmed database.
From every Pubmed register, we extracted biological concepts using two text mining tools, Biotagger and MetaMap, and we gave semantic categories to these concepts using the Unified Medical Language System (UMLS). Then, for two concepts that are co-cited, we use different measures of association based on the contingency table between them. On the other hand, for two concepts (C1 and C2) that are not co-cited directly, we create a Bayes Network that includes all the concepts that connect C1 with C2. We feed the network with the conditional frequencies among concepts that are co-cited and then we estimate the joint distribution among C1 and C2. From this joint distribution we can calculate the same association measures that we use with two concepts that are co-cited. SalamboMiner is an application that allows questions like ''which are the diseases related with this gene/protein?'' or ''which are the genes/proteins related with this disease? In both cases, the application returns a sorted list of concepts to answer the question. SalamboMiner is able to discover relationships between concepts that are not co-cited.

188

PATH2: Analysis of Genetic Pathways with Prior Information From Electronic Databases

Denise Daley (1), Ben Tripp (1), David Zamar (1)
(1) University of British Columbia

Statistical models are becoming more complex as we seek to better understand biologic disease pathways. For common diseases disease susceptibility includes biological interactions (gene-gene or gene-environment). Pathway based methods are an example one of the many approaches to test interactions. To help determine which SNP-SNP interactions to test, we developed Path; a software application designed to help researchers interface their data with biological information from several bioinformatics resources. In the ''information age'', there is a vast quantity of knowledge that can be mined from electronic databases, but the challenge is how to automate data retrieval and appropriately incorporate this information into the statistical model. We have developed Path (Path: a tool to facilitate pathway-based genetic association analysis), to aid researchers in the incorporation of prior information in statistical analyses, the software can be freely obtained from http://genapha.icapture.ubc.ca/index.php/research/software/. Path is designed to help researchers interface their data with biological information from several bioinformatics resources. We are currently expanding Path to include additional data sources to guide users in the incorporation of prior information gathered from external sources.

189

A Comprehensive Model for DNA Repair Genes and Radiation in Second Breast Cancers: The WECARE Collaborative Study Group

Duncan C. Thomas (1), Ake Borg (2), Marinela Capanum (3), Patrick Concannon (4), David V. Conti (1), Robert W. Haile (1), Xiaolin Liang (3), Anne S. Reiner (3), Marilyn Stovall (5), Sharon N. Teraoka (4), Jonine L Bernstein (3)
(1) University of Southern California
(2) Lund University
(3) Memorial Sloan Kettering Cancer Center
(4) University of Virginia
(5) MD Anderson Cancer Center

We studied the effects of radiotherapy (RT) and genes involved in DNA damage responses in 708 women with bilateral breast cancer and 1399 controls who survived a first breast cancer a comparable length of time, matched on age, date of diagnosis, race, and center and counter-matched on RT. Radiation doses (RD) were estimated by phantom dosimetry. Among 357 SNPs in 25 genes, stepwise logistic regression found associations with one or more SNPs in *ATM*, *BRCA1/2*, *MDC1*, *CHEK2*, *NBN*, *MRE11A*, *LIG4*, and *RAD51.* Interactions with RT and/or RD were also seen for *NBN*, *BRIP1*, *CHEK2*, and *LIG4.* We are now building a comprehensive model using hierarchical Bayes regression with prior information extracted from the Gene Ontology (GO). The first level is a logistic regression of disease status on latent genotypes in the 25 genes, with the SNPs being in linkage disequilibrium with them. The second level treats these coefficients as multivariate normally distributed with means depending on selected GO terms and covariances related to the correlations across all 470 GO terms related to any of these genes. The model was fitted using the WinBUGS software. Prior covariates for three GO terms relating to DNA damage response were highly significant and the spatial autocorrelation induced by the GO adjacency matrix was estimated at 0.73 (95% CI 0.52?0.88). We describe extensions of the model that will incorporate gene-RT interactions and stochastic search variable selection.

190

Scaling Up the Investigation of Disease Pathways

James W. Baurley (1)
(1) University of Southern California

Various analysis approaches for association studies are emerging that focus on pathways, incorporating prior biological evidence and model uncertainty. We have developed a Markov Chain Monte Carlo (MCMC) method that samples from the posterior distribution of pathway topologies and estimates Bayes factors for pathways of greatest interest. As more genetic data becomes available, the space of possible structures grows large, and any algorithm aimed at discovering pathways must run long enough to adequately sample from the posterior distribution of topologies. Genetic and environmental variables and their interactions can be annotated in a hierarchy, such as the directed acyclic graph (DAG) structures of Gene Ontology (GO). By utilizing such DAGs, a large search space could be partitioned into smaller manageable spaces. Each node of the DAG has an associated MCMC process searching over topologies relevant to that node. At each cycle a new topology is proposed by modifying the current topology from a child MCMC process. Topologies are passed up the DAG to the root process and recorded. The hierarchical structure of this approach allows for

computational parallelism, with communication occurring among multiple processors. We demonstrate the feasibility of this technique by discovering a simulated pathway from approximately 100 genes annotated across a portion of the GO DAG and applying our method to oxidative stress genes extracted from an asthma genome-wide association study.

191

A Catalog of Natural Polymorphisms and their Predicted Effects on Metabolic Syndrome Candidate Genes and Proteins in the Human Sequences

Aldi T. Kraja (1), Joanne Nelson (2), D. C. Rao (3), Victor G. Davila-Roman (4), Elaine Mardis (5), Michael A. Province (1)
(1) Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO
(2) GEMS Training Program and Genome Center, Washington University School of Medicine, St. Louis, MO
(3) Division of Statistics, Washington University School of Medicine, St. Louis, MO
(4) Cardiovascular Division, Washington University School of Medicine, St. Louis, MO
(5) Genome Center, Washington University School of Medicine, St. Louis, MO

Metabolic syndrome (MetS) is becoming a pandemic in the developed countries. The next generation sequencing data, most notably the 1,000 Human Genomes project, provide an opportunity for an in-depth investigation of the genomic variation in genes associated with MetS. We recently performed a search of the literature and identified 123 candidate genes implicated in MetS. We are using the complete genomes currently available, Human Ref., Watson, Venter sequences and The 1,000 Genome Project data, as preliminary datasets to build the software necessary to interrogate the positional variation in these candidate genes. We have begun to annotate the data based on the SNPs that belong to a specific MetS candidate gene, and we will extend such annotation to INDELs. We are working to annotate the distribution of these variations in different compartments of the genes (as enhancers, promoters, introns, exons, and UTR). Also we are in the process of investigating the frequency of missense mutations in these genes across the genomes for POPRES study (6,000 subjects), GENEVA study (3,419 subjects, classified for type 2 diabetes, 0/1), HyperGEN (2,584 subjects), The Framingham Heart Study (6,000 subjects), The Family Heart Study (2,700 subjects), last three classified based on MetS (0/1). Finally, we cluster these genomes by the patterns of the missense mutations happening in the 123 candidate genes of MetS. Such clustering may provide insightful information for future research.

192

Essentially Yours …and Mine: Privacy and Cohort Pedigree research

Christine M. van Vliet (1)
(1) Department of Pathology, University of New South Wales

While much has been written about the ethics of genetic research in relation to the nature of genetic information, its potential for misuse, and its protection, little attention has been given to the ethical issues related to its conduct. Pedigree research raises ethical questions because it requires as complete and as accurate, family history as possible, to ensure valid estimates of modes of inheritance. This necessitates the collection of sensitive, identifiable data on family members without consent. Current legal and ethical guidelines in Australia and America allow consent to be waived if it is impracticable to obtain and certain other criteria are met as determined by an ethics committee. These guidelines are problematic for cohort pedigree research, as not only do they impose undue burdens and restrictions owing to their complexity and inconsistency, their focus is on individual consent for use of information, however it may be argued that family history is jointly owned and relevant to each individual's participation. As a result, it is unclear to what extent information collected on family members who decline to participate, withdraw or from whom consent is not sought, may be used.
This paper will discuss the justification and ethical and legal implications for the collection, use, disclosure and retention of identifiable, sensitive information without consent, in cohort pedigree research in Australia and America.

193

Prediction Modeling in the Context of Pharmacogenomic Genome-wide Association Study: Identifying Non-responders to PegINF-alpha/ribavirin Treatment Among Chronic Genotype 1 Hepatitis C Virus Infected Individuals

Max Moldovan (1), Vijay Suppiah (2), David Booth (2), Jacob George (2), Melanie Bahlo (1)
(1) The Walter and Eliza Hall Institute of Medical Research
(2) Westmead Hospital

Currently, about 50% of chronic genotype 1 hepatitis C virus (HCV) positive individuals do not respond to the standard of care treatment of combined pegylated interferon-alpha and ribavirin (PINFR) therapy. In our study, we analyzed a genome-wide scan of 300K+ single-nucleotide polymorphisms (SNPs) taken from 293 individuals (131 responders and 162 non-responders) with known treatment response. Our aim was to build a prediction model that can accurately classify non-responders to PINFR treatment using information on a relatively small number of pre-specified genetic markers.
In particular, we applied a newly designed three-step analytical methodology, which can be seen as a pharmacogenomic extension of a typical genome-wide association analytical methodology. Firstly, we applied an exact and computationally efficient form of efficiency robust significance testing (Sladek et'al., 2007; Nature 445:881–885) in order to select a moderately large number of SNPs (e.g. 300 to 5000) that are likely to be useful in predicting the PINFR treatment response. Secondly, a lasso penalized regression was used for detection of relevant predictive markers (Hoggart et'al., 2008; PLoS Genetics 4:1–8). Finally, a prediction model was selected based on a cross-validation type algorithm assisted by the biological knowledge regarding the genetic markers under consideration. The selected model was then validated on the independent cohort of 555 individuals (261 responders and 294 non-responders).

194

A Bagging Optimal ROC Curve Method for Predictive Genetic Tests

Qing Lu (1), Yuehua Cui (2), Chengyin Ye (1), Changshuai Wei (1), Robert C Elston (3)
(1) Department of Epidemiology, Michigan State University
(2) Department of Statistics and Probability, Michigan State University
(3) Department of Epidemiology and Biostatistics, Case Western Reserve University

With the numerous findings discovered from recently accomplished genome-wide association studies, translation studies have been initiated to assess the combined effect of new genetic loci and the existing risk factors for early disease prediction. The statistical challenge we faced in evaluating these early prediction tools is how to optimally combine a large number of genetic variants, clinical risk factors and their interactions for disease prediction. For that purpose, we propose a bagging optimal ROC curve method, based on the concept of the optimal ROC curve that has many ideal properties. The method incorporates a bootstrap aggregation procedure to provide powerful and robust performance for the test. Through simulation and real data application, we compared the new method with the commonly used allele counting method and logistic regression, and found that the new method yields a better performance. The new method was applied on the Wellcome Trust GWAS dataset to form a predictive genetic test for Rheumatoid Arthritis. The formed test reaches an AUC value of 0.7043.

195

Evaluating Variance in Liability Explained by Individual Genetic Variants and Relationship to Individualized Risk Prediction

Hon-Cheong So (1), Stacey S. Cherny (1), Pak C. Sham (1)
(1) Department of Psychiatry, University of Hong Kong

An increasing number of susceptibility genes have been identified for complex diseases in recent years. However, how much the candidate genes discovered to date could explain the total genetic component of a disease is unknown. We developed a statistical framework to address this problem focusing on dichotomous disease traits and applied it to real examples of complex diseases. The genes were mainly selected based on results from meta-analyses of association studies. The total variance contributed by known candidate genes for each disease is generally not high, implying that a substantial part of heritability for most complex diseases remains unexplained. We also extended our model to deal with multi-allelic loci, haplotypes as well as gene-gene and gene-environmental interactions. In addition, we derived methods for calculating the variance explained for continuous predictor variables. We further found that the variance explained is closely related to the ability of risk prediction for individuals. We developed a methodology to quantify the absolute disease risk from liability measures. Specificity and sensitivity could be calculated for every cutoff of the absolute disease risk, hence the receiver operating characteristic curves and areas under the curve may be computed. Finally, we developed an approach to incorporate family history of the individual into known genetic factors when predicting disease risks.

196

Utilization of HER2 Genetic Testing and Impact on Treatment Decisions for Breast Cancer Patients

Katrina Goddard (1), Sheila Weinmann (1), Kathryn Richert-Boe (1), Chuhe Chen (1), Carmel Wax (1)
(1) Kaiser Permanente Northwest

HER2 testing is recommended for women with invasive breast cancer as a prognostic tool and a guide to therapy decisions. Women with a positive HER2 result have poorer prognosis, and are more likely to benefit from Herceptin use. It is unclear how HER2 test results influence treatment decisions in practice. We examined the use of HER2 testing for 3979 breast cancer patients diagnosed in 1998–2007 at a managed care organization. In the first two years, HER2 testing increased from 10% to >90% of women diagnosed with invasive breast cancer, and remained >90% until the present. The rate of HER2 testing was <3% for women with ductal carcinoma in situ (DCIS), who are not indicated for HER2 testing. Immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) testing were used. The majority of patients received the IHC test. One-fourth of the positive or negative IHC results were discordant with the FISH result. Herceptin use increased three-fold after 2004 when guidelines expanded to include the adjuvant setting for early stage breast cancer in addition to the metastatic cancer setting. The majority of patients (79%) who received Herceptin had a positive HER2 test result. Of the remainder, 38% did not have a HER2 result recorded, although most were diagnosed prior to 2000. These findings illustrate the use of HER2 testing and Herceptin in a managed care setting, and should be interpreted in a broader context including alternative methods of healthcare delivery.

