Genetic Epidemiology 34:917–992 (2010)

# ABSTRACTS FROM THE

# NINETEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

**1**

**Genetic Signatures of Exceptional Longevity**

Paola Sebastiani (1), Nadia Solovieff (1), Annibale Puca (2), Stephen W. Hartley (1), Efthymia Melista (1), Stacy Andersen (1), Daniel A. Dworkis (1), Jemma Wilks (1), Richard H. Myers (1), Martin H. Steinberg (1), Monty Montano (1), Clinton T. Baldwin (1), Thomas T. Perls (1)
(1) Boston University
(2) CNR, Italy

Like most complex phenotypes, exceptional longevity (EL) is thought to reflect the combined influence of environmental and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of EL in 1,055 centenarians and 1,267 genetically matched controls. We conducted 3 different analyses. First, we identified and replicated 33 single nucleotide polymorphisms (SNPs) that met genome-wide significance using both frequentist and Bayesian analyses. We then built a genetic risk model to evaluate, in silico, the effect of combinations of SNP alleles to predict EL and to explore the hypothesis that varying combinations of SNPs characterize different pathways to EL. Our model is based on an ensemble of 150 nested Bayesian networks and predicts EL with 77% accuracy in an independent set of centenarians and controls. The advantage of an ensemble of nested models is that it can be used to generate a genetic risk profile for each study subject. These profiles can then be clustered to discover prototypical signatures of EL. This in-silico analysis revealed that 90% of centenarians can be grouped into 19 clusters characterized by different genetic signatures of varying predictive value. The different signatures correlated with differences in the prevalence and age of onset of age-associated diseases (e.g., dementia, hypertension, cardiovascular disease) and may help dissect this complex phenotype into sub-phenotypes of healthy aging.

**2**

**Fishing for Disease Genes in the Random Forest of GWAS SNPs**

Wei W. Yang (1), Chi C. Gu (1)
(1) Washington University School of Medicine

In Genome wide association studies (GWAS), the vast number of all possible interaction models makes exhaustive search for important interactions infeasible. The approach of Random Forest (RF) provides a potential solution. But the method is unsuitable for GWAS because of the huge number of predictors and high noise level. We propose a novel method using a genetic-algorithm based on RF, termed random forest fishing (RFF). It repeatedly updates a set of core predictors until a globally important set of variables is found predictive of the disease outcome. It performs best when there are appreciable levels of interactions among the important variables. Furthermore, using guidance provided by results from tests of pairwise interactions greatly adds to its efficiency. Evaluations of RFF were done by simulation studies with 50 K GWAS SNPs, using 5 scenarios of disease models with SNP-SNP interactions. Even when none of the risk SNPs has any marginal effect, the guided RFF could identify every single risk SNP with ~50% power in a sample of 2000 cases and 2000 controls (Scenario 1). The power is dramatically improved to approaching 100% when the interacting SNPs manifested some marginal effects (Scenario 3-1). These results showed that the novel RFF method is a useful tool for identifying important genetic predictors when their interactions are substantial; it becomes especially powerful when used in combination with genome-wide tests of pairwise interactions as guidance.

**3**

**Exploiting Homozygosity Tracts to Search for Rare Recessive Variants Involved in Complex Traits**

Steven Gazal (1), Marie-Claude Babron (1), Jean-Charles Lambert (2), Dominique Campion (3), Claudine Berr (4), Christophe Tzourio (5), Didier Hannequin (3), Florence Pasquier (6), Olivier Hanon (7), Jacques Epelbaum (7), Jean-Francois Dartigues (8), Mark Lathrop (9), Philippe Amouyel (3), Emmanuelle Genin (1), Anne-Louise Leutenegger (1)
(1) INSERM U946, Paris, France
(2) Inserm U744, Lille, France
(3) Inserm U614, Faculte de Medecine-Pharmacie de Rouen, Rouen, France
(4) Inserm U888, Hopital La Colombiere, Montpellier, France
(5) Inserm U708, Paris, France
(6) Universite de Lille Nord de France, Lille, France
(7) Inserm U894, Faculte de Medecine, Universite Paris Descartes, Paris, France
(8) Inserm U897, Victor Segalen University, Bordeaux, France
(9) National de Genotypage, Institut Genomique, Commissariat a l'energie Atomique, Evry, France

Genome-wide association studies that use univariate tests are efficient to identify the variants involved in diseases. However, these tests cannot detect recessive effects, especially when the genetic variants involved are rare. Instead of looking at one marker at a time, recent works propose to detect tracts of homozygosity associated to the disease status. A novel method, implemented in the program WHAMM (Whole-Genome Homozygosity Analysis and Mapping Machina, http://www.broadinstitute.org/~bvoight/whamm/), estimates the Identity by Descent (IBD) status of each individual and compares the distribution of autozygosity between cases and controls. To estimate the IBD status by

Published online in Wiley Online Library (www.wileyonlinelibrary.com).
DOI: 10.1002/20553

taking account of the linkage disequilibrium, this approach creates segments delimited by the hotspots compiled from the HapMap data, and calculates the frequency of the haplotypes in all the population.
In this work, we studied the properties of this method and compared it to the FEstim method (Leutenegger et al., 2006) that was developed in the context of homozygosity mapping of rare recessive traits. An application on a French dataset of 1,886 affected individuals with Alzheimer's disease and 5,044 controls genotyped with Illumina Human610-Quad BeadChips (Lambert et al, 2009) will be presented.

**4**
**A Characterization of the Training Needs in Genetic Epidemiology**
Mera Krishnan (1), NIH Working Group on Statistical Genetics Training Needs (1), Alexander F. Wilson (1)
(1) National Human Genome Research Institute

Advances in DNA sequencing technology have outpaced the training of new statistical geneticists and genetic epidemiologists who will be able to analyze and develop new methods of analyzing the massive amounts of GWAS and sequence data being generated. A 61 question survey was administered to members of IGES and selected members of the American Society for Human Genetics between 2008 and 2009 to assess the state of training in genetic epidemiology and statistical genetics. Pre- and post-doctoral fellows, and faculty-level trainers were surveyed to estimate the number of future scientists and determine their academic backgrounds. Faculty were queried about challenges in recruiting qualified trainees for available training spots. The most frequently cited reasons for an inability to fill spots were a lack of qualified domestic applicants to fill positions (73%) and difficulty in recruiting foreign applicants due to residency regulations (14%). Of the 100 post-doctoral fellows surveyed, 82% expected to take an entry-level faculty position upon completion of training. Of the Pre-doctoral trainees surveyed, 74% had taken no biology-related courses as a part of their graduate training. Of these trainees, 38.9% did not have a biology-related undergraduate major. The results of this survey highlight opportunities in improving the quantity and quality of trained scientists in this field.

**5**
**On Optimal Pooling Designs to Identify Rare Variants Through Massive Resequencing**
Joon Sang Lee (1), Murim Choi (1), Xiting Yan (1), Richard P. Lifton (1), Hongyu Zhao (1)
(1) Yale University

The advent of next-generation sequencing (NGS) technologies has facilitated the detection of rare variants. Despite the significant cost reduction, sequencing cost is still high for large-scale studies. In this work, we examine DNA pooling as a cost-effective strategy for rare variant detection. We consider the optimal number of individuals in a DNA pool for a given set of minor allele frequency, coverage depth, and detection threshold. The optimal number of individuals in a pool seems to be indifferent to minor allele frequencies at the same coverage depth. In addition, when the contributions are equal, the total number of individuals required is similar for a given minor allele frequency at different coverage depths, where fewer lanes are required to identify rare variants when the coverage depth increases. When the contributions are unequal, more individuals are needed for larger pools.

**6**
**Comprehensive Approach to Analyzing Rare Genetic Variants**
Thomas J. Hoffmann (1), Nicholas J. Marini (2), John S. Witte (1)
(1) University of California San Francisco
(2) University of California Berkeley

Recent findings suggest that rare variants play an important role in both monogenic and common diseases. Due to their rarity, however, it remains unclear how to appropriately analyze the association between such variants and disease. A common approach entails combining rare variants together based on a priori information and analyzing them as a single group. Here one must make some strong assumptions about what to combine. Instead, we propose two approaches to empirically determine the most efficient grouping of rare variants. The first considers multiple different possible groupings using existing information. The second is an agnostic "step-up" approach that determines an optimal grouping of rare variants analytically and does not rely on prior information. To evaluate these approaches, we undertook a simulation study using sequence data from genes in the one-carbon folate metabolic pathway. Our results show that using prior information to group rare variants is advantageous only when information is quite accurate, but the agnostic step-up approach works well across a broad range of plausible scenarios. This agnostic approach allows one to efficiently analyze the association between rare variants and disease while avoiding strong assumptions required by other approaches for grouping such variants.