197

Family History-Based Risk of BRCA Mutations in the California Population

Nedra Whitehead (1), Yan Li (1), Linda Squires (1)
(1) RTI International

The US Preventive Services Task Force recommended in 2005 that women at a high risk of deleterious BRCA mutations be referred for genetic counseling and evaluation for testing. They estimated that 2% of U.S. women would need genetic counseling. We assessed the proportion of women in California at high risk of carrying a BRCA mutation based on family history. We analyzed data from the 2005 California Health Interview Survey CHIS), a population-based telephone survey, which included a detailed cancer family history module. We estimated the probability the respondent carried a BRCA mutation using BRCAPRO. We imputed the current age of family members and their age at cancer diagnosis as a function of the respondent's age and if the cancer diagnosis occurred before or after age 50 years. The minimum estimated probability was 0.0000014% and the maximum was 94%. The 98.5 percentile for the population was 0.63%.

Therefore, <1.5% of the population had an estimated probability of BRCA mutations >1.0%; <0.5% had a an estimated probability of BRCA mutations >10%. No information was available on BRCA mutation testing, so the predictive value of the risk estimate could not be determined. The methods we used to estimate the age of cancer diagnosis are biased towards underestimating the age of cancer diagnosis. Since younger age at diagnosis is associated with an increased risk of having a BRCA mutation, we may have overestimated the probability of carrying a BRCA mutation.

198

Genome Wide Association Studies: The Day After

Clerget-Darpoux Françoise (1)
(1) INSERM U535; Paris-Sud University

A revolution in molecular biology occurred three decades ago with the beginning of the Human Genome Project which raised many hopes and great promises in terms of Public Health. It seems important for both the scientific community and the general public to undertake a re-appraisal of these promises.
While a genome wide-association study may clearly pinpoint a region as containing susceptibility risk factors, we are still far from their identification and from the full measurement of their effects. The idea that multifactorial diseases only involve variants with moderate Genotypic Relative Risks (GRRs) is clearly wrong. In fact, the weak ORs reported by GWA studies only reflect the marginal correlation between the disease and individual tagSNPs and, at any rate, should not be interpreted as a weak effect of the genes involved in the pathological pathway. Illustration will be given on susceptibility factors of several auto-immune diseases: PTPN22 on rheumatoid arthritis, IL2Ra in multiple sclerosis, the IL2/IL21 region in celiac disease.
One should note that even variants with high GRR generally have a poor power of disease prediction. Geneticists must fight the deceit of predictive medicine while, at the same time, keeping in mind that an evaluation of relative risk genotypes could be crucial for disentangling all possible pathological pathways of a disease.

199

Weighted Kernel Fisher Discriminant Analysis (wKFDA) for Integrating Genomic and Clinical Data with Application to Breast Cancer Prediction

Jemila S. Hamid (1), Celia M.T. Greenwood (2), Joseph Beyene (2)
(1) Hospital for Sick Children
(2) Hospital for Sick Children and University of Toronto

We propose a method for integrating heterogeneous data sets for a classification task. We use kernel based statistical method where each data set is represented as a kernel matrix. Weights are assigned to each of the data sets and the kernels are combined, in a weighted fashion, to get a unified kernel representing information from all data sources. We perform kernel Fisher discriminant analysis (KFDA) using each of the data sets and define a weight based on classification error. These weights can be considered as measures of relative importance for each data set. The similarity/kernel matrices are then combined and KFDA on the combined kernel is performed to classify individuals into subclasses. We present the algorithm used to implement the proposed method and illustrate our approach by integrating gene expression and clinical data sets with the aim of improving breast cancer prediction. Our method, however, is presented in a more general setting and can be applied to combine two or more heterogeneous data sets. A preliminary analysis using our illustrative breast cancer data resulted in a weight of 0.49 and 0.51 for the clinical and gene expression data, respectively, indicating that both data sets provide comparable information for the purpose of prediction. A detailed analysis using an elaborate cross validation scheme to define the weights and combine the data optimally using wKFDA is being performed and results from this analysis will be presented.

200

Predictive Modeling of Colorectal Cancer in Case-Control SNP Studies

Niloofar Arshadi (1), Rafal Kustra (1)
(1) Dalla Lana School of Public Health, University of Toronto

Colorectal cancer is one of the leading cancers in overall mortality, but also one of the most curable ones, with over a 90% survival rate if detected early. This has invigorated a search for genetic markers, and genetic predictive models in general, that can identify individuals at increased lifetime risk for more aggressive screening. we apply a gradient boosting machine (GBM) to build a genome-wide, multilocus predictive model using GWAS colorectal cancer data with ~535,000 SNPs and 2255 subjects. The GBM prediction model is an ensemble of decision trees combined together based on the boosting paradigm. The GBM classifier also ranks variables (SNPs) in terms of their predictive power called relative influence (RI). We propose a novel technique to utilize the RI of SNPs for identification of genetic regions that may be informative of cancer risk.
SNPs were filtered based on their linkage disequilibrium ($LD <= 0.8$), low minor allele frequency ($MAF > 0.01$) and missing value rates (>10%). The GBM model is trained with all subjects in a training set, but only a subset of SNPs whose univariate *p*-value is below a threshold. Predictive accuracy was evaluated for various *p*-value thresholds using 5-fold cross-validation. The highest area under curve ($AUC = 0.76$) is achieved with *p*-value threshold [0.002,0.008], which results in about 1900 SNPs on average. Not all of these are used in building the trees, and those that are used have varying RI levels as determined by GBM.

201

A Progressive Multi-state Modelling for Predicting Disease Risk in Gene Mutation Carriers: Application to HNPCC Pedigree Data from Newfoundland

Yun-Hee Choi (1), Laurent Briollais (2), Karen A. Kopciuk (3)

(1) Department of Epidemiology and Biostatistics, University of Western Ontario, London, Canada
(2) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada
(3) Division of Population Health, Alberta Health Services, Calgary, Canada

Carriers of a major disease-causing gene are often at risk for more than one phenotype, and they may be screened to prevent some phenotypes and experience an intervention during the course of their disease process. To better understand the complexity of the disease process, we investigate disease risk or penetrance estimators for transitions from one phenotype to second using event times. We estimate transition intensity probabilities for risk of developing successive phenotypes in pedigree data and also provide age-specific risk estimates associated with subsequent cancers. A progressive multi-state model approach is adopted using both parametric and weakly parametric baseline hazards. We also develop an Expectation-Maximization algorithm to estimate the parameters of the model which can infer missing genotypes based on observed genotype and phenotype information of other family members. Our method is applied to data from a retrospective cohort of 12 independently ascertained Hereditary Non Polyposis Colorectal Cancer (HNPCC) families from Newfoundland who harbour a founder MSH2 mutation and who often experienced more than one HNPCC-related cancer. Our study confirms high-penetrance (65% by age 50 and 99% by age 70) in mutation carriers for developing a first HNPCC cancer and reports on high risks for developing a second HNPCC cancer, following a first one (e.g. 39% by age 70 if the first cancer occurred at age 40).

202

Risks of cancer for MLH1 and MSH2 mutation carriers

James G. Dowty (1), Aung K. Win (1), Daniel Buchanan (2), Stephen N. Thibodeau (3), Graham Casey (4), Noralane M. Lindor (5), Steve Gallinger (6), Loic Le Marchand (7), Robert Haile (4), Polly Newcomb (8), John L. Hopper (1), Mark A. Jenkins (1)
(1) Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Parkville, Victoria, Austra
(2) Familial Cancer Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia
(3) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota
(4) Department of Preventive Medicine, University of Southern California, Los Angeles, California
(5) Departments of Medical Genetics, Mayo Clinic, Rochester, Minnesota
(6) Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada
(7) University of Hawaii Cancer Research Center, Honolulu, Hawaii
(8) Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington

Background: Germline mutations in the DNA mismatch repair genes MLH1 and MSH2 are associated with a substantially increased risk of some cancers. Due to the rarity of these mutations, previous studies have been underpowered to provide precise estimates of age- or gene-specific risks.
Methods: We recruited 307 MLH1 ($n = 133$) and MSH2 ($n = 174$) mutation carriers (probands) and their relatives ($n = 13{,}226$) from Australasia, USA and Canada through the Colon Cancer Family Registry. Hazard ratios (HR) of cancer incidences for mutation carriers relative to those for the general population, and hence age-specific cumulative risks (penetrances), were estimated using modified segregation analyses conditioned on ascertainment.
Results: The HR for colorectal cancer (CRC) decreased with age ($P < 0.001$) but did not differ by gene ($P = 0.9$). The HR for endometrial cancer (EC) was higher for MSH2 than MLH1 (16.8 versus 2.8; $P = 0.01$).
For male carriers from the USA, the cumulative cancer risks to age 70 yrs (95% CI) were estimated to be 67% (55–79) for CRC and 15% (7–30) for other Lynch cancers. The corresponding risks for females were 30% (20–45) for CRC, 25% (12–49) and 5% (2–13) for EC (respectively for MSH2 and MLH1), and 24% (15–38) for other Lynch cancers.
Conclusion: The cancer risks for MLH1 and MSH2 germline mutations depend on age (CRC) and gene (EC). These findings have important implications for the clinical management of Lynch Syndrome.

203

Penetrance of CDKN2A mutations estimated using population-based Australian families

Mark Jenkins (1), Anne Cust (1), Daniel Schmidt (1), Enes Makalic (1), Elizabeth Holland (2), Helen Scmid (2), Richard Kefford (2), Graham Giles (3), Armstrong Bruce (2), Joanne Aitkin (4), Graham Mann (2), John Hopper (1), Australian Melanoma Family Study (5)
(1) University of Melbourne
(2) University of Sydney
(3) The Cancer Council Victoria
(4) The Cancer Council Queensland
(5)

Background: *CDKN2A* gene mutations increase risk of melanoma (67% by age 80 when using multiple-case families, and 28% by age 80 when using population-based families with unverified melanoma histories.
Methods: The Australian Melanoma Family Study, a population-based case-control-family study, included probands with incident cutaneous melanoma diagnosed before age 40 recruited from Brisbane, Sydney and Melbourne and their first- and second-degree relatives. We identified 12 probands with pathogenic or suspected pathogenic *CDKN2A* mutations. The hazard ratio (HR) of melanoma incidence for carriers relative to that for the general population, were estimated using a modified segregation analysis that incorporates both genotyped and ungenotyped relatives and conditions on ascertainment to produce unbiased estimates.
Results: The HR for reported melanoma was greater for females than males (HR = 34.2 and 9.9, respectively; $P = 0.02$). Combining males and females, the penetrance of melanoma for *CDKN2A* mutation carriers was 60% (95%

CI: 43 to 77%) by age 80 (females 70%, males 43%). The HRs for confirmed melanoma was 8.4 for males and females combined (c.f. 19.5 for reported melanoma). **Conclusion**: Our population-based estimates of melanoma risk which apply to the mutations identified in early-onset cases of melanoma, appear to be higher than previous population-based estimates from studying the families of cases unselected for age at onset and were higher for females than males.

204

Identify Association of Less Frequent Variants with Large Effect Sizes

Chao Xing (1)
(1) UT Southwestern Medical Center at Dallas

The genome-wide association study (GWAS) becomes the prevalent study design in human genetic mapping. However, its success so far is constrained to identifying common variants with modest effect sizes, which explain a small fraction of heritability of complex traits. Conventional methods to identify association in GWAS are to prioritize signals purely by p-values and select those meeting the genome-wide significance level, and thus they favor detecting common variants. Here we attempt to improve the power of GWAS by detecting association of low frequent variants with effect sizes comparable to or even larger than that of common variants. We first investigated the allelic distribution of SNPs from the Affymetrix Genome-Wide Human SNP Array 6.0 in the Atherosclerosis Risk in Communities (ARIC) Study population, which implied that a considerable amount of information on less frequent variants lies in the commercial arrays. Further, we conducted a genome-wide screen for glucose levels using a false discovery rate control procedure accounting for allele frequency, and another genome-wide screen for stature using a regularized statistic in the European American (EA) population of the ARIC study. Both screens identified associations for less frequent variants that were missed by conventional methods. This study reveals that there is a cache of less frequent variants in the commercial arrays, and this group of variants can be valuable if proper analytical approaches are used in GWAS.

205

More Low Hanging Fruit on the Family Tree? Searching for Rare Sequence Variants in Families.

Laura M. Yerges-Armstrong (1), Patrick F. McArdle (1), Toni I. Pollin (1), Alan R. Shuldiner (1), Braxton D. Mitchell (1)
(1) University of Maryland

The emergence of high throughput genotyping technologies and interest in identifying variation important at the population level has led to increased use of large samples of unrelated individuals, particularly relative to use of family samples. Though many suggest sequencing the extremes of such unrelated samples for rare variants, we propose using family samples to enrich a sequencing sample with likely rare variant carriers by selecting closely related individuals who have extreme trait values. To illustrate the value of using related individuals to inform selection of a sequencing set, we evaluated our ability to correctly identify subjects segregating a known rare null mutation in APOC3 (R19X) in the Lancaster County Old Order Amish associated with low fasting triglyceride (TG) levels. We first identified subjects with the lowest TG levels, representing the bottom 2.5% of the TG distribution, and within this set found 14 pairs of highly related individuals ($\geq$3rd degree). Although the Amish are a founder population, this was higher than expected by chance. The allele frequency of the R19X was 21% in the lower TG tail compared to 2% in the larger population, enhancing the likelihood of identifying this variant in a large scale sequencing effort. We conclude that identifying individuals who have extreme phenotype values and are highly related is a powerful strategy for selecting study subjects to be sequenced in order to identify rare trait-altering variants.

206

Rare and Common Variants in NRXN1 (2p16.3) Associated with Childhood onset Schizophrenia (COS)

Anjene M. Addington (1), Julie Gauthier (2), Dan Spiegelman (2), Yohan Lee (1), Robert Long (1), Guy Rouleau (2), Judith L. Rapoport (1)
(1) NIH
(2) University of Montreal

Neurexins are a family of proteins that function in the vertebrate nervous system as cell adhesion molecules and receptors. Involvement of NRXN1 has been implicated in schizophrenia, autism and mental retardation, mostly through the study of rare deletions that have been identified in these patient populations. Given the growing evidence of the involvement of NRXN1 in complex neurodevelopmental disorders, we tested for association with known SNPs using family-based TDT, as well as novel variants identified through targeted resequencing of 94 probands with childhood onset schizophrenia (COS) and 190 unrelated controls. Several SNPs within the gene were significantly associated with COS ($P < 0.001$). Resequencing of all exons, promoter, and conserved regions revealed a total of 66 variants among the COS probands: 8 missense, 10 silent, 46 intronic, and 2 in the 3'UTR. Of these, 23 are in dbSNP and 43 are unknown. Among the novel variants identified, 26 were not seen in any of the controls samples, including 5 of the 7 missense variants. To contrast, 2 of 4 novel missense variants identified in the 190 controls were not observed among the COS. All of the novel variants were inherited, though functional assays of select variants that are predicted to affect protein function are ongoing. These studies provide further evidence for the involvement of NRXN1 in early neurodevelopmental disorders, such as schizophrenia and autism.

207

A method to analyze extreme quantitative trait data attributable to rare variants: application to the analysis of next generation sequence data.

Daniel Covarrubias (1), Bingshan Li (2), Suzanne M. Leal (2)

(1) Rice Unversity
(2) Baylor College of Medicine

Although methods which are used to detect associations with common variants that influence QTL can be used to analyze rare variants, they are underpowered. We extended the combined multivariate collapsing (CMC) method (Li and Leal, AJHG 2008; 83:311-21) to analyze quantitative traits. The quantitative CMC (QCMC) method can be used to analyze data for which either all individuals regardless of their QTVs are sequenced or only individuals with extreme QTVs are sequenced. In order to evaluate the power and robustness of the QCMC method, sequence data was simulated via coalescence theory using parameters estimated from population genetic data and the QTV distributions were based-upon clinically relevant quantitative trait data. It was shown in most situations that there was only a slight gain in power when the complete sample was sequenced and analyzed, compared to when only those individuals with extreme QTVs in the upper and lower 25% were analyzed. When only individuals with extreme QTVs are sequenced and analyzed it is advantageous to analyze their QTVs instead of dichotomizing and implementing the CMC method, since power loss can be substantial. The QCMC is robust to misclassification (e.g. inclusion of non-causal variants and exclusion of causal variants). In conclusion, the QCMC method can easily be implemented to analyze rare variant quantitative trait data obtained from candidate genes or whole exome sequencing.

208
Analyzing Rare Genetic Variants Using WTCCC Data
Tao Feng (1), Yali Li (1), Xiaofeng Zhu (1)
(1) Case Western Reserve University

Genome-Wide Association (GWA) studies are less successful in detecting rare genetic variants contributing to common diseases because of the poor statistical power for most of current methods. We developed a two-stage method that we can classify together rare risk haplotypes using a relatively small sample for either of these designs at stage 1, and then has increased power to test association in the remaining sample at stage 2. Here we report the results of applying this two-stage method to the Wellcome Trust Case Control Consortium (WTCCC) dataset that include 7 complex diseases: Bipolar disorder, Cardiovascular disease, Hypertension, Rheumatoid Arthritis, Crohn's disease, Type 1 Diabetes and Type 2 Diabetes. We identified several genes reach genome wide significance for each trait, including some genes detected in the original WTCCC study. Our analysis suggests that searching for rare genetic variants is feasible in current genome-wide association studies, candidate gene studies or resequencing studies.

209
A Similarity Measure of Rare Variants Between Two Sequence Diploids
Dajun Qian (1)
(1) City of Hope

For common variants in form of single nucleotide polymorphism (SNP), genotype similarity is often measured by haplotype sharing that counts the number of identical-by-state alleles in a region, or the number of marker intervals flanking a marker. Similarity-based association analysis between genotype similarity and phenotype similarity holds a power advantage over frequency-based methods in detecting weak and noisy SNP associations complicated by allelic heterogeneity and varied causal effects. To expand the similarity-based approach, we proposed a statistic for measuring the similarity of rare variants between two sequence diploids. The performance of the similarity measurement is verified by its consistent agreement with visual judgment under a variety of scenarios.

210
Identification of Copy Number Variation in High-risk African-American Men with Prostate Cancer
Elisa M. Ledet (1), Xiaofeng Hu (2), Marilyn M. Li (2), Diptasri M. Mandal (1)
(1) Louisiana State University Health Sciences Center, New Orleans, LA
(2) Hayward Genetics Center, Tulane University School of Medicine, New Orleans, LA

Prostate cancer (PCa) is a complex multi-allelic disease and the most common malignancy in men. Disease susceptibility loci have been identified for this cancer but no definite locus-specific information is established due to genetic heterogeneity. Genomic copy number variations (CNVs) have been detected in prostate tumors and several other cancers, but changes in copy number have not been studied in germ-line DNA from hereditary African-American PCa cases. To identify PCa associated CNVs, we have recruited ten African-American families with at least 3 affected individuals. From these families, 30 individuals, including 21 affected males and 9 unaffected males, were selected for CNV analysis. Array comparative genomic hybridization was used to determine specific CNVs that may predispose African-American men to PCa. We have used a combined targeted/whole-genome array system using the Agilent 4 X44 K format; 50% of the probes were selected from the Agilent eArray system and targets known cancer-associated chromosome regions as well as over thirty reported PCa associated regions. Data was analyzed using CGH Analytics 3.5 (Agilent Technologies). Novel CNVs were identified on chromosomes 1, 13, and 20 in prostate cancer cases. These CNVs may represent a component of genetic variability which contributes to the high prevalence and mortality of PCa in African American men. Detailed analysis of our array CGH data and validation study will be presented.

211
Segmentation and Estimation for SNP Microarrays: a Bayesian Multiple Change Point Approach
Yu Chuan Tai (1), Mark N. Kvale (1), John S. Witte (1)
(1) University of California, San Francisco

High-density SNP microarrays provide a useful tool for the detection of copy number variants (CNVs).

The analysis of such large amounts of data is complicated, especially with regard to determining where copy numbers change and their corresponding values. In this work, we propose a Bayesian multiple change point model (BMCP) for segmentation and estimation of SNP microarray data. Segmentation concerns separating a chromosome into regions of equal copy number differences between the sample of interest and some reference, and involves the detection of locations of copy number difference changes.
Estimation concerns determining true copy number for each segment. Our approach not only gives posterior estimates for the parameters of interest, namely locations for copy number difference changes and true copy number estimates, but also useful confidence measures. In addition, our algorithm can segment multiple samples simultaneously, and infer both common and rare CNVs across individuals. Finally, for studies of CNVs in tumors, we incorporate an adjustment factor for signal attenuation due to tumor heterogeneity or normal contamination that can improve copy number estimates.

212
Scoring and Calling Copy Number Variants
Glen A. Satten (1), Andrew S. Allen (2), Morna Ikeda (3), Jen G. Mulle (3), Stephen T. Warren (3)
(1) Centers for Disease Control and Prevention
(2) Dept. of Biostatistics and Bioinformatics, Duke University
(3) Dept. of Human Genetics, Emory University

Background: Copy number variants (CNVs) are currently under intensive investigation. Many methods detect CNVs, but every method calls false positives. Distinguishing true calls requires experimental validation.
Methods: We developed a CNV calling algorithm based on a simple score for each called CNV. We show this score predicts the chance a CNV is experimentally validated. Based on this score, we developed a simple backward elimination algorithm for calling CNVs: starting with a jump in log-intensity ratio (LIR) at each probe (each jump is a putative CNV breakpoint), we remove jumps one at a time (while combining intensities between jumps) until all remaining called CNVs have a score larger than a user-specified cutoff. For robustness we use medians rather than means of adjacent probe intensities when combining LIRs. Genomic covariates such as GC content may be added to our scoring and CNV-calling algorithms. Our method can also be restricted to inherited CNVs in case-parent trios.
Results: We validated our score using samples from males with autism (Autism Genetic Resource Exchange) analyzed with a Nimblegen CGH array having 2,020,823 probes on the X chromosome. We experimentally validated 91 CNVs called in 41 persons (43 were validated, 48 were not). The mean score for validated CNVs was 4.5 with standard deviation (SD) 2.9. The mean score for non-validated CNVs was 2.3 with SD 1.6.

213
Genome Wide Characterization of CNVs in African Americans from HyperGEN
Nathan Wineinger (1), Nicholas Pajewski (1), Richard Kennedy (1), Mary K. Wojcznski (1), Steven C Hunt (2), C. Charles Gu (3), Ulrich Broeckel (4), D.C. Rao (3), Donna K. Arnett (1), Hemant K Tiwari (1)
(1) University of Alabama at Birmingham
(2) University of Utah
(3) Washington University in Saint Louis
(4) Medical College of Wisconsin

With the availability of the latest high-throughput commercially available genotyping arrays, the ability to accurately measure copy number variation (CNV) in population-based studies has improved dramatically. This study is the first to characterize CNVs in African Americans using one of the recent genotyping platforms specifically designed for CNV analysis (Affymetrix 6.0). We present the results on CNV data from 446 unrelated African American subjects selected from the HyperGEN cohort. The data were analyzed using three freely available calling algorithms designed for current high-resolution SNP arrays: Birdsuite, PennCNV, and VanillaICE. The results show 37 to 46 CNV segments, covering 3172 to 3189 kb per person (medians), depending on the algorithm used. In general, there were over twice as many deletions than duplications. However, the duplicated segments tended to be larger than the deleted segments. We also distinguished between rare and common CNVs and investigated the behavior of SNPs in CNV regions.

214
Assessing Copy Number Variation within a Genome-Wide Association Study
Sara Lindstrom (1), Marilyn C. Cornelis (2), Majken Jensen (2), Constance Chen (1), Frank B. Hu (2), Peter Kraft (1)
(1) Department of Epidemiology, Harvard School of Public Health
(2) Department of Nutrition, Harvard School of Public Health

The new generation of whole-genome single-nucleotide polymorphism arrays has made it possible to study copy number variants (CNVs) on a genome-wide basis. CNVs have been shown to influence human traits, but there is limited knowledge about how to accurately assess them. CNVs could either be studied by genotyping already mapped CNVs or by identifying de novo or rare CNVs to study the aggregated burden. In the first approach, the uncertainties in copy number assignment should be considered (cf. ''haplotype dosage''). The second approach includes measuring probe intensities to identify deviations that indicate a variant copy number (this is inherently sensitive to false positives). Several statistical tools for analyzing CNVs have been propposed, but there is no consensus about optimal strategies. Empirical data is sparse and several technical issues in terms of data quality analysis, signal:noise ratios and computational burden need attention. We will apply different methods for testing (Birdsuite/PLINK, CNVtools, raw intensities) and discovering (Birdseye, PennCNV) CNVs in a genome-wide association study of type 2 diabetes, In total, 3,000 cases and 3,000 controls were genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0. We will address

concordance in CNV discovery between different methods and also evaluate the association between type II diabetes and CNVs. We will also discuss the importance of quality control and computational burden.

215

Copy Number Variation in Candidate Regions in Extended Families with Autism Spectrum Disorder

Daria Salyakina (1), Holly Cukier (1), Dequiong Q. Ma (1), Antony Griswold (1), Ioanna Konidari (1), James M. Jaworski (1), Patrice Whitehead (1), Harry H. Wright (1), Ruth K. Abramson (1), Michael L. Cuccaro (1), John R. Gilbert (1), Margaret A. Pericak-Vance (1)
(1) Miami Institute for Human Genomics

Recent studies have identified several copy number variations (CNVs) that are associated with increased risk of autism spectrum disorder (ASD). Most of these findings describe rare CNVs and have not been replicated by other research groups. Recently Glessner et'al. [2009] reported novel and known CNV candidate regions for ASD.
We have conducted analysis of these validated candidate regions in a cohort of 69 extended ASD families with multiple affecteds including 117 individuals diagnosed with ASD and 354 healthy relatives. We examined if any of these CNVs account for familial ASD. We tested the hypothesis that a consistent CNV contributing to ASD risk is shared in the affecteds in the family. Samples were genotyped using the Illumina Human 1M Beadchip. We used the PennCNV algorithm for CNV calling in our cohort.
For the 26 candidate regions listed in the paper, we identified CNVs in 16 of these regions. 14 of them were inherited and two could be de novo
The most CNV rich region was on chromosome 15q11-13; however, no segregation in affected individuals within families could be shown.
One locus on chromosome 3q26.31, had a duplication overrepresented in ASD affecteds (8.5% vs. 5.3%). CNVs in remaining candidate regions also did not show segregation in affecteds with ASD. In conclusion, no one candidate region mentioned by Glessner et'al. [2009] showed segregation in multiple affected individuals within our families. These CNVs probably do not contribute to risk of ASD.

216

CNVis: Visualization and Analysis of Copy Number Data

Nathan Pankratz (1)
(1) Indiana University

The analysis of copy number (CN) variation is a relatively new field of study. I have developed CNVis, an open-source Java tool set, to visualize CN data and thereby allow investigators to compare results and easily identify problems using an expanding collection of modules. *Parse* imports raw data and exports to many file formats (PLINK, PennCNV, QuantiSNP, etc). *GenderChecks* identifies markers co-hybridizing to sex chromosomes, which can affect Log R ratios (LRRs) and B Allele Frequencies (BAFs) resulting in spurious CNV calls (~1% over-called, ~1% under-called). *Scatter* plots coordinates (i.e. X/Y, Theta/R, BAF/LRR) and can color code any user-defined variable (gender, DNA source, etc) allowing investigators to detect if an association is due to artifact. *Centroids* reclusters common CNV probes properly, allows manual re-positioning of centroids, and dynamically updates LRRs. *Trailer* plots LRR and BAF for every marker on a chromosome and overlays gene annotation and results from one or more CN calling algorithms, which is particularly useful for determining why different algorithms produce different breakpoints or CN. *Mosaic* identifies somatic mosaicism, typically a deletion or duplication of an entire chromosome, which can confound analysis when the DNA source tissue is not affected by the disease. Only with tools such as these can investigators separate true positives from false positives *in silico*. Binaries, documentation and source code at: http://cnvis.sourceforge.net/.