**7**
**Genome Wide Meta-Analysis of Joint Tests for Genetic and Gene-Environment Interaction Effects**
Hugues Aschard (1), Dana B. Hancock (2), Stephanie J. London (2), Peter Kraft (1)
(1) Harvard School of Public Health, Department of Epidemiology
(2) National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services

There is a growing interest for the study of gene-environment ($G \times E$) interactions in the context of genome-wide association studies (GWAS), which may uncover new causal loci. However, statistical tests incorporating $G \times E$ interaction require large sample sizes to have reasonable power, which will likely necessitate meta-analytic approaches.
Although meta-analysis of a single $G \times E$ interaction parameter is straightforward, meta-analysis of multiple parameters is less well known. We describe an approach for meta-analysis of a joint test for a genetic and $G \times E$ interaction, based on estimates for multiple parameters and their variance-covariance matrix, and we show that the meta-analytic joint test has appropriate Type I error rates.
Using simulation studies across a broad range of genetic main effects and $G \times E$ interaction effects, we show that in many cases the joint test can be more powerful than the marginal test of genetic association and the standard 1 d.f. interaction test, the latter having less than 1% power in almost all situations. The largest gains of power (?50%)

compared to the marginal test are seen when the genetic effect is weak in one exposure stratum but strong in another, or when the main and interaction effects are both strong and in opposite directions. Lastly, regardless of the test used, we have also shown that sample sizes far exceeding those of a typical GWAS ($N \gg 2{,}000$ *cases*) will be needed to reliably detect genes with subtle G × E interaction patterns.

**8**

**In Silico Genotype Imputation on Large Pedigrees**

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

Availability of dense genotypes or sequence data is commonplace but expensive. Many pedigree studies contain existing samples with sparse genotypes. In small pedigrees, it is already feasible to impute dense genotypes from the existing data and a few densely-genotyped individuals. Here we show that "in silico" dense genotyping on large pedigrees is also now computationally practical. Using MORGAN, MCMC-based sampling conditional on genotype data provides realizations of inheritance vectors (IVs) on large pedigrees. We introduce GIGI (Genotype Imputation Given Inheritance) which imputes dense genotypes using sampled IVs and the few observed dense genotypes. GIGI uses probabilistic inference based on IVs, allele frequencies, and pedigree structure, and uses threshold-based allele calling. Coupled with the IBDgraph program, which identifies equivalent sampled IVs, this provides rapid and efficient imputation. We used GIGI on a 95-member multigenerational real pedigree with ~60 individuals observed for STRs and 323 dense SNPs over a ~50 cM region. SNPs from 13 individuals were used for imputation with the remaining data used for validation. Using a stringent genotype-calling threshold, 68% of alleles were called in the validation sample, of which 97.6% were correct. These results demonstrate that accurate imputation in large pedigrees is practical, providing a cost-effective sequential approach to obtain dense marker genotypes from platforms including next-generation sequencing.

**9**

**A Flexible Likelihood Framework for Dissecting Gene Pleiotropy Combining Non-randomly Ascertained Samples: Application to Sequence Data**

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

Studying gene pleiotropy using exome data will provide insight into underlying genetic pathways and lead to powerful mapping of complex traits. Since many important secondary traits are often measured in addition to the main phenotype under study, it is greatly beneficial to combine samples for phenome mapping. Sample ascertainment can be complex for existing sequencing studies and involve family histories, multiple phenotypes or sub-phenotypes. Due to phenotypic correlations, the analyses of gene pleiotropy can be seriously biased if sample ascertainment is not properly adjusted. Existing methods are limited for joint modeling phenotypic correlations and complicated ascertainment schemes. We propose a modified liability-threshold likelihood framework (MoLTeF) for mapping genes with pleiotropic effects using selected samples. MoLTeF is very flexible and allows efficient inferences and estimations of genetic parameters of interest. Extensive simulations under a rigorous population genetic model are carried-out with phenotypic parameters estimated from complex traits. It is shown that the power for mapping secondary traits can be greatly improved when multiple cohorts are combined. In the presence of pleiotropy, if a selected sample is used where the gene locus is associated with both main and secondary phenotypes, the power will be further increased. In conclusion, MoLTeF will play an important role in dissecting human phenomes in sequencing based studies.

**10**

**A Bayesian Partitioning Model for Detection Of Multilocus Interaction in Case-Control Studies**

Xiang Li (1), Saonli Basu (1)
(1) University of Minnesota, Division of Biostatistics

Studying one single nucleotide polymorphism (SNP) at a time may not be sufficient to understand complex diseases. A SNP alone may have little or no effect on risk of disease, but together may increase the risk substantially. The joint behavior of genetic variants is often referred to as epistasis or multilocus interaction. We have proposed here a Bayesian partitioning model to detect such multilocus interaction. Our model clusters genotypes according to the direction of association, and computes, via Markov chain Monte Carlo, the posterior probability that each marker set is associated with the disease. Since the number of parameters to model multilocus interaction grows exponentially with the number of SNPs, we propose a pair-wise scoring approach to approximate high order interactions. We illustrate and compare our model with existing approaches, such as Multifactor Dimensional Reduction and Penalized logistic regression, through extensive simulations. In the seven interaction models we considered, our method had uniformly higher power to detect a rare SNP; and comparable power for a common SNP. We also applied our method to detect significant SNPs associated with acute rejection using the genotype data from a custom chip with 3,590 SNPs in a cohort of 271 transplant patients. We jointly analyzed SNPs within different pathways and identified several SNPs including the ones that cannot be detected through single SNP analysis.

**11**

**Realities and Limitations of Coverage in Current "Whole"-Exome Sequencing Capture Approaches**

Kevin B. Jacobs (1), Meredith Yeager (1), Michael G. Cullen (1), Xijun Zhang (1), Joseph Boland (1), Jennifer Bacior (1), Victor Lonsberry (1), Casey Matthews (1), David Roberson (1), Quan Chen (1), Laurie Burdett (1), Idan Menashe (2), Xiaohong R. Yang (2), Lynn R. Goldin (2), Mary L. McMaster (2), Neil E. Caporaso (2), Philip R. Taylor (2), Maria Teresa Landi (2), Joshua Sampson (2), Nilanjan Chatterjee (2), Michael L. Nickerson (3), Kate McGee (3), Michael C. Dean (3), Javed Khan (4), Margaret A. Tucker (2), Stephen J. Chanock (2), Alisa M. Goldstein (2)
(1) Core Genotyping Facility, SAIC-Frederick, Inc., NCI-Frederick; Div Cancer Epidemiol & Genetics, NCI, NIH
(2) Div Cancer Epidemiology & Genetics, NCI, NIH
(3) Laboratory of Experimental Immunology, Center for Cancer Research, NCI, NIH
(4) Oncogenomics Section, Pediatric Oncology Branch, Center for Cancer Research, NCI, NIH

New technologies such as ''whole''-exome sequencing capture methods have led to renewed excitement about discovering more rare, Mendelian, high-risk susceptibility gene variants in humans. To assess ''whole''-exome sequence capture approaches with respect to coverage, we evaluated the content and the empirical performance of the currently available ''whole''-exome sequence capture methods [NimbleGen Sequence Capture 2.1 M Human Exome Array; Agilent SureSelect Human All Exon Kit] on three sequencing platforms [454 FLX Ti (4 runs); ABI SOLiD (1 quadrant); Illumina GA II (2 lanes)]. The protein coding sequences (CDS) reported in the RefSeq database (build 36.3) served as the gold standard. NimbleGen capture probes currently target 77% of CDS bases and Agilent capture probes target 83% of CDS bases. We observed 21.5 Mbps (65%), 25.0 Mbps (76%) and 23.4 Mbps (71%) of the 33.0 Mbps of CDS with ?8x sequence depth for Nimblegen/454, Agilent/SOLiD, and Agilent/Illumina, respectively. Since identification of rare gene variants requires high per-gene coverage, we also computed the proportion of genes with >90% of CDS bases covered with ?8X sequence depth. Only 42%, 55%, and 45% of genes were covered by Nimblegen/454, Agilent/SOLiD, and Agilent/Illumina, respectively. Coverage of entire gene CDS is presently incomplete, thus negative results must be interpreted cautiously since it is not currently possible to fully exclude most genes from consideration of harboring causal mutations.