217

Developing Quality Control Standards for CNV analysis

Elizabeth W. Pugh (1), Hua Ling (1), Kurt Hetrick (1), Audrey H. Schnell (1)
(1) Center for Inherited Disease Research (CIDR), IGM, Johns Hopkins University SOM

Many *de novo* copy number variation (CNV) algorithms compare a sample's intensity data to a reference group mean at consecutive SNPs. A series of low or high intensity data points suggests a decrease or increase in copy number. CNV calling results in large GWAS datasets can be inconsistent. We are developing guidelines for flagging samples that perform poorly for CNV analysis.
We examined 44 HapMap subjects genotyped 2-5 times on the Illumina® 1M yielding 143 non-independent duplicate pairs. The mean genotype call rate for these samples was 99.7% (range 98.8–99.9). The genotype concordance rate was 99.98% for 63 independent pairs (range 99.868–9.999). Though all samples had high quality genotypes, their intensity variation for the autosomal SNPs and probes differs (mean standard deviation (SD) of log R ratio (LRR) 0.20, range 0.14–0.47).
cnvPartition 1.2.1 was used to call CNVs using the clustering file from a large GWAS project (defined with 96% of the samples). For the 143 pairs, the mean number () of CNVs called per sample was 39 (range 16–272). The overlap (of CNVs of the same type with at least 1bp overlap for the pair/mean of CNVs for the samples in the pair) ranged from 3% to 82%. Using a cutoff of $>.25$ for the SD of the LRR for autosomal markers flagged 23 pairs (concordance ranges 3–53%). The concordance for the remaining 120 pairs ranged from 32 to 82%). Additional QC measures may further improve performance of CNV calling.

218

Estimating the Dataset-specific Parameters for HMM in Detecting CNVs

Yaji Xu (1), Emily Lu (1), Bo Peng (1), Christopher I. Amos (1)
(1) University of Texas M.D. Anderson Cancer Center

SNP genotyping arrays have been developed to characterize single nucleotide polymorphisms (SNPs) and can also be applied to type for DNA copy number variations (CNVs). The quality of the inferences about copy number can be affected by many factors including batch effects, DNA sample preparation, signal processing, and analytical approach. Nonparametric and model-based statistical algorithms have been developed for detecting CNVs. In these methods, the generalized genotyping approach, including QuantiSNP and PennCNV which involve both log R ratio and B allele frequency, has shown advantages in recent studies. Both QuantiSNP and PennCNV implement a Hidden Markov Model (HMM) and fixed the (hyper) parameters beforehand. We have been comparing results for duplicated samples using different analytical procedures. We have found in general that PennCNV provides a more reliable assessment of copy number variation in these duplicated samples. To further investigate the behavior of CNV procedures we are developing reagents and approaches for copy number assessments for which we have known copy number changes. In one approach, we define an algorithm which mixes the chromosome X raw data from males and females to create pseudo-CNV-data as the training sample for specific dataset. The dataset-specific parameters estimated using the mixing algorithm provides more accurate results than those from fixed parameters.

219

How Much Variation is there Among Copy Number Variation Algorithms? A Case Study Using the Affymetrix SNP Array 6.0

Elizabeth J. Atkinson (1), Mariza de Andrade (1), Jeanette Eckel-Passow (1), William Bamlet (1), Martha Matsumoto (1), Sooraj Maharjan (1), Sharon Kardia (2)
(1) Mayo Clinic
(2) University of Michigan

As the number of genome wide association studies (GWAs) increase, an additional benefit is the ability to investigate copy number variation (CNV). CNV data obtained from the GWA SNP chips is a relatively new technology and there are an increasing number of software packages devoted to extracting CNV data. However, there has been no formal comparison exploring the impact that these different software packages have on the corresponding results. We extracted CNV data from the Affymetrix SNP Array 6.0 on 854 samples of hypertensive sibships from Rochester MN using PennCNV, Canary, and the R packages CRLMM and aroma.affymetrix. Variations among normalization, impact of ''bad'' samples, and choice of a reference sample were explored. A comparision of these four software packages will be discussed using the sibship data.

220

Variability in Copy Number Variation: Detection and Comparison Across Platforms

Betty Q. Doan (1), Robert Scharpf (2), Ashley O'Connor (1), Rafael Irizarry (2), Aravinda Chakravarti (1)
(1) Genetic Medicine, Johns Hopkins Medicine
(2) Biostatistics, Johns Hopkins Public Health

Multiple methods exist for calling CNVs on genome wide SNP data, but are limited for fine mapping platforms. We have developed an algorithm to call allele-specific copy number (CN) using Affymetrix Targeted Genotyping SNP microarrays and generalized it for any genotyping platform that provides raw intensities and genotypes. Because not all SNP intensity data are informative on the copy number scale, we have established a pipeline to remove such SNPs to improve data quality. Our CN estimates identified the deletion in our positive control sample, and were consistent across a series of replicates. Comparing the data across all 800 samples, we noticed a portion of samples were too noisy to be properly segmented (normal range of 1 to 3 copies). We are able to show that the variance in the copy number estimates is surprisingly a simple non-linear function of the raw intensities, and standard normalization methods could not control this variability. This observation was present in our Affymetrix 6.0 data, where several samples, which passed all genotyping QC metrics, had a high variance on the copy number scale. Consequently, we have developed CNV-specific QC metrics that assess the quality of the signal to noise ratio in the SNP and sample data. We are applying and comparing these QC metrics to regions identified on Affymetrix 6.0 with complementary data on NimbleGen 2 M and Agilent 1 M aCGH arrays and AB SOLiD next-gen sequencing technologies.

221

Copy Number Variation of the CCL3-related Chemokine Gene Cluster and Kawasaki Disease Susceptibility

Sadeep Shrestha (1), Howard W. W. (1), Aditi Shendre (1), Aaron K. Olson (2), Mary Beth Lee (2), Michael Portman (2)
(1) Department of Epidemiology, University of Alabama at Birmingham
(2) Department of Pediatrics, University of Washington, Seattle Children's Hospital

Four closely related chemokine genes, CCL3, CCL3L1, CCL3L2, and CCL3L3 are localized in a segmental duplication at chromosome 17q12. CCL3L1 is implicated in Kawasaki Disease (KD) pathogenesis. CCL3 has two gene copies in a diploid genome whereas the other genes have multiple copies. We measured total gene copy numbers (GCN) of the three genes in 85 complete trios and 72 parent-patient pairs using RT-PCR. We used a method that we recently developed to estimate child-parent GCN as a five element vector for trios where the first element is the number of trios for which the child had a lower GCN than both parents while the fifth element is where the child had a greater GCN than both parents, and a similar three element for parent-child pairs. The maximum copy number observed was 23, requiring a maximum of 12 gene copies in each chromosome. This observed and expected values, which was derived from the underlying distribution of copy numbers on individual chromosome using the EM algorithm, were significantly different for complete trios ($P = 2.93E-06$, 4 df), for single-parent-child pair ($P = 0.04$, 2 df) and for combined ($P = 7.98E-07$, df using Satterthwaite approximation) suggesting a difference in the total GCN distribution between parents and patients. However, assays for GCN of three

genes separately are warranted to determine whether this association is from he combined GCN effect or from individual genes.

222

Single Locus Bayesian Modelling of Nondisjunction in Trisomy 21

Nokuthaba Sibanda (1)
(1) Victoria University of Wellington

Meiotic nondisjunction leads to aneuploid gametes, which may result in trisomy in fertilised cells. Trisomy of chromosome 21 causes Down Syndrome and is one of the most commonly studied genetic disorders. A Bayesian approach combined with Markov Chain Monte Carlo was used to develop a probability model for estimating relative frequencies of the parent and meiotic stage of nondisjunction origin. Genotype data for both parents and their trisomy 21 child at a single locus on chromosome 21 were used. Two ways of determining the likelihood were compared: a direct approach and an augmented data likelihood approach. The probability model was applied to simulated and to real data at the D21S11 and EcoRI loci. For simulated data, the direct likelihood approach gave better point estimates with narrower 95% posterior credible intervals. Smaller relative frequencies had the worst estimates and a larger number of alleles resulted in improved estimate precision. Estimates from real data were compared to published results. Point estimates from the augmented data likelihood approach were closer to published frequencies, particularly for the more centromeric D21S11 locus. This study shows that a statistical modelling approach based on data from a single locus can be used to estimate relative frequencies of parent and meiotic stage of the origin of nondisjunction.

223

Hypocholesterolemic Autism: Follow-up Association Studies on the Linkage Regions of Chromosomes 5 and 10

Yoonhee Kim (1), Elaine Tierney (2), Forbe D Porter (3), Eli DO Roberson (4), Joan E Bailey-Wilson (1)
(1) National Human Genome Research Institute, NIH, Baltimore, MD
(2) Kennedy Krieger Institute, Baltimore MD
(3) National Institute of Child Health and Human Development, NIH, Bethesda, MD
(4) Johns Hopkins University School of Medicine, Baltimore, MD

Previously we found 2 nominally significant linkage regions (5q11.1–11.2 and 10p14) in the Autism Genetic Resource Exchange (AGRE) data using 390 microsatellites under the hypothesis that hypocholesterolemic autism may have different genetic susceptibility loci. We performed follow-up association studies on chromosomes 5 (ch5, 45–59 Mbp, 1552 SNPs) and 10 (ch10, 4–14 Mbp, 2248 SNPs) using high-density SNPs (average spacing 7.1 kb) from the 500 K Affy panel provided by AGRE. We defined 2 subgroups: probands' cholesterol levels were 1) >2 SD (standard deviation) below the mean (47 families), and 2) less than 100 mg/dl (28 families). In each group we analyzed different subsets depending on disease diagnosis: 1) both sibs had autism or 2) one or more sibs had "Not Quite Autism" or "Broad Spectrum" using family based association TDT tests in PLINK 1.04. We narrowed the interval into 5 Mbps (53~58 Mbp) on ch5 with 7 nominally significant SNPs (<0.0001 P value) and 3 Mbps (5.3~8.4 Mbp) on ch10 with 2 nominally significant SNPs. Even though we cannot conclude that these SNPs have significant evidence for association after correcting for multiple testing, genes near these loci are intriguing: interleukin related genes (IL2RA and IL6ST), PDE4D, and GATA3. These findings suggest that this clinical subset may have different genetic risk factors than other autistic children and that further analyses with more samples of hypocholesterolemic ASD families are required.

224

A Multi-stage/multi-design Strategy Identifies a New QTL for VWF on Chromosome 6 — a Possible Link with Venous ThromboEmbolism (VTE)?

Guillemette Antoni (1), Noemie Saut (2), Yiqiang Luo (3), Gwenaelle Burgos (2), Christine Biron-Andreani (4), Jean-François Schved (4), Gilles Pernod (5), Irene Juhan-Vague (2), Marie-Christine Alessi (2), Ludovic Drouet (6), Sophie Visvikis-Siest (7), Philip Wells (8), Joseph Emmerich (9), David-Alexandre Tregouet (6), Pierre-Emmanuel Morange (2)
(1) INSERM UMR_S 937, UPMC Univ Paris 06, France; Dalla Lana School of Public Health, University of Toronto, Canada
(2) INSERM, UMR_S 626, Université de la Méditerranée, Marseille, France
(3) Dalla Lana School of Public Health, University of Toronto, Canada
(4) Laboratoire d'Hématologie, CHU Montpellier, France
(5) Service de Médecine Vasculaire, CHU Grenoble, France
(6) INSERM UMR_S 937, UPMC Univ Paris 06, France;
(7) Equipe INSERM "Génétique Cardiovasculaire" CIC 9501, Nancy, France
(8) Ottawa Health Research Institute, Ottawa, Canada
(9) INSERM U765,Université Paris-Descartes, France

A multi-stage/multi-design strategy was used to identify new loci that would contribute to VTE susceptibility by modulating FVIII and/or VWF levels, two known quantitative risk factors for VTE. A pedigree linkage analysis was first performed on 5 extended French-Canadian families including 261 individuals genotyped for 1079 microsatellites and identified 4 putative regions linked to FVIII and vWF levels. These regions, located on chromosomes 2, 6, 9 and 12, were then evaluated by in *silico* association analysis of published GWAS data on VTE with the aim of narrowing the linkage signals by focusing on candidate regions for VTE. Four SNPs in the chromosome 6 region were in particular *in silico* associated with VTE at $P<10^{-4}$ and were then investigated for association with FVIII and VWF in a sample of 123 healthy French nuclear families. One SNP was associated with VWF levels ($P=0.0018$) in these families and also associated with VWF ($P=0.009$) in an independent sample of 823 patients with VTE at young age (<50 years). Besides, the allele associated with

decreased VWF levels tended to be associated with a lower risk of VTE (OR = 0.70 [0.47–1.04], $P = 0.081$) in an independent French case-control study for VTE of 607 cases and 607 controls.
In conclusion, strong evidence for a new QTL for VWF levels on chromosome 6 was obtained using a multi-stage approach including pedigree linkage analysis, association analysis in nuclear families and case-control data.

225
Identification of the Biologically Relevant HLA DR-DQ Amino Acids in Type 1 Diabetes (T1D)
Glenys Thomson (1), David R. Karp (2), Nishanth Marthandan (2), Paula A. Guidry (2), Steven J. Mack (3), Richard M. Single (4), Ana M. Valdes (5), Richard H. Scheuermann (2), Wolfgang Helmberg (6), T1D Genetics Consortium (7)
(1) Univ California
(2) UT Southwestern Medical Center
(3) Children's Hospital Oakland Research Institute
(4) Univ Vermont
(5) King's College, London
(6) Univ Graz, Austria
(7) Consortium

The immune response HLA class II DRB1-DQB1 genes are the major T1D genetic susceptibility loci, with a hierarchy of haplotype and genotype effects from very predisposing to very protective. With T1D and the many other HLA associated diseases, it is difficult to identify the combinations of biologically relevant amino acids directly involved in disease given the high level of HLA polymorphism and the pattern of amino acid variability, including varying degrees of linkage disequilibrium. Using a suite of complementary methods, we have analyzed ~1400 Caucasian pedigrees with high resolution HLA DR-DQ typing from the T1D genetics consortium. As well as standard analyses, we applied two new analytic tools which were very informative: an asymmetric measure of LD which more accurately detects the correlation between amino acid sites than standard LD measures, and the sequence feature variant type (SFVT) method. With SFVT analyses, association tests are performed on variation at biologically relevant SFs based on structural (e.g. beta-strand 1) and functional (e.g. peptide binding) features of the protein, and combinations thereof. We demonstrated that amino acid variation within the peptide binding sites (PBSs) of the HLA DRB1 and DQB1 proteins explains the complex DRB1-DQB1 haplotype associations with T1D. Further, for DQB1, variation in pocket 9 of the PBS is the main contributor to T1D risk, whereas for DRB1 variation covering a number of pockets in the PBS is required.

226
Azathioprine Induced Severe Pancytopenia in One Patient with Crohn's Disease: Identification of a Novel Thiopurine S-Methyltransferase Allelic Variant IVS8-1G > A
Tiphaine Adam de Beaumais (1), May Fakhoury (1), Benedicte Pigneur (2), Sheila Viola (2), Yves Medard (1), Franck Broly (3), Evelyne Jacqz-Aigrain (1)
(1) Robert Debré hospital
(2) Trousseau hospital
(3) CHRU Lille

The thiopurine S-methyltransferase (TPMT) polymorphisms are a major factor responsible for large individual variations in thiopurine toxicity due to excessive accumulation of cytotoxic metabolites. This present clinical observation describes a 14-year-old girl with Crohn's disease who developed severe pancytopenia during her Azathioprine (AZA) treatment. Viral infections were excluded. Red blood cell (RBC) thioguanine nucleotides (6-TGN) and 6-methylmercaptopurine ribonucleotides (6-MMP) concentrations were measured by high-performance liquid chromatography. The open reading frame of TPMT gene (exons 3 to 10) and their consensus flanking sequences were sequenced (Applied Biosystems 3130 XL). Genotyping for the three predominant TPMT mutations showed a heterozygous TPMT*2 genotype but 6-TGN levels were unexpectedly elevated (3414 pmol/$8*10^8$ RBC) and the concentration of 6-MMP was inferior to 20 pmol/$8*10^8$ RBC. Familial genetic analysis showed that the patient was carrier of a heterozygous composite TPMT genotype for two non-functional mutations. We identified a novel TPMT variant allele IVS8-1G > A corresponding to a G > A transition at the splice acceptor site of the intron 8 probably responsible of TPMT deficiency. In cases of TPMT phenotype/genotype discordance concomitant with the occurrence of a severe adverse effect, sequence analysis of the complete open reading frame of the gene is interesting to identify rare inactivating variants.

227
Protein Kinase C ? (PRKCH) Gene Polymorphism is Associated with Sero-negative Severe Gastric Atrophy
Yasuyuki Goto (1), Asahi Hishida (1), Keitaro Matuo (2), Kazuo Tajima (2), Emi Morita (1), Mariko Naito (1), Kenji Wakai (1), Nobuyuki Hamajima (1)
(1) Department of Preventive Medicine/Biostatistics and Medical Decision Making, Nagoya University Graduate School of Medicine
(2) Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan

This study aimed to investigate the associations of the risks of gastric atrophy and gastric cancer with a *G/A* single nucleotide polymorphism (rs3783799) of *PRKCH* gene, which encodes the ? isozyme of protein kinase C (PKC ?). The gene is supposed to affect the development of inflammation through iNOS and cell proliferation through ERK/Elk-1 and Akt pathway. The subjects consisted of 583 cases from first-visit outpatients at Aichi Cancer Center Hospital aged 27 to 80 years, who were diagnosed as gastric cancer from 2001 to 2005, and 1,742 controls frequency-matched for age and sex. 1638 controls were available for measurement of both anti-H. pylori IgG antibody and pepsinogens (PGs). Among them, 57.3% were sero-positive and 33.0% had gastric atrophy (PG1 < 70 ng/dl and PG1/PG2 < 3). When compared to healthy controls who were not infected with H. pylori or did not have gastric atrophy, the *AA* genotype was significantly associated with sero-negative severe atrophy

(PG1 < 30 ng/dl and PG1/PG2 < 2); OR = 4.81, and 95% confidence interval, 1.15–20.1, relative to the *GG* genotype. The genetic polymorphism was not associated with the risk of gastric cancer. This was the first study to examine the associations of the *PRKCH* polymorphism with gastric atrophy and gastric cancer, suggesting that the *AA* genotype may be a higher risk for sero-negative severe gastric atrophy.

228

Analysis of Overdispersion Patterns in the LOH of Biallelic Markers in a Study into the Polyclonal Origin of Multiple Sporadic Basal Cell Carcinomas.

Franz Quehenberger (1), Ellen Heitzer (2), Peter Wolf (3)
(1) Department of Medical Informatics, Statistics and Documentation, Medical University of Graz
(2) Department of internal medicine, Medical University of Graz
(3) Department of Dermatology, Medical University of Graz

Introduction: There have been reports that not only basal cell carcinomas are of monoclonal origin, but also multiple lesions within a patient [Shulman et'al., 2006]. They found identical LOH patterns near the PTCH gene. In our study, however, by comparing the mutation patterns and LOH at the PTCH gene, we found no case which would have suggested monoclonal origin of lesions [Heitzer et'al., 2009]. Here we present the conclusions from the analysis of overdispersion from that study.
Methods: Tissue from three basal cell carcinomas and normal skin from each of six patients was genotyped at six SNPs near the PTCH gene. Overdispersion or clustering is observed if LOH occurs more often within the same patient than by chance alone or, if LOH has occurred, the same allele is lost more often than expected. The permutation test based on score statistic of Tarone [1979] was used to test for overdispersion.
Results: There was no significant overdispersion in the occurrence of LOH, but significant overdispersion in the allele that was lost.
Conclusions: Although the study included only a small number of subjects, a small number of lesions per subject and not all of the SNPs were informative, statistically significant overdispersion could be established. If a part of a chromosome is lost, there is a strong tendency that it occurs at the same chromosome on different lesions. Loss of the same allele does not imply monoclonal origin of multiple basal cell carcinoma.

229

Variation of Breast Cancer Risk in the French National BRCA1/2 Carrier Cohort (GENEPSO)

Julie Lecarpentier (1), Catherine Noguès (2), Emmanuelle Mouret-Fourme (2), GENEPSO French National BRCA1/2 carrier cohort (3), Rosette Lidereau (4), Nadine Andrieu (1)
(1) Institut National de la Santé et de la Recherche U900 - Mines parisTech - Institut Curie - Paris - France
(2) Centre René Huguenin - Saint Cloud - France
(3) Groupe Génétique et Cancer - FNCLCC - Paris - France
(4) Inserm U735 - Laboratoire d'Oncogénétique - Centre René Huguenin - Saint-Cloud - France.

Germline mutations in *BRCA1* and *BRCA2* confer high risk of breast cancer (BC), but the magnitude of this risk varies according to different factors. Although controversial, there are data to support the hypothesis of allelic risk heterogeneity.
We assessed variation in BC risk according to location of mutations from the study GENEPSO. Since the women in this study were selected from high-risk families, an oversampling of affected women was eliminated by using a weighted Cox-regression model. Women were censored at the date of diagnosis when affected by any cancer, or the date of interview, when unaffected. 990 women were selected for this analysis (379 affected, 611 unaffected).
BRCA1 and *BRCA2* were divided in 20 regions corresponding to the vigiciles among unaffected women. The adjacent regions with similar hazard-ratio (HR) were combined. For *BRCA1*, there was some evidence of a lower risk of BC when the mutation was located between codon 374 and 1161 (HR:0.65, $P = 0.037$) as compared with mutations outside this region. For *BRCA2*, there was strong evidence for a region at decreased risk (codons 957 to 1827) (HR = 0.33, $P = 0.004$) and for one at increased risk (codons 2546 to 2968) (HR = 3.37, $P = 0.003$).
Our findings are consistent with those suggesting that central region of *BRCA1/2* are at lower risk of BC. However, borders of these regions varied according to studies and the true variation may be more complicated and involved also ''environmental'' factors.

230

State of the Genetic Disorders in the Pakistani Population

Shahid Mahmood Baig (1)
(1) National Institute for Biotechnology and Genetic Engineering (NIBGE)

The population of Pakistan is 170 million with highest rate of consanguineous marriages in the world (>60%). At National Institute for Biotechnology and Genetic Engineering (NIBGE) more than 700 β-thalassemia families with at least one affected child were analyzed for mutation screening and/or prenatal diagnosis and 400 large consanguineous families with at least 3–4 affected births with rare monogenic disorders were studied for linkage analysis. β-Thalassemia is the most common genetic disorder with carrier frequency of more than 5.6%. Of the 700 β-thalassemia families analyzed, 90% (630) were consanguineous. In the last 15 years, less than 3,000 prenatal diagnoses have been carried out in Pakistan whereas approximately 6,000 new thalassemia major children are born annually in this country. Our data shows that 70–80% of these families have to continue the tradition of consanguineous marriages due to various social and cultural reasons. More than 90% of these inbred families are willing to have the carrier screening and prenatal diagnosis to prevent affected births but unable due to cost or unavailability of tests. NIBGE is already providing free prenatal diagnosis of β-thalassemia to a significant section of the population. There is genuine need to establish nationwide, accessible and cost effective preventive programs to control the genetic diseases in Pakistan.

231

A Bayesian Adjustment for an Ascertainment Bias in Human Genetics

Hongyan Xu (1), Jai W. Choi (1), Balgobin Nandram (2)
(1) Medical College of Georgia
(2) Worcester Polytechnic Institute

When there is a rare disease in a population, it is inefficient to take a random sample to estimate a parameter. Instead one takes a random sample of all nuclear families with the disease by ascertaining at least one sibling (proband) of each family. In these studies, if the ascertainment bias is ignored, an estimate of the proportion of siblings with the disease will be inflated. The problem arises in human genetics, and it is analogous to the well known selection bias problem in survey sampling. For example, studies of the issue of whether a rare disease shows an autosomal recessive pattern of inheritance, where the Mendelian segregation ratios are of interest, have been investigated for several decades, and corrections have made for the ascertainment bias using maximum likelihood estimation. Here, we develop a Bayesian analysis to estimate the segregation ratio in nuclear families when there is an ascertainment bias. We consider the situation in which the proband probabilities are allowed to vary with the number of affected siblings, and we investigate the effect of familial correlation among siblings within the same family. Simulation results show that the Bayesain approach can overcome the difficulties associated with maximum likelihood estimation for the ascertainment bias problem. We discuss an example on cystic fibrosis and a simulation study to assess the effect of the familial correlation.

232

A Comparison of Methods for Simulating a Gene Region with a Specified LD Structure

Audrey E. Hendricks (1), Richard H. Myers (2), Kathryn L. Lunetta (1)
(1) Boston University School of Public Health
(2) Boston University School of Medicine

The foundation of genetic association studies as well as the basis for most of the new complex analyses that use Genome Wide Association (GWA) data presumes the ability to detect genetic risk variants through linkage disequilibrium (LD) with genotyped markers. Thus, accurately modeling LD in simulations is essential to correctly evaluate the power and type I error of new methods. Here we compare an extension of the method used by de Bakker et'al. [2005] to the method used by Li and Stephens [2003] in simulating LD for a specific gene region using 430 phased haplotypes from the HapMap. Both methods sample from a set of phased haplotypes. Further, Li and Stephens use a Markov chain to recombine portions of the region across various haplotypes. We compare this to an extension of de Bakker et'al.'s method where recombination is simulated within the set of haplotypes by applying the recombination rate estimated by the HapMap project using McVean et'al.'s [2004] coalescent method. We evaluate the ability of each of these methods as well as the impact of different recombination levels on the ability to accurately simulate the LD for a specific gene region. We compare the HapMap sample used for the simulation to the simulated data itself with respect to the pairwise LD distribution and the SNP Hardy Weinberg Equilibrium test statistics and minor allele frequencies. For both simulation models we find that overall LD decreases as the overall probability of recombination increases.

233

Celestial3D: Designing a Tool for 3D Visualization of Familial Data from a Bioinformatics Perspective

Eric Lam (1), Rebecca J. Webster (1), Jon Emery (2), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, The University of Western Australia
(2) School of Primary, Aboriginal and Rural Health Care, The University of Western Australia

Whilst software packages already exist to facilitate the visualization of pedigrees, many if not most will follow the traditional approach for visualizing the pedigree in two-dimensions (2D). The Celestial3D project started as a proof-of-concept for arranging pedigrees in three-dimensional (3D) space using X3D with the primary purpose of improving the visualization and analysis of complex data for large/complex pedigrees over that of the typical 2D format. However, the requirements for Celestial3D have since grown significantly with overall plans for it to serve three major applications: genetics education, epidemiological/pedigree research and disease risk assessment in a range of clinical settings. Here we present our approach to the design and implementation of such a tool. The data structures and algorithms have been designed for a great amount of flexibility to support interoperability across many different datasets and future pedigree editing/building features. This aspect is prominently demonstrated where the design allows many arbitrary attributes to be tagged to each individual. Additionally, common requirements such as enhanced viewing capabilities with searching/filtering functionality and pedigree editing support have led us to adopt a 3D games engine (Unity), which provides data abstraction advantages and opportunities to incorporate new user-interactivity not previously possible.

234

Improved Combined Family and Case-control Haplotype Analyses

Ryan Abo (1), Nicola J. Camp (1)
(1) University of Utah

Methods to exploit combinations of genetic data from mixed resources for increased power have recently been sought. Here we describe an improved algorithm for the analysis of mixed case-control and family data and an improved method to perform haplotype analyses. The new method considers non-transmitted alleles ('pseudo' controls) in place of parental genotypes (explicit controls) when combining family and unrelated samples. For better haplotype analyses, we implemented a phasing algorithm which allows a mixture of family and unrelated samples.