**12**

**Replication Strategies and Rare Variants Discoveries in Genetic Studies of Complex Traits using Next Generation Sequencing Technologies**

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

For association studies, significant signals identified in the exploratory sample (stage 1) need to be replicated in an independent sample (stage 2). Two strategies can be used: (a) variant-based replication: nucleotide sites uncovered in the stage 1 are genotyped and analyzed in stage 2 or (b) gene-based replication: the gene regions implicated in stage 1 are sequenced in stage 2 and both known and novel variants are analyzed. The efficiency of the two strategies is affected by the proportions of variants discovered in stage 1 and sequencing errors. Using a rigorous population genetic framework, we demonstrate that gene-based replication is consistently more powerful. For small studies with a few hundred individuals usually <50% of the causal nucleotide sites are uncovered, but >80% of the locus population-attributable-risk can be explained by these variants. For large studies with thousands of individuals, a fairly large portion of causative variants sites can be observed. Customized genotyping can thus be a temporal solution for replicating large genetic studies if an ethnically matched stage 2 sample is analyzed. However, sequencing is a necessity if stage 1 sample is small and the discovery of novel variants is also of interest. Using an error model for downstream analysis of sequencing data, we show that the impact of current sequencing error levels on the power for mapping rare variants is minimal and the advantage of gene-based replication remains.

**13**

**Enriching Targeted Sequencing Experiments for Rare Disease Alleles**

Todd L. Edwards (1) Zhuo Song (2) Chun Li (3)
(1) Vanderbilt Epidemiology Center, Division of Epidemiology, Vanderbilt University, Nashville, TN 37203
(2) Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 3721
(3) Center for Human Genetics Research, Department of Biostatistics, Vanderbilt University, Nashville, TN 37212

Next-generation targeted resequencing of GWAS-associated genomic regions is a standard approach for discovering rare disease alleles. Many studies may use next-generation sequencing for SNP discovery, with the intent to genotype candidate SNPs captured by resequencing. This approach is reasonable, but inefficient for rare alleles if samples are not carefully selected for the resequencing experiment. We have developed an approach to estimate expected count of captured rare disease alleles, SampleSeq, to select samples for a targeted resequencing experiment. SampleSeq requires specification of prevalence and a target rare disease allele frequency range, and assumes genotypes are available at previously associated SNPs, HWE, and no recombination. SampleSeq was compared to random cases or controls, and selecting subjects for burden of nearby high-risk alleles. We simulated 1000 replicates of 2000 cases and 2000 controls for 15 regions with disease allele frequencies 0.001–0.01 and prevalences 0.01, 0.05, 0.1, 0.2. SampleSeq provided >35% higher yields of rare disease alleles, acquired >35% more samples with at least 1 disease allele, and required >33% less samples to capture the same number of disease alleles in all scenarios compared to the best alternative. This allows for smaller sample sizes in resequencing experiments, or captures rarer risk alleles. SampleSeq also can calculate the sample size needed to ensure capture of rare alleles of desired frequencies.

**14**

**An Efficient Test of Gene-Environment Interaction for Genomewide Association Studies**

Melanie Sohns (1) Juan P. Lewinger (2) Heike Bickeboller (1) Duncan C. Thomas (2)
(1) Department of Genetic Epidemiology, University Medicine Gottingen, Germany
(2) Department of Preventive Medicine, University of Southern California, USA

Complex diseases are caused by the interplay of multiple genetic and environmental factors. However, genomewide association studies (GWAS) have typically focused on genetic main effects. Classical approaches to detect gene-environment interactions ($G \times E$) such as the standard case-control test (CC) have low power in the GWAS context because of the stringent multiple testing correction required. The case-only test (CO) (Piegorsch et al, Stat Med 13:153–162) has higher power but it is not valid in the presence of population-level gene-environment associations. The standard case-control statistic can be viewed as a case-only statistic adjusted by subtraction of a control-only statistic. We propose an alternative adjustment of the standard case-only statistic where the control-only adjustment is computed from the hierarchical model proposed by Lewinger et al. (Genet Epidemiol 2007;

31:871–882) based on the entire ensemble of SNPs. We evaluated the power and type I error of our proposed G × E test under a wide range of scenarios and compared it to the standard case-control, case-only, and the G × E tests of Mukherjee & Chatterjee (MUK, Biometrics 2008;64:685–694) and Murcray et al. (MUR, Am J Epidemiol 2009;169:219–226). We found that our proposed test preserves the type I error and is more powerful than the MUK and MUR tests in most of the scenarios we considered.

**15**

**Detecting Gene-Gene/Gene-Environment Interactions for Quantitative Traits with U-Statistics**

Ming Li (1), Wenjiang Fu (1), Qing Lu (1)
(1) Michigan State University

The genetic etiology of complex human diseases has been commonly viewed as a process that involves gene-gene and gene-environment interactions. Detection of the interaction effects among genes and environmental factors remains a great challenge due to the high order epistatic complexity. Statistical approaches, such as multifactor dimensionality reduction (MDR) method and generalized MDR (GMDR), have recently been proposed to test the association of a set of genetic markers with either dichotomous or continuous traits. In this paper, we proposed a U-statistics-based method to evaluate the combined effect of multiple loci on quantitative traits with the consideration of gene-gene interactions. We used the *U*-statistics-based forward algorithm to select the potentially associated genetic markers and test the significance level of the association by permutation. Our method has the following advantages: (1) It is an non-parametric method without any assumption of the distribution of the trait or the mode of inheritance. (2) It does not require any cut off to determine the levels of the quantitative traits. (3) It is computationally efficient and can potentially be applied to the genome-wide scale. Through the simulation study, we found our method outperformed the existing approaches with both greater power and higher selection accuracy. We also illustrate our method with an application to Nicotine Dependence, using the data from the Gene Environmental Association Study Initiative.

**16**

**Gene-Environment Interactions in Genome-Wide Association Studies: A Comparative Study of Tests Applied to Empirical Studies of Type 2 Diabetes**

Marilyn C. Cornelis (1), Eric J. Tchetgen (1), Liang Liming (1), Lu Qi (1), Nilanjan Chatterjee (2), Frank B. Hu (1), Peter Kraft (1)
(1) Harvard School of Public Health
(2) National Cancer Institute

The most effective statistical approach to investigating gene-environment (G × E) interactions in the context of genome-wide association studies (GWAS) remains unresolved.Using two case-control GWAS of type 2 diabetes, we present a comparative study of five tests for interactions: (i)standard logistic-regression-based case-control; (ii) case-only; (iii)semiparametric maximum-likelihood estimation (semiMLE) (iv) an empirical-Bayes(EB) shrinkage estimator;and (v) a two-stage test.We also compared two joint tests of genetic main effects and G × E interaction: (i)joint and ii) semiMLE joint tests.Elevated body mass index was the exposure of interest.Single nucleotide polymorphisms (SNPs) with the most significant G × E interactions using the standard test were also strongly correlated with the exposure among controls. A similar, but less dramatic,pattern was observed for the EB and two-stage tests,while the case-only and semiMLE interaction tests were not correlated with tests of G–E independence among controls. Both joint tests detected markers with known marginal effects. Our findings suggest that methods which exploit G–E independence are efficient and robust options to investigating G × E interactions in GWAS.In contrast, tests that incorporate the standard G × E interaction parameter are liable to detect markers that are associated with the exposure by chance. Finally, joint tests of genetic main effects and G × E interaction can be powerful approaches to enhancing the detection of disease loci.

**17**

**An Integrative Genomic Strategy Combining Linkage and Association Analysis With Expression Profile Analysis for Localizing Genetic Variants Influencing Quantitative Traits: An Example From the San Antonio Family Heart Study**

Eugene Drigalenko (1), Anthony G. Comuzzie (1), Joanne E. Curran (1), Matthew P. Johnson (1), Melanie A. Carless (1), Jack W. Kent Jr (1), Juan Peralta (2), Thomas D. Dyer (1), Shelley A. Cole (1), Laura Almasy (1), Michael C. Mahaney (1), Eric K. Moses (1), John Blangero (1), Harald H.H. Goring (1)
(1) Southwest Foundation for Biomedical Research
(2) Universidad de Costa Rica, San Pedro, Costa Rica

We have pursued an integrative genomics approach to identify genetic variants influencing quantitative traits in the San Antonio Family Heart Study. The following data was used: (1) SNP genotype data (Illumina; 542,944 SNPs), (2) expression profiles on lymphocytes (Illumina; 16,681 transcripts), and (3) many quantitative traits. The analyses relating SNP genotype, expression level, and trait to one another involved (1) joint linkage and association study of a trait of interest, (2) association analysis of SNP hits with transcripts to reveal *cis*-regulatory potential, (3) correlating the transcripts associated with a SNP to the trait of interest, and (4) evaluating the results for consistency in directionality of effect. SOLAR program package was used. An example of the successful application is with resistin, a cytokine thought to link obesity with type 2 diabetes. Data were available for 598 participants. We identified two candidate regions: (1) not surprisingly, the structural gene RETN (chromosome 19; the *p*-values are $5.04 \times 10^{-9}$ for joint linkage and association test, $2.56 \times 10^{-6}$ for *cis* association, and $1.96 \times 10^{-5}$ for the transcript as a predictor of trait), indicating that this gene influences the level of both mRNA and protein; (2) OASL (chromosome 12, the *p*-values are $3.57 \times 10^{-8}$, $2.56 \times 10^{-6}$, $2.02 \times 10^{-3}$). Our results suggest that a multi-pronged integrative genomics approach may be of great utility for identifying candidate genes and variants influencing human complex traits.

**18**

**Bayesian Meta-Analysis of Trans-Ethnic Genome-Wide Association Studies: Application to Fine-Mapping**

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

Genome-wide association studies (GWAS) have been successful in identifying novel loci contributing effects to complex human traits. However, the extent of linkage disequilibrium (LD) across these loci means that locating specific causal variants is very difficult. One approach that may help to resolve this problem is meta-analysis of trans-ethnic GWAS, taking advantage of the differences in patterns of LD between distantly related populations for fine-mapping. Fixed-effects meta-analysis assumes allelic effects to be the same in each population. However, we would expect heterogeneity in these effects between different ethnic groups because: (i) causal variants may not be the same; (ii) exposure to interacting environmental risk factors may vary; or (iii) allele frequencies at interacting variants may not be the same. To address this challenge, I have developed novel meta-analysis methodology, TRANSMAP, which clusters populations according to relatedness and similarity in allelic effects via a Bayesian partition model. Simulations suggest that, when allelic effects are the same across ethnic groups, there is only minimal loss in power of TRANSMAP compared to fixed-effects meta-analysis. However, in the presence of trans-ethnic allelic effect heterogeneity, there are substantial gains in power for TRANSMAP, and clear improvements in the resolution of fine-mapping within associated loci.

**19**

**Phenotype-Wide Association Study (PheWAS) for Exploration of Novel SNP and Phenotype Relationships within PAGE**

Sarah A. Pendergrass (1), Kristin D. Brown-Gentry (1), Scott Dudek (1), Jose L. Ambite (2), Christy L. Avery (3), Steve Buyske (4), Congxing Cai (2), Gerardo Heiss (3), Lucia Hindorff (5), Charles Kooperberg (6), Yi Lin (6), Teri A. Manolio (7), Tara Matise (8), Lynne Wilkens (9), Megan D. Fesinmeyer (6), Chun-Nan Hsu (2), Dana C. Crawford (1), Marylyn D. Ritchie (1)
(1) Center for Human Genetics Research, Vanderbilt University, Nashville, TN
(2) Information Sciences Institute, University of Southern California, Marina del Rey, CA
(3) Department of Epidemiology, University of North Carolina, Chapel Hill, NC
(4) Department of Genetics and Department of Statistics, Rutgers University, Piscataway, NJ
(5) National Human Genome Research Institute Bethesda, MD
(6) Public Health Sciences, Fred Hutchinson Cancer Research Institute, Seattle, WA
(7) National Human Genome Research Institute, Bethesda, MD
(8) Department of Genetics, Rutgers University, Piscataway, NJ
(9) Cancer Research Center, University of Hawaii, Honolulu, HI

The current paradigm of genome-wide association studies (GWAS) tests thousands of markers for association with a single phenotype. Discovery may be enhanced by testing hundreds or thousands of markers for associations with hundreds to thousands of phenotypes. The Population Architecture using Genomics and Epidemiology (PAGE) network was established in 2008 to unite diverse, population-based studies that have a wealth of phenotypic and genotypic data for post-GWAS characterization and discovery. Using three studies (National Health and Nutrition Examination Surveys (NHANES); Women's Health Initiative (WHI); and Atherosclerosis Risk in Communities (ARIC)), we performed a Phenotype-Wide Association Study (PheWAS). Association tests stratified by race/ethnicity were performed for NHANES (93 SNPs, 152 phenotypes, ~7,000 participants), WHI (95 SNPs, 1,457 phenotypes, ~2,000–13,000 participants), and ARIC (69 SNPs, 611 phenotypes, ~4,000–10,000 participants), and results shared across studies ($P < 0.01$) were identified. Multiple novel associations were observed across studies. We have developed an analysis pipeline, including high-throughput phenotype QC and a "PheWAS-View" web interface to display these results and provide a resource for exploring novel SNP-Phenotype relationships. PheWAS has the potential to uncover novel associations, identify pleiotropy, and provide an efficient and useful resource for hypothesis generation for future studies.

**20**

**Estimation of Odds Ratios of Genetic Variants for the Secondary Phenotypes Associated with Primary Diseases**

Jian Wang (1), Sanjay Shete (1)
(1) UT MD Anderson Cancer Center

Genetic association studies for binary diseases are designed as case-control studies: the cases are those affected with the primary disease and the controls are free of the disease. At the time of case-control collection, information about secondary phenotypes is also collected. To study the secondary phenotypes, investigators are using standard regression approaches, where individuals with secondary phenotypes are coded as cases and those without secondary phenotypes are coded as controls. However, using the secondary phenotype as an outcome variable in a case-control study might lead to a biased estimate of odds ratios (ORs) for genetic variants. This is because the secondary phenotype is associated with the primary disease of interest; therefore, individuals with and without the secondary phenotype are not sampled following the principle of a case-control study design. In this article, we first demonstrated that such analyses will lead to a biased estimate of OR and proposed an approach to provide a more accurate OR estimate of genetic variants associated with the secondary phenotype. We also proposed a bootstrapping method to estimate the empirical confidence intervals for the corrected ORs. The performance of our approach was demonstrated via simulation studies as well as a real data analysis of single-nucleotide polymorphisms associated with chronic obstructive pulmonary disease using lung cancer genome-wide association data.

**21**

**A Powerful Multi-Phenotype Approach on Genome-Wide Association Studies (GWAS) to Identify Novel Pleiotropic Genes that Affected Multiple Quantitative Traits**

Yi-Hsiang Hsu (1), Xing Chen (2), Mayetri Gupta (3), David Karasik (1), James Meigs (4), L. Adrienne Cupples (3), Douglas P. Kiel (1)
(1) Hebrew SeniorLife Institute for Aging Research and Harvard Medical School, Boston, MA
(2) Harvard School of Public Health, Boston, MA
(3) Boston University, School of Public health, Biostatistics Dept., Boston, MA
(4) Massachusetts General Hospital, Boston, MA

Pleiotropy occur when multiple phenotypes were affected by same genetic variants. Previous studies proposed to simply look-up on the overlaps of association signals among univariate GWAS across multiple traits to identify pleiotropy. However, the correlation among traits may also lead to overlap in false positive/negative signals. Due to moderate genetic effects, it is inefficient to detect pleiotropy by univariate analytical framework. We propose here a new approach to test pleiotropy on GWAS using a two-stage strategy: in the first stage, we performed a multi-phenotype GWAS by modeling traits simultaneously using our newly developed empirical-weighted linear-combined test statistics (eLC); and then, we tested the pleiotropy using a simplified structure-equation-modeling on selected SNPs from the first stage. eLC directly combines correlated test-statistics with a weighted sum of univariate statistics to maximize the heritability of the overall association test. Using GWA16 simulated dataset, our eLC approach has outperformed the simple look-up on the overlaps among univariate GWAS and other multivariate methods (such as MANOVA, GEE and PCA). We applied our approach to data from the GEFOS ($n = 32{,}000$) and MAGIC ($n = 36{,}610$) consortia to identify pleiotropy on both bone density and glycemic phenotypes. Several pleiotropic effects were found, i.e. SNPs in or near *NPSR1, TNFRSF11B and TGFBI* genes that may potentially link skeleton to the energy metabolism.