To test the new method, we simulated data under the null and alternative hypotheses for TDT/case-control (TDTCC) and affected-sib-pair/case-control (ASPCC) mixed resource types. A set of 500 cases/controls was combined with 500 case-parent trios for the TDTCC set and with 250 affected sibpairs with parents for the ASPCC set. A ten-locus haplotype was designed to tag a single disease variant in the alternative hypothesis. The genetic models considered included varying minor allele frequencies (6–17%) and relative risks (1.2–3.0) with a 5% sporadic rate.
The new method achieved consistently higher power over the explicit control method across the genetic models with both datasets. There were correct type I error rates using both methods and resources. These results illustrate the validity of pseudo controls with our improved haplotype analyses and demonstrate the new algorithm is more powerful than the use of explicit controls.

235
Strong Bias for P-falciparum Parasites Carrying the Wild-type PfCRT Allele in the Placenta
Heiko Becher (1), Nadja Oster (1), Petra Rohrbach (1), Cecilia Sanchez (1), Boubacar Coulibaly (2), Gabriele Stieglbauer (1), Michael Lanzer (1)
(1) University of Heidelberg, Germany
(2) Centre de Recherche en Santé à Nouna, Burkina Faso

Resistance to chloroquine, the former standard treatment for malaria, has been linked to polymorphisms within the pfcrt gene of the human malarial parasite *P. falciparum*. The prevalence of the pfcrt K76T polymorphism gene has shown to be correlated with chloroquine use in a population. Because of increasing treatment failure rates, chloroquine has been replaced by artemisinin-based combination therapies as recommended treatment.
We performed a study with 27 pregnant women from a maternity ward aged 17–34 and 20 men aged 16–38 years in Nouna, Burkina Faso/West Africa who were found to be infected with *P. falciparum*. The distribution of the pfcrt K76T and the pfmdr1N86Y polymorphism in the parasite from the peripheral blood (both sexes) and from the placenta was investigated to show whether the placenta provides a better microenvironment for the wild-type allele.
We found an unequal distribution of pfcrt haplotypes in the parasite from the peripheral blood and from the placenta (48% versus 15%, $P<0.03$) of women diagnosed with a *P. falciparum* infection. No difference was seen between sexes for the pfcrt haplotypes. In comparison, no differences were seen with regard to pfmdr1 polymorphisms.
Our data suggest a selective disadvantage of the polymorphic and a selective advantage of the wild-type pfcrt haplotype in the placenta, supporting a model in which the human host provides various microenvironments that favor genetically distinct P. falciparum populations.

236
Disentangling Genetic, Prenatal and Postnatal Environmental Effects
Jin J. Zhou (1), Suzanne Pelka (2), Kenneth Lange (3), Christina G.S. Palmer (4), Janet S. Sinsheimer (5)
(1) Departments of Biomathematics, UCLA
(2) Center for Society and Genetics, UCLA
(3) Departments of Biomathematics, Human Genetics, UCLA
(4) Departments of Psychiatry and Biobehavioral Sciences, Human Genetics, UCLA
(5) Departments of Biomathematics, Human Genetics, Biostatistics, UCLA

Maternal behaviors, environment and genetics combine to affect fetal development and produce human phenotypic variability. These combined effects lead to trait variability in adults as well as neonates. However trait heritability is misattributed if inherited genetic effects and prenatal effects are conflated. Disentangling inherited effects from prenatal or postnatal effects is possible in animals using prenatal (in-vitro fertilization) and postnatal (adoption) cross fostering designs. In humans, studying individuals conceived by assisted reproductive technologies (ART) and their families may provide a way to disentangle these effects. However, researchers lack a refined study design combined with an appropriate analysis.
We describe a novel approach using a Genetic-Gestation study design and an analytical method for disentangling inherited effects from prenatal or postnatal effects. The analytical method, an adaptation of variance components, can be applied to analyze data from an arbitrary mixture of traditional and ART families. We implement our approach by modifying the Polygenic QTL option of Mendel. Besides inherited effects, we estimate prenatal and postnatal effects. We present data on the statistical properties of our modified variance component approach based on carefully constructed Genetic-Gestation study design simulations.

237
QT Interval Duration in the Jackson Heart Study, a Segregation Analysis
Sarah G. Buxbaum (1), Sara Tribune (2), Lynette Ekunwe (1)
(1) Jackson Heart Study/Jackson State University
(2) Jackson Heart Study/Tougaloo College

The QT interval, measured in electrocardiograms has been shown to be heritable, with an estimate of 41% in the Jackson Heart Study. We followed this work with a segregation analysis of the QT interval, and showed suggestive evidence of Mendelian segregation of a major gene underlying this trait. A Box-Cox transformation was applied while simultaneously estimating the parameters of the model using the SEGREG program in the S.A.G.E. package. An environmental model with two means was significantly better than one mean ($P = 4.3 \times 10^{-7}$). The most parsimonious genetic model that best fitted the data was a dominant Mendelian model with an allele frequency of 0.18; however, it was not significantly different from the two mean environmental model. A three mean genetic model was not quite significantly different from this environmental model ($P = 0.07$). There was no significant residual spousal correlation. Additionally, the two mean

models, both environmental and genetic, had significant and equal residual familial correlations of parent-offspring and sibs, which may be interpreted as a polygenic effect, or as a shared family environmental effect.
Findings of polygenic effects (multiple loci with small effects) have previously been reported in genome wide association studies. Segregation analysis allowing for multiple loci with major effects is underway, using the lm_twoqtl program in MORGAN and SEGREG in S.A.G.E..

238

Evaluation of Genotype-specific Age Patterns of Hazard Rates Using Joint Analysis of Genetic and Non-genetic Subsamples of Longitudinal Data

Konstantin G. Arbeev (1), Svetlana V. Ukraintseva (1), Liubov S. Arbeeva (1), Alexander M. Kulminski (1), Igor Akushevich (1), Anatoli I. Yashin (1)
(1) Duke University

Comparison of age patterns of hazard rates for carriers of selected alleles/genotypes evaluated under various conditions can help better understand the role of genetic factors in risks of death and development of aging-related diseases and disability. The sample size of genetic data is often a limiting factor for the desirable accuracy of analyses of genetic effects on age patterns of the hazard rates. Longitudinal data provide an additional reserve for increasing the accuracy of respective estimates as they contain information on survival (or disease-free life span) for the participants of the study for whom genetic data were not collected. We recently proposed a method for joint analyses of genetic (i.e. genotype-specific hazard rates and cross-sectional data on genotype frequencies) and non-genetic (hazard rates in the entire sample) subsamples of longitudinal data, which is extended to accommodate possible trends in initial frequencies of genotypes and hazard rates (in cohort and cross-sectional designs). We performed simulation studies using data sets structurally similar to the Framingham Heart Study data (the original cohort, FHS-C, and the Offspring study, FOS) that showed a substantial increase in the accuracy of estimates in joint analyses of genetic and non-genetic subsamples compared to analyses based on genetic subsample alone, in different scenarios. The approach is illustrated by application to the FHS-C/FOS data on APOE and ACE D/I polymorphisms.

239

Parent-of-origin Effects in Lynch Syndrome

Christine M. van Vliet (1), James G. Dowty (2)
(1) Department of Pathology, University of New South Wales
(2) Centre for Molecular, Environmental, Genetic, Analytic Epidemiology, The University of Melbourne

Anticipation in Lynch Syndrome (defined by the presence of a germline mutation in a mismatch repair gene) has been contentious since it was first reported. Parent-of-origin effects (POE), in which disease risks depend on the sex of the parent from whom a genetic variant is inherited, often accompany anticipation and are less subject to statistical biases. We tested for a POE in Lynch Syndrome using a novel method which avoids the limitations of previous approaches.
Our study was based on 17 families from the Victorian Colorectal Cancer Family Study. Each pedigree was ascertained via a proband who had Lynch Syndrome and was diagnosed with primary colorectal cancer before age 45 years. Cancer histories and blood for mutation testing were obtained by sequential ascertainment of relatives. Risks for carriers were estimated using all relatives, whether genotyped or not, by modified segregation analyses in which conditional likelihoods (given ascertainment) were maximized. The POE was quantified as a hazard ratio (HR), being the ratio of cancer incidence for carriers who inherited their mutations maternally to the corresponding paternal incidence.
We found that maternally-inherited mutations were more pathogenic than paternally-inherited mutations for both colorectal cancer (HR 2.5, 95% CI 1.6–3.9) and extra-colonic Lynch Syndrome cancers (HR 2.9, 95% CI 1.7–4.9). If confirmed, these findings will have important biological and clinical implications.

240

Polymorphisms in the Angiotensin-Converting Enzyme (ACE) Gene as Determinants of Cardiovascular Diseases Risks and Short Life in Framingham Heart Study

Alexander Kulminski (1), Svetlana Ukraintseva (1), Konstantin Arbeev (1), Irina Culminskaya (1), Anatoli Yashin (1)
(1) Duke University, Center for Population Health and Aging

We have investigated the effects of eight rarely-studied SNPs of the ACE gene, which are not in strong linkage disequilibrium either with each other or with the I/D polymorphism, on incidence of CVD and on chances to live short life in a sample of 1755 genotyped participants (mean age = 36.2 and standard deviation = 9.6 years at baseline) of the Framingham Heart Study Offspring cohort followed for 35 years. The Cox regression analysis of the risks of CVD for each SNP (adjusted for age, sex, blood pressure, diabetes, cholesterol, BMI, and smoking) revealed significant genetic effects for four SNPs (rs4305; rs4363; rs9896208; rs12449782). Two SNPs (rs4329; rs7221678) showed marginal significance ($0.05 < P < 0.1$). The effects of rs4305 ($P = 0.004$), rs4329 ($P = 0.001$), rs4363 ($P = 0.010$), and rs12449782 ($P = 0.014$) SNPs on risks of CVD and short life were significant in multivariate analysis that included all six SNPs. The logistic regression analysis of the associations of four CVD-predictive SNPs with short life (defined as dying before age 76 years, i.e., as 20% of the shortest-lived individuals in this sample vs. those who survived the cut-off age) revealed significant impact of rs4305 ($P = 0.005$) on chances to reach the old age in the model with the same adjustments as in the case of CVD. Additional adjustment by CVD did not diminish the effect of rs4305 ($P = 0.006$) SNP on survival. Possible systemic role of the ACE gene in healthy aging is discussed.

241

Genetic Mechanisms for Venous Thromboembolism (VTE)

John A Heit (1), Petterson Tanya (1), Armasu Sebastian (1), Elysia Jeavons (1), Joseph Larson (1), Julie M Cunningham (1), Mariza de Andrade (1)
(1) Mayo Clinic College of Medicine

VTE consists of deep vein thrombosis and its complication, pulmonary embolism. We conducted a candidate gene association study from 780 genes in four pathways (anticoagulant, procoagulant, fibrinolytic and innate immunity) relevant to the pathogenesis of VTE using a haplotype tagging algorithm that incorporated Illumina's design score for iSelect. A total of 12,313 SNPs and 2070 unique individuals (49% VTE subjects; 48% males) were available for the analysis. The analyses were performed with PLINK using an additive genetic model and adjusted for age, sex, state of residence, and prior myocardial infarction or stroke (yes/no). The most significant results included two SNPs in the F5 (procoagulant Factor V) gene including the Factor V Leiden mutation, five SNPs in the ABO gene including blood group type, and one SNP in the F2 (procoagulant prothrombin) gene. We also performed an analysis stratified on sex, F5 carrier, F2 carrier, and blood type O status (yes/no). Significant results identified for F5 non-carriers mutation were SNPs in the ABO, F2, and IL4R genes; for F2 non-carriers mutation SNPs in ABO and F2 genes, and for non-O blood type carriers SNPs in the F5 and PLSCR1 genes. In conclusion, SNPs within the F5, F2, and ABO genes are susceptibility variants for VTE, and the joint population-attributable risk of the most significant risk alleles from F5, ABO and F2 was 0.21.

242

A Bias Correction of False-negative Outcomes in Tumor Characterization Studies

Cyril Rakovski (1), Daniel J. Weisenberger (2), Paul Marjoram (2), Peter W. Laird (2), Kimberly D. Siegmund (2)
(1) Chapman University
(2) University of Southern California

Etiologic studies increasingly use tumor characteristics such as DNA methylation and sequence mutation to sub classify cancers of a given organ. An issue that is typically ignored in measuring tumor characteristics is the sampling error due to the quantity of input DNA. As the amount of input DNA decreases, the ability to detect targeted molecular features within a tumor decreases, which can result in false-negative outcomes. We propose to use approximate Bayesian computation, a simulation-based approach, to incorporate into the lab measurement estimation process uncertainty that arises when sampling small quantities of DNA fragments. We find that by using information on DNA quantity and simulating the DNA sampling, we are able to estimate a posterior distribution for the lab measurement that captures sampling variation. A summary from this posterior distribution (e.g. mode, probability of exceeding a threshold) is a better estimate of the true tumor characteristic than a summary based on the lab measurement, which represents only a single realization from the sampling process.

243

Identification of KCNQ5 Gene Variants Associated with Lung Function Using a Two-stage Multi-marker Strategy

Hugues Aschard (1), Emmanuelle Bouzigon (1), Hélène Tharrault (1), Marie-Hélène Dizier (1), Regis Matran (2), Mark Lathrop (3), Francine Kauffmann (4), Florence Demenais (1)
(1) Inserm U946, Paris, France
(2) Univ Lille Nord de France, Lille, France
(3) CEA-CNG, Evry, France
(4) Inserm U780, Villejuif, France

A previous genome-wide linkage scan conducted in French EGEA families detected linkage of 6q14 to a measure of lung function (FEV1), with a higher signal found in adult offspring. We investigated further this region by genotyping a panel of 399 SNPs (spanning 30 Mb) in 203 EGEA families (337 adult offspring). To reduce the problem of multiple testing, we used a two-stage approach based on two multi-marker methods, the Local Score, and multi-marker FBAT (FBAT-M). This strategy allows detection of sets of adjacent markers showing aggregation of high statistical scores while taking into account linkage disequilibrium between markers. We identified five marker sets associated with FEV1 (P-values from 0.005 to 0.0008) that remained significant after Bonferroni's correction. The most significant marker set was located within KCNQ5 gene and included 11 SNPs of which four were strongly associated with FEV1 ($P \leq 0.0007$). These associations were replicated in a set of 267 adult French controls (P-values from 0.03 to 0.002 for four SNPs). The combination of P-values from our two samples using Fisher's combined probability test, enhanced the evidence for association of FEV1 with three SNPs (combined P-values from 6×10^{-4} to 5×10^{-5}). KCNQ5, which is expressed in bronchial epithelium and is a determinant of airway-surface liquid, is a strong candidate for lung function regulation.