**22**
**Identifying Rare Haplotypes Associated with Common Diseases through Bayesian Lasso**
Swati Biswas (1), Shili Lin (2)
(1) University of North Texas Health Science Center
(2) The Ohio State University

Identifying rare variants associated with a common disease is a challenging problem even with enormously large sample sizes. To detect associated haplotypes under the case-control sampling design, we propose a Bayesian regularized approach based on a generalized linear model (Bayes-rGLM). Bayes-rGLM employs retrospective likelihood, that is, the probability of haplotype and covariates, if available, given disease status. This formulation is more appropriate in haplotype analysis than the commonly used prospective likelihood as haplotypes cannot be always inferred exactly with phase unknown SNP data. The regularization in Bayes-rGLM is carried out through Bayesian Lasso in which the coefficients are penalized using appropriate prior distributions. The penalization of coefficients results in weeding out haplotypes that are not associated with the disease so that the associated ones, especially those that are rare, can stand out and be accounted for more precisely. We have conducted simulations under various settings involving different combinations of truly associated haplotypes, both rare and common, to investigate the Bayes-rGLM approach and compare with Hapassoc, a standard software for detecting haplotype association. Our results show that Bayes-rGLM is much more powerful in identifying associated haplotypes, especially the rare ones, when the false positive rates for the both are kept the same.

**23**
**A Powerful Approach for Rare Variants Analysis in Quantitative Traits Based Association Studies**
Dalin Li (1), David V. Conti (1)
(1) University of Southern California

Rare variants may, in part, explain some of the missing hereditability unexplained in current GWAS. The SNPs-collapsing (SC) approach has been proposed for rare-variant analysis in a gene region. However this method is based on the hypothesis that effects of the causal rare-variants are all deleterious due to natural selection. This might be untrue in practice and even slight violation of its hypothesis can greatly reduce its power. Here we propose a new approach for rare variant analysis in quantitative traits. We first show that for quantitative traits, even with small sample size and low MAF the likelihood-ratio test statistic still approximately follows chi-square distribution. Then the likelihood-ratio test statistics for all the rare variants in the region are combined to generate a sum statistic which represents the association of the given region with the outcome. The sum statistics is then compared to its null distribution which can be generated based on the correlation structure between variants. Simulations show that power of this approach is close to the SC approach when the causal variants are all deleterious, and can be much more powerful when a proportion of the variants are protective. A unique advantage of this approach is that, it allows the investigator to combine rare and common variants to get a more comprehensive view on the effect of a gene region. We propose to apply this approach in future sequencing-based studies

**24**
**Methods for Identifying Rare Variants with Bidirectional Effects on Quantitative Traits**
Qunyuan Zhang (1), Ingrid B. Borecki (1), Michael A. Province (1)
(1) Washington University School of Medicine

Collapsing multiple variants into one variable is a useful method for rare variants association analysis. Direct collapsing, however, is not valid or may significantly lose power when a collapsed group contains variants with bidirectional (i.e. some positive and some negative) effects on target traits. This can be true for quantitative traits (such as blood pressure and body mass index), regardless of whether subjects are sampled randomly from population or selectively from two extreme tails of trait distribution. We propose two possible solutions for this problem. One is splitting the variants of interest into two groups according to their effect directions and then performing collapsing analysis for the two groups separately; another is calculating a weighted collapsing score based on *p*-values from single-variant tests of effect directions and

then performing test using the score. We investigate the receiver operating characteristics (ROC) of the proposed methods using simulated data under various configurations, and show that the proposed methods achieve higher statistical power, in comparison with the regular direct collapsing method. We also demonstrate the applicability of the proposed methods using real resequencing data of candidate genes. Finally, we discuss how to reduce the bias of *p*-value calculation in the proposed methods through random permutation.

**25**

**Deep Re-Sequencing to Identify Functional Variants at the CRP Gene**

Christina T.L. Chen (1), Zach Stednick (1), Andrew N. McDavid (1), Orsalem J. Kahsai (1), Ahmad S. Zebari (1), Dylan O'Shea (1), Christopher S. Carlson (1)
(1) Fred Hutchinson Cancer Research Center

Plasma C-Reactive Protein (CRP) level is a biomarker that predicts future risk of cardiovascular disease. To better predict CRP levels, studies have looked at common single nucleotide polymorphisms (SNPs) in the CRP region. These common variants, however, only explain very little variance in CRP levels among individuals. We therefore hypothesize that there exist some rare SNPs in this region which play a functional role in determining CRP levels. To investigate whether there is any rare variant, we first used long-ranged PCR to amplify the 6 kb region spanning CRP in each of the 2000 individuals from the CARDIA cohort separately. Then we pooled groups of 96 individuals for re-sequencing on the Illumina GAIIx. We obtained an average coverage of 48,000 per base in each pool, counting only bases with high quality scores on high quality reads. In total, we identified 139 polymorphisms with 46 of these already known. The minor allele frequencies range from 0.04% to 36.8%, suggesting that we were able to confidently identify variants as rare as 1 copy among 2000 individuals. We are currently genotyping these variants and investigating whether any of these is associated with CRP levels. Our results for the CRP region suggest that this type of comprehensive, pooled sequencing of thousands of samples may prove an efficient technique to use in replicating rare variant enrichment studies such as exome re-sequencing.

**26**

**Detecting Rare Genetic Variants Associated with Complex Traits Using Resequencing Data**

Tao Feng (1), Robert C. Elston (1), Xiaofeng Zhu (1)
(1) Department of Epidemiology and Biostatistics of Case Western Reserve University

Detecting rare genetic variants underlying complex traits is topical but challenging in aspects of both design and analysis. Several statistical methods have been developed, but how successful these current methods can be in practice still needs to be tested. Rare variants segregate in families and this information has not yet been used in existing statistical methods for detecting rare variants. Here we introduce a novel statistical method for testing association that uses both data from both families and independent samples. Our simulations of a variety of disease models suggest that our method is more powerful than current existing methods. The method can be applied to resequencing data. Our study also suggests that family data are extremely useful in searching for rare variants underlying complex traits.

**27**

**The Use of Whole Exome Sequencing to Identify Rare Susceptibility Variants in Cancer Prone Families**

Lynn R. Goldin (1), Xiaohong R. Yang (1), Mary L. McMaster (1), Kevin B. Jacobs (1), Meredith Yeager (1), Michael G. Cullen (1), Xijun Zhang (1), Joseph Boland (1), Quan Chen (1), Laurie Burdett (1), Phillip R. Taylor (1), Maria T. Landi (1), Neil E. Caporaso (1), Alisa M. Goldstein (1), Stephen J. Chanock (1), Margaret A. Tucker (1)
(1) Division of Cancer Epidemiology and Genetics, National Cancer Institute

In order to identify susceptibility loci for specific cancers using families at high risk, we conducted whole exome sequencing on 3–4 members in each of 6 cancer pedigrees (3-melanoma, 2-Hodgkin lymphoma, 1-Waldenstrom Macroglobulinemia). To maximize accuracy of variant calling and fitting genetic models, we included a trio and a distantly related case per family for a total of 24 individuals. We used Nimblegen SeqCapEZ Exome V1 capture, and conducted 3–4 sequencing runs with Roche/454 FLX Ti. The gsMapper program from Roche/454 was used for alignment and variant calling. There were on average 75,000 high-confidence non-reference variants detected, 6,683 (8.9%) were novel (not in dbSNP 130), and 405 are believed to be novel non-synonymous protein coding variants per individual. Assuming a dominant model, we selected variants present as heterozygotes in all of the cases and not in any of the unaffected individuals across all pedigrees. We annotated the list of candidate genes and computed the likely effect of variants using SIFT, POLYPHEN, and SNPs&GO. Some variants are likely to be spurious because of ambiguous sequence alignment due to many homologs in the genome. Our final list of candidates for each family varied from 0 to 100. We will follow-up each of these genes in the remaining individuals from our families. Limitations of this method include incomplete coverage and the small number of samples that can be screened due to cost considerations.

**28**

**Statistical Models for Detecting Rare Variants Associated with a Disease**

Jeesun Jung (1), Inyoung Kim (2), Deukwoo Kwon (3)
(1) Indiana University, School of Medicine
(2) Virgina Tech University, department of statistics
(3) National Cancer Institute, Division of Cancer Epidemiology and Genetics

Detecting common variants associated with common disease has been successful in a framework of genome-wide association studies (GWAS) where rare variants have not been focused on to study. Due to recent advance in sequencing technologies, deep re-sequencing data becomes available to discover the rare variants influencing a disease. In this study, we propose a novel statistical method for identifying disease associated rare variants with two scenarios: (1) 1–5% of rare variants, and (2) less than 1% of rare variants. We assume the variants have

poisson or zero-inflated poisson distribution depending on rate of variants. We derive a score test statistic for testing association of rare variants. Based on simulation studies, we performed power and type I error rate studies and we have demonstrated that our proposed method is statistically robust and achieves a good power to detect association. In addition, we compared the proposed method with EM approach by the simulation studies.

**29**

**Rare Genomic Variants Contribute to Systolic Blood Pressure Variation in The Framingham Heart Study**

Berit Kerner (1), Bengt O. Muthen (1)
(1) UCLA

Hypertension is a growing health problem with serious long term consequences including heart disease, stroke and renal failure, affecting millions of individuals worldwide. We explored heterogeneity in systolic blood pressure (BP) development in 1060 males from 692 families at four different time points spanning thirty years of observation. Growth mixture modeling was used to define specific risk subclasses according to age of onset and developmental trajectories taking covariates, such as body mass index and treatment for hypertension into account. Then 550,000 SNPs were used in a genome-wide association analysis using class membership as phenotype. Taking this approach we found significant genome-wide associations between high BP early in life and the rare intronic SNP rs9928761 in the gene *CDH13* ($p = 6.4 \times 10^{-11}$), as well as with SNP rs8018017 in the gene *SLC24A4* ($p = 1.4 \times 10^{-8}$). Both genes had been previously associated with BP in other data sets. In addition, we found evidence of association with the very rare genomic variant rs10494067 in the gene *NTNG1* ($p = 5.5 \times 10^{-10}$). Stratification of patients into high and low risk groups in regard to developmental trajectories might lead to increased homogeneity in the sample and increased power to detect rare variants that might contribute to the phenotype.

**30**

**Natural Selection and Efficient Phenotype Association in Resequencing Studies**

Christopher R. King (1), Paul J. Rathouz (1), Dan L. Nicolae (2)
(1) University of Chicago Department of Health Stuies
(2) University of Chicago Department of Statistics

Projects to determine the impact of rare genetic variation on human disease with base-pair by base-pair resequencing of many participants are underway. Recent authors have recognized the importance of using population genetics to inform the hypothesized causal model, particularly with respect to the role of rare variation. We present a generalized linear mixed model which uses population genetic theory to suggest a pooling strategy for all information within a gene for efficient testing. We demonstrate analysis of whole exome sequencing of a case-control study of asthma phenotypes. In simulation we show power improvements over alternative methods in scenarios of interest. We show how to incorporate additional prior information on SNP effects from mutation type, location and phylogenetics. Our method produces easily interpreted parameter estimates and model summaries.

**31**

**Variant Calling from Low-Pass Next Generation Sequence Data in Families**

Bingshan Li (1), Wei Chen (1), Goncalo Abecasis (1)
(1) University of Michigan

Next generation sequencing is currently being employed to uncover rare variants associated with human complex traits. It becomes challenging due to high error rates of next-gen sequencing technologies. Although population based studies are powerful for common variants, family designs offer an efficient approach for association studies of rare variants. By utilizing transmission constraints within families, accuracy can be greatly increased by jointly calling variants within families. In addition, if families with multiple affected individuals are collected, rare variants are greatly enriched, making variant calling more accurate and association studies more powerful. Here we developed a likelihood framework and implemented it in software for calling variants from low-pass next-gen sequencing data in families, based on information encoded in Genotype Likelihood Format (GLF) and rigorous population genetics theories. Simulation using different error rates (0.01∼0.001) and sequence coverage (2X∼8X) show that more variants can be called with higher accuracy in families than in unrelated individuals, especially when the error rate is high and coverage is low. For the situation when only polymorphic sites in a sample are considered, family designs always have better call rate and accuracy. We are currently sequencing families in the SardiNIA project and the new software will be applied to the sequencing data for variant calling and association studies.

**32**

**An Association Detection Method for a Hybrid of Common and Rare Variants**

Yi Li (1), Jian Jun Liu (1)
(1) Genome Institute of Singapore

With the advance of next generation sequencing technology, identification of rare variants in human genomes has become practical. However, the number of rare variants in the human genome is so huge that it is seldom to achieve genome-wide association significance if we adopt single variant test method. This forces us to perform collapsing association tests, where all the variants in a region are grouped together, and tested as one variant. Since common neutral variants diminish the allele frequency difference between cases and controls greatly, common and rare variants need to be dealt with differently. Currently, two strategies were suggested: one was to assign to each variant a continuous grouping weight that depends on the allele frequency and favors rare variants by design; the other strategy groups the rare variants together, treats them as a common variant, and adopts multi-variant methods to test all the common variants in the region. One of the challenges in the second strategy is to decide on a frequency threshold to tell which variants are common or rare. Usually this threshold is disease and data dependent; and a simple brute-force strategy is to try various thresholds and report the one with the most significant association *p*-value. Here we report a different strategy which makes use of the characteristics of common and rare

variant association patterns. Simulation data shows that our method is more powerful when the signal/noise ratio is high.

**33**

**Systematic Interrogation for Mutations in Nearly 1000 Cilia Related Genes in Patients with Primary Cilia Dyskinesia Using Targeted Sequence-Capture and Massively Parallel Sequencing**

You Li (1), Michael M. Barmada (2), Steve Sabol (3), Bishwanath Chatterjee (3), Gregory Pazour (4), Margaret W. Leigh (5), Thomas W. Ferkol (6), Janette Lamb (7), Michael Knowles (8), Maimoona Zariwala (8), Cecilia Lo (1)
(1) Department of Developmental Biology, University of Pittsburgh School of Medicine, Pittsburgh PA
(2) Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh PA
(3) Laboratory of Developmental Biology, NHLBI/NIH, Bethesda MD
(4) Program in Molecular Medicine, University of Massachusetts Medical School, Worcester MA
(5) Department of Pediatrics, University of North Carolina School of Medicine, Chapel Hill NC
(6) Department of Pediatrics, Washington University School of Medicine, St. Louis MO
(7) Genomics and Proteomics Core Lab, University of Pittsburgh, Pittsburgh PA
(8) Pulmonary and Cystic Fibrosis Center, University of North Carolina, Chapel Hill NC

Primary Cilia Dyskinesia (PCD) is a genetically heterogeneous recessive disorder associated with repetitive bronchial infections and degenerative airway disease due to structural and functional defects of motile cilia in the airway. As disease causing mutations have been found in only ~30% of PCD patients, we have designed a custom sequence capture microarray containing probes for highly conserved cilia related genes which, together with massively parallel sequencing, can be used to identify novel PCD causing mutations. As proof of principle, we examined a patient with PCD of indeterminate genetic etiology using our capture array and ABI SOLiD sequencing. Sequence data were analyzed using CLC Genomic Workbench software. Data analysis pipeline and filtering parameters were optimized based on validation results obtained by Sanger sequencing. Approximately 40% of reads were mapped to the targeted region with an average coverage >40x. Over 90% of the identified variants matched content in dbSNP and the 1000 Genomes database, and were filtered out. From the remaining variants we recovered a DNAI1 mutation known to be present in this patient, and also two novel DNAH3 missense mutations - all present in a heterozygous state. In addition, 48 coding variants in other cilia genes were identified as potential novel disease-causing mutations. These findings suggest that next-generation sequencing holds promise for systematic searches for mutations in diseases affecting the cilium.

**34**

**GRANVIL: Rare Variant Analysis of Genome-Wide Association Studies**

Reedik Magi (1), Andrew P. Morris (1)
(1) WTCHG, University of Oxford

Genome-wide association (GWA) studies have been very successful in identifying common alleles contributing moderate effects to complex traits. However, a large portion of heritability remains unexplained, but could be attributed to rare variants. However, alternative analysis methods must be used for gaining information about rare variants. In GWA analysis, the power of detecting true positive associations decreases in the case of causal variants with low minor-allele frequency. Therefore methods, which combine the information of several markers in genomic regions, should be preferred rather than using single-marker data.
GRANVIL (Gene- or Region-based ANalysis of Variants of Intermediate and Low frequency) is an implementation of a method to perform rare-variant analysis of binary or quantitative phenotypes. The method is based on accumulation of minor alleles of rare or uncommon markers discovered through dense genotyping or resequencing data. Both directly genotyped markers and imputed marker data can be used for analysis. Association analyses are based on gene- or other pre-defined regions, determined by analyst. GRANVIL is an open source software package and can be downloaded from http://www.well.ox.ac.uk/GRANVIL/.