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

244

A Multivariate Growth Curve Model for Ranking Genes in Replicated Time Course Microarray Experiments

Jemila S. Hamid (1), Joseph Beyene (2)
(1) Hospital for Sick Children
(2) Hospital for Sick Children and University of Toronto

The Growth Curve Model is used for ranking genes and estimating average gene expression profiles in replicated time course microarray data. The approach takes the within individual correlation as well as the temporal ordering into consideration. Moreover, time is included as a continuous variable in the model to account for the temporal pattern. Polynomial profiles are assumed to describe the time dependence and a transformation that incorporates information across all genes is applied. A moderated likelihood ratio test is then applied to the transformed data to get a statistic for ranking genes according to the difference in expression profiles among biological groups. The methodology is discussed in a general setup and could be used for one sample as well as more than one sample problems. The estimation is done in a multivariate framework in which information from all the groups involved is used for better inference. Moreover,

the within individual correlation as well as information across genes entered in the estimation through a moderated covariance matrix. We assess the performance of our method using simulation and illustrate the results with real time course microarray data taken from clustered human lung. Simulation results show that our approach performs uniformly b etter than the MANOVA approach. For the illustrative data, our method identified important genes that are previously known to associated with lung cancer and some novel genes are also identified.

245

Search for Association Between Allergic Diseases and 11p14 Genetic Variants: Indication of Association with NELL1 Gene

Marie-Hélène Dizier (1), Michel Guilloud-Bataille (2), Patricia Jeannin (1), Isabella Annesi-Maesano (3), Jocelyne Just (4), Francine Kauffmann (5), Mark Lathrop (6), Emmanuelle Bouzigon (1), Florence Demenais (1)
(1) INSERM UMRS946, Univ Paris Diderot
(2) INSERM UMR-S535, Univ Paris Sud
(3) INSERM UMR-S 707, Paris
(4) Höpital Trousseau, Paris
(5) INSERM UMR-S780, Univ Paris Sud
(6) CNG, CEA, Evry; CEPH, Paris

A previous genome-wide linkage scan conducted in French EGEA families (Epidemiological study on the Genetics and Environment of Asthma) detected the 11p14 region linked to three allergic diseases: asthma, atopic dermatitis and allergic rhinitis. Our aim was to further investigate this region using a panel of 306 SNPs genotyped in 365 EGEA families, spanning a 20 Mb region. To test for association of these SNPs and the three diseases, we used three different methods: two family-based tests (FBAT), the first one (FBAT1) applied to affected sibs only, the second one (FBAT2) to affected and unaffected sibs and logistic regression (LR) applied to all siblings while accounting for family dependence.
The strongest association signal was found between asthma and two adjacent SNPs, rs2282661 and rs95119 ($r^2 = 0.14$) using FBAT1 ($P = 0.0002$ and 0.008 respectively). One of these SNPs, rs2282661, showed also association using FBAT2 ($P = 0.002$) and a marginal signal with LR ($P = 0.09$). Using the parents of the nuclear families as an independent sample, we replicated the association of asthma with rs95119 ($P = 0.009$). The two SNPs showing association belong to NELL1, a gene which encodes a protein that contains epidermal growth factor (EGF)-like repeats and is a potential candidate for allergic diseases. This gene has been recently found associated with Crohn's disease which shares genetic determinants with asthma. Further analyses using multi-marker methods are underway to confirm our findings.

246

The Role of Polymorphisms in the ACE and ADRB2 Genes in Incidence of Myocardial Infarction in Framingham Heart Study Participants

Alexander Kulminski (1), Svetlana Ukraintseva (1), Irina Culminskaya (1), Konstantin Arbeev (1), Igor Akushevich (1), Anatoli Yashin (1)
(1) Duke University, Center for Population Health and Aging

In this study we focus on the interplay between the effects of the rs1042714 SNP in the beta2-adrenergic receptor (ADRB2) gene and six rarely-studied SNPs in the angiotensin-converting enzyme (ACE) gene, which are not in strong linkage disequilibrium with the ACE I/D polymorphism, on incidence of myocardial infarction (MI) in a longitudinally followed sample (N = 1755) of participants of the Framingham Heart Study Offspring cohort. The ADRB2 SNP (rs1042714) and two of the ACE SNPs (rs4363 and rs12449782; $D' = 0.77$, $r^2 = 0.42$) were significantly associated with incidence of MI over the entire follow up (starting from the first examination in 1971 and through 2007; $P = 0.013$, 0.045, and 0.003, respectively) and over the shorter-term period (starting from the fourth examination in 1987; $P = 0.011$, 0.026, and 0.002, respectively) in the best-predicting Cox regression models with all these SNPs included and with adjustment for age, sex, blood pressure, diabetes, cholesterol, BMI, and smoking. The effect of heterozygous genotype on the incidence of MI was opposite for rs4363 and rs12449782 SNPs of the ACE gene, i.e. protective for the rs12449782 and detrimental for the rs4363. Analysis of the effects of compound genotypes of these two ACE SNPs (rs4363 and rs12449782) and all three SNPs (including rs1042714) better delineated joint contribution of different genetic markers into the risk of MI. The role of integrative systemic genetic studies of aging-related diseases is discussed.

247

Entropy Based Marker Selection and Mantel Statistics in Candidate Gene Analysis of Oxidative Stress Related Mechanisms and Breast Cancer Risk in the MARIE Study

Anke Schulz (1), Petra Seibold (1), Rebecca Hein (1), Dieter Flesch-Janys (2), Christine Fischer (3), Lars Beckmann (1), Jenny Chang-Claude (1)
(1) Department of Cancer Epidemiology, German Cancer Research Centre, Heidelberg, Germany
(2) Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
(3) Institute of Human Genetics, University of Heidelberg, Heidelberg, Germany

Haplotype analysis is an established option for testing association in candidate genes, yet there is no gold standard for marker selection prior to haplotype estimation. We combined an entropy based marker selection algorithm (EMS) with Mantel statistics using haplotype sharing, a test for marker-trait association, which performed well for simulated data. Here, we apply the new approach to a real data set of 131 SNPs, mainly tagging SNPs ($r^2 > 0.8$), in 20 candidate genes involved in oxidative stress related mechanisms, genotyped in 1628 postmenopausal breast cancer cases and 1947 matched controls from the German case-control study MARIE. For each tested marker, the other markers in the same gene are screened by an entropy-based criterion of multilocus linkage disequilibrium (LD) and selected if they increase the LD. Haplotype sharing analysis is applied to the haplotypes estimated from these selected markers. This approach is

compared to pointwise conditional logistic regression (LR). LR yielded 10 nominally significant variants ($P < 0.05$), mostly in genes coding for members of the thioredoxin system and in the *MT2A* gene. The EMS Mantel statistics yielded significant results for 4 variants, 2 in *TXN2* and 2 in *MT2A*, corresponding to the 4 most significant variants for LR. The most significant one, rs2281082 in *TXN2* for LR, was inversely associated with breast cancer risk ($p$trend = 0.002). Using EMS Mantel statistics, the most significant was rs1580833 in *MT2A* ($P = 0.0007$).

248

Polymorphisms in Adipokine Genes and Prostate Cancer Risk and Aggressiveness

Nora L. Nock (1), Jennifer Beebe-Dimmer (2), Andrew Rundle (3), Christine Neslund-Dudas (4), Cathyrn Bock (2), Deliang Tang (3), Benjamin A. Rybicki (4)
(1) Case Western Reserve University
(2) Wayne State University
(3) Columbia University
(4) Henry Ford Health System

Studies examining the association between obesity and prostate cancer risk have produced inconsistent results; however, recent reports suggest that obesity may only increase the risk of aggressive forms of prostate cancer. The molecular mechanisms driving this putative association are not clear but may involve alterations in molecules released from adipose tissue (''adipokines'') such as leptin and tumor necrosis factor-alpha (TNF-a), which have both been shown experimentally to enhance cell proliferation and angiogenesis and, therefore, may play a role in accelerating prostate tumor growth. Common variants in adipokine genes have not been well explored in prostate cancer. Therefore, we evaluated the potential association between polymorphisms in several adipokine genes including leptin (LEP), leptin receptor (LEPR), adiponectin (ADIPOQ), adiponectin receptors (ADIPOR1, ADIPOR2), interleukin-6 (IL-6) and TNF-a and prostate cancer risk in a case-control study of Caucasian and African-American men treated at Henry Ford Health Systems. We found that SNPs in LEPR (e.g. rs1887285), ADIPOQ (e.g. rs822391) and ADIPOR1 (e.g. rs1342387) were associated with prostate cancer risk. Additional analyses stratified by ethnicity, disease aggressiveness and body composition are underway and will be presented.

249

Tiled Logistic Regression for Response to Antidepressant Treatment

Heejong Sung (1), Yoonhee Kim (1), Juanliang Cai (1), Alexa J.M. Sorant (1), Francis J. McMahon (2), Alexander F. Wilson (1)
(1) Genometric section, Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD
(2) Genetic Basis of Mood and Anxiety Disorders, NIMH, NIH, Bethesda, MD

A new method called ''Tiled Regression'' has been developed for quantitative traits [Wilson et'al., pers. comm.]. The genome is divided into independent regions, or tiles, and the SNPs or sequence variants within each tile are analyzed with multiple and stepwise regression to identify markers within the tile having independent and significant additive effects on the phenotype. Markers identified within tiles are then reanalyzed together with stepwise regression. In this study, a version of tiled regression using logistic regression was applied to the binary trait ''responder status'' in data from the STAR∗D (Sequenced Treatment Alternatives to Relieve Depression) study [Rush et'al., 2004]. This trait was derived from the relative improvement in a standard depression score over a period of antidepressant treatment. 1096 white subjects, classified according to high or low levels of improvement, were analyzed with 705 SNPs in candidate genes. Tiles were defined in two different ways: by hotspot (within and between high-recombination ''hotspots'', 197 tiles) and by gene (within and between known genes, 102 tiles). At a significance level of 0.05, 7 SNPs were selected for the final model using hotspot-based tiles and 8 for gene-based tiles, with overlap of 5 SNPs. The SNP rs7997012 on chromosome 13 with the most significant p-value in the final model (0.00116 and 0.00163 using hotspot- and gene-based tiles, respectively) was the one previously noted by McMahon et'al. [2006].

250

Measuring the Contribution of Genetic Variants in Association Analysis with Survival Outcome

Martina Mueller (1), Helmut Kuechenhoff (2), Claudia Lamina (3), Doerthe Malzahn (4), Heike Bickeboeller (4), Arne Pfeufer (5), Thomas Illig (6), H.-Erich Wichmann (7), Iris M Heid (8)
(1) Department of Medicine I, Klinikum Grosshadern, Munich, Germany
(2) Ludwig-Maximilians-Universität, Munich, Institute of Statistics, Germany
(3) Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria
(4) Gustav-August-Universität, Göttingen, Department of Genetic Epidemiology, Göttingen, Germany
(5) Institute of Human Genetics, Helmholtz Center Munich, Germany
(6) Institute of Epidemiology, Helmholtz Zentrum München
(7) Ludwig-Maximilians-Universität, Munich, Inst. of Medical Informatics, Biometry and, Epidemiology, Chair of Epidemiology, Germany
(8) Institute of Epidemiology and Preventive Medicine, University of Regensburg, Germany

Introduction: In genetic association analyses, it is of increasing interest to report measures that quantify the impact of the genetic variant on the phenotype by a percentage. Typically, measures of explained variation like R^2 fulfil this purpose. For survival outcomes, however, the ongoing development of statistics leads to a variety of proposed criteria. Categorical covariates like genetic variants further restrict applicability of suggested criteria.
Methods: In order to compare the performance of different criteria to judge the impact of genetic variants on survival outcomes, we defined the following

requirements: (a) limitation to the range [0;1], (b) robustness against censoring, (c) values increasing with the associated increasing hazard ratio. With respect to these, criteria based on three different approaches have been compared in simulation studies: deviance residuals (criterion 1), variation of individual survival curves (criterion 2), Schoenfeld residuals (criterion 3).
Results: All criteria strongly depended on the underlying effect size. Criterion 1 was highly dependent on censoring percentage and showed a tendency to systematically exceed the desired range of values in extreme scenarios. Criterion 2 tended to yield generally low values. Our requirements were best fulfilled by criterion 3.
Conclusion: The criterion based on Schoenfeld residuals is identified as a powerful tool for judging the contribution of genetic variants and a wide range of possible applications.

251

A Bayes Model Averaging Approach Using Haplotypes

Nadine Cremer (1), Jenny Chang-Claude (1), Lars Beckmann (1), David V. Conti (2)
(1) German Cancer Research Center
(2) University of Southern California

Dealing with multivariate SNP data one must decide whether to base the analysis on (subsets of) SNPs or whether to use estimated haplotypes to capture dependencies between polymorphisms due to linkage disequilibrium.
We present a Bayes Model Averaging approach averaging over all haplotype models for each subset of SNPs to address this question. The posterior distribution of models is determined from the regression likelihoods and a prior influencing model parsimony. Final inference for the most parsimonious SNP set is determined via posterior probabilities for each model and for any given SNP.
For a limited number of SNPs we are able to enumerate all models. For larger numbers of variables Markov Chain Monte Carlo techniques can be used to efficiently sample the model space.
Using coalescent simulations, we show that if a causal SNP is genotyped, the approach weights more heavily on the corresponding single SNP model. In contrast, if the true causal SNP is not genotyped, but captured by a haplotype of genotyped SNPs, a parsimonious set of SNPs is selected to describe the haplotype association. The advantage of such a method is the ability to identify SNPs that drive the association, while also capturing haplotype dependencies. Via extensive simulations, type I error and power of the proposed method are investigated, and the additional information gained by using this approach is discussed across several real data examples.