**35**

**Refining Genetic Associations with Hepatitis C Treatment Response Using Massively Parallel Sequencing of Pooled DNA Following a Genome Wide Association Study**

Katherine R. Smith (1), Vijayaprakash Suppiah (2), Graeme Stewart (2), David R. Booth (2), Jacob George (2), Melanie Bahlo (1)
(1) The Walter and Eliza Hall Institute of Medical Research, Melbourne
(2) Westmead Millennium Institute, Sydney

Four independent genome-wide association studies (GWAS) have identified associations between response to treatment of chronic Hepatitis C infection and the IL28A/IL28B region on chromosome 19. We aimed to further refine association signals in this region using massively parallel sequencing (MPS) of individuals from one of these studies (Nat Genet 2009;41:1100–1104).
We separately pooled DNA from 100 responders and 99 non-responders. For each pool, we amplified a continuous 100,000 base pair (bp) region around IL28A/IL28B and generated around 17 million pairs of 75 bp sequence reads using Illumina's Genome Analyzer II technology. Following alignment to the human genome and variant calling, we used an allele-based test to compare minor allele frequencies (MAFs) between the two cohorts at each variant.
Around 40 variants showed stronger association with treatment response than the best variant identified by the original GWAS achieved in this study. We will present these results as well as a comparison of MAFs to those obtained from the original genotyping arrays and from HapMap, where possible. In addition, we will discuss some statistical considerations that arise from association testing of MPS data generated from pooled DNA. These include the non-independence of reads and the choice of association test considering that genotype data is not readily available.
This study provides insight into the feasibility of using MPS of pooled DNA to identify causal variants following a GWAS.

**36**

**Tiled Regression Applied to Cafe-Au-Lait Macule Burden in Neurofibromatosis Type 1 Sequence Data**

Heejong Sung (1), Alexander Pemov (2), Yoonhee Kim (1), Juanliang Cai (1), Alexa J. Sorant (1), Jennifer L. Sloan (3), Sarah L. Coombes (2), Jim Mullikin (4), Pedro Cruz (4), Douglas R. Stewart (2), Alexander F. Wilson (1)
(1) Genometrics Section, Inherited Disease Research Branch, NHGRI/NIH, Baltimore, MD
(2) Genetic disease research branch, NHGRI/NIH, Bethesda, MD
(3) Genetics and Molecular Biology Branch, NHGRI/NIH, Bethesda, MD
(4) NIH Intramural Sequencing Center (NISC), NHGRI/NIH, Bethesda, MD

The cause of the variation in phenotypic severity in neurofibromatosis type 1 (NF1) is unknown. The method of "tiled regression" is applied to sequence data in order to identify all independently significant sequence variants (of those tested) that affect the number of cafe-au-lait macules (CALM) in NF1. Five candidate genes for modifiers in NF1 (*DPH2, MSH2, MSH6, MLH1, MED21*) were sequenced in 99 Caucasians. One-hundred eighteen sequence variants were identified: 70 were rare variants (RV, defined as MAF$<$0.05) and 48 were common variants (CV). These variants were grouped into "independent" tiles, defined by linkage disequilibrium (LD) block, hotspot region and gene. Using tiled regression, multiple and stepwise regression models were applied to each tile separately; higher level stepwise regression models were then applied across tiles to identify variants that independently predict macule number. Several coding schemes were considered: (1) number of minor alleles for each CV and RV, (2) presence or absence of minor allele for each CV and RV, (3) presence or absence of the minor allele for each CV, and a new collapsed variant for each tile, coded as presence or absence of a minor allele at any RV within the tile, in other words, collapsing multiple RVs into a single tile-wide variant. One CV (rs4660761) was selected in the final model in nearly all models. A collapsed variant was selected only with LD-based tiles.

**37**

**Ascertained Samples for Targeted Resequencing Increases Power When Identifying Causal Variants**

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

Genome wide association studies continue to identify Single Nucleotide Polymorphisms (SNPs) associated with disease. However, often these associated SNPs are not the causal SNPs, and usually investigators employ targeted resequencing to close in on the causal variant. Typically a subset of affected individuals is randomly selected from the cases used for the GWAS. Here we propose a design to sample the individuals to be used for the targeted resequencing that increases the power to detect the causal variant. We show that ascertaining from a subset of cases who possess the risk allele increases power to detect a causal variant. To evaluate this method, we simulated a disease with a causal SNP. Our simulation studies reveal that ascertaining individuals for targeted resequencing substantially increases the power to detect a causal SNP without increasing the false-positive rate.

**38**

**Quality Score Recalibration In Second Generation DNA Sequencing Data**

Xiting Yan (1), Murim Choi (2), Wei Zheng (1), Richard P. Lifton (2), Hongyu Zhao (3)
(1) Keck Laboratory, Yale University
(2) Department of Genetics, Yale University
(3) Department of Epidemiology and Public Health, Yale University

Second generation sequencing technology has become commonly used in genomics studies. For example, it has been used to sequence individual genomes and detect sequence variations among them. To reduce false positive results due to sequencing errors, the base quality score generated from the sequencer, which is designed to quantify the probability of a base being wrongly called, is often taken into account when detecting sequence variations. However, these scores may not correspond to the true calling rates, leading to complications in detecting sequence variations. In this presentation, we discuss potential biases in quality scores of the Illumina Solexa sequencing data and introduce methods to recalibrate the base quality score to improve statistical power to detect sequence variations.

**39**

**Genetic Association Studies of Copy-Number Variation: Should Assignment of Copy Number States Precede Testing?**

Patrick Breheny (1), Brooke Fridley (2)
(1) University of Kentucky
(2) Mayo Clinic

Recently, structural variation in the genome has been implicated in many complex phenotypes. Using current SNP arrays, researchers are able to investigate the impact not only of SNP variation, but also of copy-number variants (CNVs) on phenotypes. The most common approach to analysis involves estimating, at the level of the individual genome, the underlying number of copies present at each region. Once this is done, tests are performed for an association between copy number state and phenotype. Uncertainty in copy-number assignment is often not incorporated into the testing procedure, and can diminish the power of the subsequent association test. An alternative approach is to conduct hypothesis testing for an association between copy number and phenotype using continuous measures of copy number at the individual marker level, followed by the aggregation of neighboring test results to determine CNVs associated with the phenotype. Here, we explore the strengths and weaknesses of these two approaches using both simulated and real data from a pharmacogenomic study of the chemotherapeutic agent gemcitabine. Our results indicate that pooled single-locus testing is capable of offering a dramatic increase in power ($>$10-fold) over CNV-level testing, particularly for small CNVs. However, there are also settings in which CNV-level testing is superior; understanding these trade-offs is an important consideration in conducting association studies of structural variation.

**40**

**A Novel Copy Number Estimation Method Removes Artifacts in GWAS Studies**

Wenjiang J. Fu (1), Ming Li (1), Yalu Wen (1), Qing Lu (1),
(1) Dept Epidemiology, Michigan State Unversity

Analysis of GWAS data often suffers from a number of array artifacts, even after careful array normalization procedures. Among them, population stratification, batch effect and genomic waves are major barriers in association analysis and CNV studies, and frequently lead to false positives or false negatives. Although a number of methods have been introduced to remove these artifacts, very often, special procedures are required before association analysis. In this paper, we present a novel method based on oligoarrays, which yields cleaned copy number estimation and accurate SNP genotype calls through a newly developed model—the probe intensity composite representation model (Wan et al., 2009, Nucleic Acids Research doi:10.1093/nar/gkp559) and a recent improvement using an MA ratio method. This novel approach not only removes array artifacts, including unequal footing and unequal scaling, without requiring across-array normalization, but also removes the genomic waves, batch effect and population stratification. We demonstrate with the Wellcome Trust Case-Control Consortium data and an NCI breast cancer GWAS study data that the novel method leads to a powerful association test that reveals new findings in previous GWAS studies.