252

High-density SNP Genotyping of DNA Extracted from Buccal Cells

Stephen W. Erickson (1), Mario A. Cleves (1), Stewart L. MacLeod (1), Charlotte A. Hobbs (1)
(1) University of Arkansas for Medical Sciences

Genetic studies conducted in geographically diverse regions may require participants to collect DNA samples and mail them to a remote study center. DNA extracted from self-collected buccal cell samples is typically of lower quantity and quality than blood samples and is subject to contamination by oral bacteria, with subsequent degradation and fragmentation of human DNA.
The National Birth Defects Prevention Study is an ongoing population-based study in which more than 47,000 buccal cell samples have been collected from case-parental and control-parental triads. In a candidate-gene study of congenital heart defects, a subset of these samples was genotyped for 1,536 SNPs in 61 candidate genes using a customized Illumina GoldenGate SNP panel.
We show how novel approaches are applied to fluorescent intensity data to produce genotype calls. First, a carefully fitted probabilistic model of intensity data yields objective, quantitative criteria that can be applied across all SNPs and samples within a study. For a significant number of samples at each SNP, the model can only eliminate one of the three possible genotypes and not definitely assign cluster membership. For many of these cases, pedigree and linkage disequilibrium structure can be used to augment the fluorescent intensity data and resolve ambiguous SNP calls. In this manner, we aim to make maximal use of a valuable research resource, with the goal of elucidating the genetic basis for nonsyndromic birth defects.

253

An Association Study of Candidate Regions that Emerged from Linkage Analyses with Schizophrenia: an Efficient Approach to Control for Multiple Testing and Investigate Interactions

Chantal Mérette (1), Valérie Jomphe (1), Aurélie Labbe (2), Yvon Chagnon (1), Marc-André Roy (1), Michel Maziade (1)
(1) Laval University
(2) McGill University

Based on the linkage signals from the genome scan reported in Maziade et'al. (Mol Psychiatry, 2005), we defined candidate regions of ± 5 Mb around the markers that yielded maximum lodscores over 2.6. We studied the association between SNPs located in these linked regions and schizophrenia (SZ) in a sample of 247 cases and 137 controls independent from the family sample but from the same Eastern Quebec population. We selected 19964 SNPs and performed $\times^2$ testing. We reported odds ratios (OR) and p-values that we interpreted against an FDR correction and a threshold for tendency ($FDR_p < P < .001$). We obtained evidence of association that remained significant after controlling for multiple testing ($P = 7 \times 10^{-6}$; $FDR_p = .0193$; $OR = 2.0$) on chr 3 involving SNP rs2613946 located near the BOC gene. Other suggestive results included an OR of 17.4 ($P = .0002$) on chr 6 within the CDKAL1 gene involved in type 2 diabetes, and one on chr 18 with rs9950834 within the ST8SIA5 gene ($OR = 4.65$; $P = .0007$). Two genes yielded suggestive association with SZ in the area of linkage on chr 13 that we had previously clearly replicated in a second sample (Maziade et'al., Eur J Hum Genet, 2009): $OR = 3.45$, $P = .0004$, with rs9548509 within FREM2 and, within ENOX1, the SZ subjects never

showed the GG genotype that was seen in 5.15% of the control group. We will investigate the potential interactions among the associated SNPs using statistical learning tools such as classification and regression trees (CART).

254

The PhenX Toolkit—Get the Most from Your Measures

Carol M. Hamilton (1), Peter Kraft (2), Lisa Strader (1), Joseph Pratt (1), Jane Hammond (1), Tabitha Hendershot (1), Wayne Huggins (1), Dean Jackman (1), Richard Kwok (1), Deborah Maiese (1), Destiney Nettles (1), Helen Pan (1), Diane Wagener (1), Mike Zmuda (1), Heather Junkins (3), Rongling Li (3), Erin Ramos (3), William Harlan (4), Jonathan Haines (5)
(1) RTI International
(2) Harvard University
(3) NHGRI
(4) Retired, NIH
(5) Vanderbilt University

Despite the vast potential for cross-study comparisons, the lack of standardized or comparable phenotypic and environmental measurements has limited the ability to combine data from GWAS and other genomic studies. To enhance cross-study analyses, RTI International and the National Human Genome Research Institute (NHGRI) are collaborating on a three- year project called PhenX (consensus measures for Phenotypes and eXposures). The goal of PhenX is to identify 15 high-priority, well-established, measures in each of 20 research domains. Working Groups (WGs) of domain experts select these measures via a consensus-building process that also considers input from the broader scientific community. The measures selected by the WGs are made publicly available via the PhenX Toolkit. Researchers will want to visit the PhenX Toolkit to select measures as they plan a new study or to add measures to an existing study. By the end of 2009, the PhenX Toolkit is expected to include measures that were selected by 11 WGs, including Demographics, Anthropometrics, Alcohol, Tobacco and Other Substances (ATOS), Cardiovascular, Nutrition and Dietary Supplements, Environmental Exposures, Oral Health and Cancer. Broad acceptance and use of PhenX measures should greatly facilitate cross-study analysis. You can visit the PhenX Toolkit at https://www.phenx-toolkit.org/. Supported by: NHGRI, Award No. 1U01 HG004597-01.

255

Clinical Implications: ApoE Genotyping as a Progression-Rate Biomarker in Phase 2 Disease Modification Trials for Alzheimer's

Cliona Molony (1), David Stone (2), Christine Suver (3), William Potter (2), Eric Schadt (3)
(1) Rosetta Inpharmaics
(2) Merck Research Laboratories
(3) Rosetta Inpharmatics

Current studies suggest that apolipoprotein E (apoE) genotype influences the rate of progression in Alzheimer's disease (AD), with patients carrying the e4 allele progressing faster than non-carriers. Results from some clinical trials demonstrate trends for apoE status and disease progression that disagree with trends in epidemiological studies, however, raising questions as to whether populations defined for phase II clinical trials conform to the results seen in epidemiological studies, and raising concerns over the use of apoE genotyping as a progression rate pharmacogenetic marker in AD disease-modification trials. We examined the cognitive subset of ADAS-Cog from 436 placebo patients who had been genotyped for the e4 allele in two phase II clinical trials. In one trial e4 carriers showed a faster rate of decline ($P < 0.001$) that was evident by 9 months after baseline; in the second trial e4 carriers showed worse cognitive scores at enrollment ($P < 0.05$)and a trend towards faster rate of cognitive decline at 12 months ($P = 0.041$). These results suggest that populations defined for clinical trials do in fact conform to the results seen in epidemiological studies, and studies showing a slower rate of disease progression in e4 carriers probably represent chance finding due to low power. Furthermore we propose that apoE genotyping may be only useful as a disease-progression biomarker in larger studies with >200 patients e4 carriers per treatment group.

256

Kinship Testing goes Linkage

Michael Nothnagel (1), Michael Krawczak (1)
(1) Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany

Practical applications of kinship testing in humans so far have largely ignored linkage. This is mainly due to the fact that, until quite recently, genetic markers routinely used for identification purposes were not (closely) linked. Furthermore, over 95% of kinship cases in humans involve a disputed paternity in a trio of mother, child, and alleged father. Here, inter-marker linkage is indeed computationally irrelevant in the absence of linkage disequilibrium. However, with the introduction of a substantial number of novel markers into forensic practise, linkage has become an issue in kinship testing. Furthermore, the availability of more comprehensive markers sets has rendered increasingly more complex kinship cases tractable with fewer individuals tested, eventually involving only pairs of supposed relatives. We performed extensive simulations of genotypes for the 34 autosomal STRs widely used in forensic practise today to assess the current limits of the resolution of pair-wise kinship testing and the effects of an appropriate consideration of linkage between STRs on the ensuing likelihood ratios. The results reveal a clear-cut difference between the power of the 34 STRs to resolve first degree and second degree relatedness. They also demonstrate the need to take inter-marker linkage appropriately into account.

257

FCN2 Genetic Variation and L-ficolin Serum Levels in African Malaria Patients

Imad Faik (1), Bertrand Lell (1), Sanjeev Krishna (1), Segun I. Oyedeji (1), Idris Zulkarnain (1), Peter G. Kremsner (1), Jürgen F.J. Kun (1)
(1) Institute for Tropical Medicine, University Tübingen, Germany

The completion of the human genome project promoted a new area of genetic studies. Single nucleotide polymorphisms (SNPs) are one of the most prevalent sources conferring inter/intra population genetic alterations that have been associated with different phenotypes of diseases in humans including malaria. The human Ficolin-2 (L-ficolin) is a serum protein that binds to sugar arrays of different human micro-pathogens forcing phagocytosis. 112 severe and 170 mild malaria cases from Gabon were recruited. Here, we investigate for the first time the clinical significance of SNPs in the FCN2 gene in African malaria patients by correlating polymorphisms with the malaria clinical manifestation severe/mild, and with the circulating L-ficolin serum concentrations. 3 promoter SNPs (−986G>A, −602G>A and −4A>G) and 1 SNP in exon 8 (+6424 G>T) were screened using the TaqMan real-time PCR. Furthermore, the promoter region of the same gene was sequenced in 40 healthy Gabonese. Additionally, we measured the L-ficolin serum levels at different time points (t_0 admission time, and t_6 six months later) in severe ($t_0 = 70$, and $t_6 = 68$) and mild ($t_0 = 88$, and $t_6 = 54$) malaria cases using Elisa. Linkage disequilibrium data reveal polymorphic allelic combination patterns in the FCN2 promoter region of the analyzed healthy Gabonese, where strong allelic combinations at (−986 and −4) such as at (−557 and −64) were found. Further analysis showed the following results: The L-ficolin concentration [μg/ml] was at t_0 significantly higher when compared with t_6 in the mild (mean 11.8 vs. 8.2) and in the severe cases (mean 14.6 vs. 11.1), (*t*-test, $P<0.001$, respetively, $P<0.001$). When the mean of the L-ficolin concentration was compared between patients, the severe group showed at both time points t_0 and t_6 a significantly higher mean values (14.60 vs. 11.83) and (11.12 vs. 8.24), respectively, (*t*-test, $P<0.001$, respectively, $P<0.001$). A significant association of L-ficolin levels with the constructed haplotypes at (986G>A, −602G>A, −4A>G and +6424 G>T) was observed at t_0, when all patients were included in one group: the GGAT haplotype correlated with the lowest L-ficolin concentration (oneway analysis, $P = 0.007$). Genotypes analysis showed an association of L-ficolin levels with the FCN2 polymorphism at position + 6424 G>T (oneway analysis, $P = 0.002$). The variation was in a gene-dose dependent manner, decreasing from the homozygous over the heterozygous to the homozygous mutant. These data showed that L-ficolin levels are variable among African malaria patients according to their FCN2 haplotype, and show that FCN2 is an acute phase protein such as mannose-binding lectin (MBL) and C-reactive protein (CRP) and other serum components.

258

Efficient Genomewide Association Analysis in Small Families

Michael B. Miller (1)
(1) Minnesota Center for Twin and Family Research, Department of Psychology, University of Minnesota

Most genome-wide association studies (GWAS) include unrelated individuals whose phenotype data are theoretically independently distributed so that ordinary least-squares (OLS) methods of regression analysis can be used to assess genotype-phenotype associations. When observations are correlated because families were sampled, the OLS method loses power. The generalized least-squares (GLS) method of regression analysis can be used to take familial correlations into account and provide statistically-valid results, but computational demands of the method can become extraordinary in GWAS because of the large number of tests. We present a modified GLS method that gives essentially the same result as GLS but with a great increase of computational speed. The method is demonstrated on simulated data for 8000 subjects in 2000 families based on HapMap Phase-1 haplotypes. Using code written in Octave, an ordinary desktop PC can complete analysis of a trait with five covariates for 811,889 SNPs in less than 17 minutes. Such rapid analysis allows use of random permutation tests on larger multi-core computer systems.

259

Multilocus Bayesian Meta-analysis of Gene-disease Associations

P.J. Newcombe (1), C. Verzilli (1), J. Pablo-Casas (1), Aroon Hingorani (2), L. Smeeth (2), J. Whittaker (1)
(1) Non-Communicable Disease Epidemiology Unit, NCDEU, London School of hygiene & Tropical Medicine, London, United Kingdom
(2) Epidemiology & Public Health, UCL Division of Population Health, London, United Kingdom

Meta-analysis is a vital tool in genetic epidemiology, however, there are a number of factors which limit its effectiveness. We consider two such limitations and present newly developed possible solutions. Meta-analyses to identify gene-disease associations are compromised when contributing studies have typed partially overlapping sets of markers. Currently, only marginal analyses are possible, and these are restricted to the subset of studies typing that marker. This does not allow full use of available data, and leads to confounding of marker effects by closely associated markers. A second limitation of genetic meta-analysis arises when looking for gene-environment interactions. There will often be additional studies which have looked at the gene-disease association alone, and as such report data marginalised over the environmental variable of interest. These studies can still contribute information on the gene-environment interaction, but due to the missing data there is currently no method to incorporate them into the same analysis.
We present a Bayesian approach to the first problem which exploits prior information on underlying haplotypes to allow multi-marker analysis incorporating data from all relevant studies of a gene or region, irrespective of the markers typed. We present results from application of this approach to data on a possible association between *PDE4D* and ischemic stroke, and a known association between *LPL* and coronary heart disease, in which we take advantage of the models ability to perform a multivariable analysis to make inference on the number of causal sites. We also present an extension to this methodology aimed at tackling the second problem, which enables simultaneous analysis of gene-environment interaction data and gene-disease data, by exploiting prior information on the population prevalence of a binary environmental factor. We use simulations to demonstrate a scenario in which this method can contribute an important increase in power.