**41**

**Batch Effects with Illumina 1M Onmi-Quad Chip in CNV Detection**

Sulgi Kim (1), Debby Tsuang (2), Ellen Wijsman (1)
(1) University of Washington Medical Center
(2) Veterans Affairs Puget Sound Health Care System

Copy number Variation (CNV) is structural variation that is suspected to be causal in many diseases. CNV detection with some SNP genotyping platforms (e.g. Affymetrics) is known to be sensitive to various normalization procedures. This has not been assessed with the Illumina platforms including the 1M Omni-Quad panel. We performed a CNV study with 90 trios sampled through schizophrenic offspring. The Illumina Omni-Quad 1M chip accommodates up to 4 samples per chip with 8 chips genotyped at the same time in a batch. In addition, trios were genotyped by one of two scanners in two separate batches at different time points. When we jointly normalized the data from the different batches, SNP call rates were not affected. However, CNV calls identified patterns among different batches leading to investigation of the existence and effects of batch effects. One Quality Control (QC) issue is the presence of Genomic Wave (GW, pattern of Log R Ratio (LRR) along the chromosome), which can lead to false positive CNV calls. In our sample, batch-specific normalization reduced GW and correlation in CNVs called across batches. Our study illustrates the underappreciated problem of batch effects in the Illumina 1M panel. These results also suggest that even with an Illumina panel, more effort should be expended on initial QC, before devoting effort to different CNV detection algorithms since different algorithms applied to biased data can give correlated, but false CNV calls.

**42**

**Copy Number Variations in germ-line DNAs in high-risk African-American men with prostate cancer**

Elisa M. Ledet (1), Xiaofeng Hu (2), Marilyn Li (2), Diptasri M. Mandal (1)
(1) Department of Genetics, Louisiana State University Health Science Center- New Orleans, LA
(2) Hayward Genetics Center, Tulane University School of Medicine, New Orleans, LA

Prostate cancer (PCa) is the most common type of cancer affecting African American men. Genomic copy number variations (CNVs) have been detected in prostate tumors, but changes in copy number have not been studied in germ-line DNAs from high-risk African-American PCa cases. To identify PCa associated CNVs, Array comparative genomic hybridization (aCGH) was used to analyze 30 individuals (21 affected males and 9 unaffected males) from 10 families. A combined targeted/whole-genome array system was used with Agilent $4 \times 44$ 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. Initial analyses using CGH Analytics 3.5 (Agilent Technologies) showed CNVs on chromosomes 1, 13, 14, 15, and 17 in prostate cancer cases, which were not reported in the Database of Genomic Variants (DGV). A unique CNV, found in 17 affected males, (spanning 34 Kb) on chromosome 14 has been validated with qPCR. These variations may contribute to the high prevalence and mortality of PCa in African American men. Validation on other CNV regions is ongoing.

**43**

**Accounting for SNP and CNV Information in Association Studies**

Gaelle Marenne (1), Nuria Malats (2), Emmanuelle Genin (3)
(1) CNIO, Spain; INSERM UMR-S946, France.
(2) CNIO, Spain.
(3) INSERM UMR-S946, France.

Apart from Single Nucleotide Polymorphisms (SNPs), there also exist in the Human Genome structural variations such as Copy Number Variants (CNVs) that can influence phenotypic traits. Despite the possibility that the SNP-arrays offer to access both the number of copies and the alleles present at SNPs located in a copy-number variable region, few efforts have been made so far to combine information. Indeed, association studies usually focus either on the effect of the allele or on the effect of the number of copies on disease risk.
In this work, we studied the power of these approaches, as well as the combined approach that considers CNV-specific genotypes (A, AA, AAB…), in situations where the disease risk is influenced by both the number of copies and the allele type. We simulated case-control data at a locus under different scenarios. To take into account CNV-specific genotypes effects, different logistic models were considered. Among them, a model where a trend is assumed on the number of B alleles in the genotype (Trend) and the model implemented in Plink with two explanatory variables, the sum of the number of A and B alleles and their difference. Interestingly, providing that the SNPs are not excluded because of calling uncertainties, we found that tests based on the bi-allelic genotypes show little power losses in comparison with those combining SNP and CNV information. Among the models that combine both information, Trend has the highest power under most scenarios.

**44**

**Investigating the Association Between Rare Copy Number Variation and Developmental Anomalies in Autism Spectrum Disorders**

Alison K. Merikangas (1), Elizabeth A. Heron (1), Richard J.L. Anney (1), Aiden P. Corvin (1), Louise Gallagher (1), The Autism Genome Project (2)
(1) Trinity College Dublin
(2) The AGP Consortium

Studies of recurrent copy-number variation (CNV) in Autism Spectrum Disorders (ASD) have provided potential insight into its broader phenotypic manifestations such as learning disability, physical characteristics, and seizures as opposed to the strict autism phenotype. The goal of this study is to investigate whether individuals carrying rare CNVs that impact genes implicated in ASD or intellectual disability (ID) are more likely to manifest general developmental anomalies and associated syndromes than those with other rare CNVs. This analysis will include participants with rare CNVs in the Autism Genome Project. Subjects were genotyped on the Illumina 1M SNP microarray, and CNVs were called using combinations of three algorithms. Detailed phenotypic information was collected using a broad range of measures of adaptive and cognitive functioning. The prevalence of developmental anomalies among those individuals having rare CNVs that impact ASD-implicated and ID genes will be compared to those who have rare CNVs that do not impact these gene sets. Such differences could identify specific genetic variants underlying the broader phenotypic manifestations of ASD and may provide insight into the genetic and phenotypic heterogeneity of this highly heritable and complex disorder.

**45**

**Copy Number Variants Segregate in Extended Families with Autism Spectrum Disorder**

Daria Salyakina (1), Holly N. Cukier (1), Deqiong Ma (1), James M. Jaworski (1), Michael L. Cuccaro (1), John R. Gilbert (1), Scott M. Williams (2), Ram K. Menon (3), Margaret A. Pericak-Vance (1)
(1) John P. Hussman Institute for Human Genomics, University of Miami
(2) Center for Human Genetics Research, Vanderbilt University, Nashville, TN
(3) Department of Epidemiology and Department of Obstetrics & Gynecology, Rollins School of Public Health, Emory University, Atlanta

In the multifaceted etiology of autism spectrum disorder (ASD), CNVs must now be incorporated into our understanding of this complex disease. We sought to take advantage of our unique resource of 46 extended ASD families to detect potential CNVs that segregate to all autism family members and thus may contribute to ASD susceptibility. Using the genotyping of over 1 million sites from the Illumina Human genotyping array, we applied the PennCNV algorithm to recognize deletions and duplications. Families were then evaluated for co-segregation of CNVs in ASD patients. We were able to identify and validate nine deletions and seven duplications that were segregating in multiple affected individuals within 13 extended families. Eight genes were disrupted by 6 of these CNVs; deletions were identified on chromosomes 1p34.1 (*ZSWIM5*), 4q31.3 (*LRBA*) and 6q11.1(*KHDRBS2*) and duplications were located on 4p16.3 (*TNIP2*), 7p21.2 (*ICA1*, *NXPH1*) and 10q 23.2 (*BTAF1*, *FGFPB3*). In addition, our CNV findings overlap with larger, previously reported genomic rearrangements in patients with ASD, ADHD, developmental delay, mental retardation, and various congenital defects at 3p26.3–p26.2, 3q26.1, 4p16.3, 10p12, 12q24.31?qpter, 13q, 15q11.2–q.13.3. Our data support the hypothesis that the misregulation of genes either by altering their dosage, affecting their regulatory elements, or creating novel functional regions, may be an important genetic mechanism in ASD.

**46**

**The Role of Common Copy Number Variation in Amyotrophic Lateral Sclerosis (ALS)**

Louise V. Wain (1), Nick R.G. Shrine (1), Christopher Shaw (2), John F. Powell (2), John Hardy (3), Pamela Shaw (4), Karen E. Morrison (5), Robert H. Brown (6), Richard Orrell (3), Boniface Mok (3), Lyle J. Palmer (7), Jennie Hui (8), Alan L. James (9), Bill Musk (10), Ammar Al-Chalabi (2), Martin D. Tobin (1)
(1) Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK
(2) MRC Centre for Neurodegeneration Research, King's College London, Institute of Psychiatry, London, UK
(3) Department of Molecular Neuroscience and Reta Lilla Weston Laboratories, UCL Institute of Neurology, London, UK
(4) Academic Unit of Neurology, University of Sheffield Medical School, Sheffield, UK
(5) Department of Clinical Neurosciences, School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK
(6) Day Neuromuscular Research Laboratory, Department of Neurology, University of Massachusetts Medical School, Worcester, MA, USA
(7) Centre for Genetic Epidemiology and Biostatistics, University of Western Australia
(8) Molecular Genetics, PathWest Laboratory Medicine WA, Nedlands, Western Australia
(9) Department of Pulmonary Physiology/West Australian Sleep Disorders Research Institute, Sir Charles Gairdner Hospital, Australia
(10) Busselton Population Medical Research Foundation, Sir Charles Gairdner Hospital, Australia

The underlying genetic causes of amyotrophic lateral sclerosis (ALS) have not yet been fully elucidated. Although previous studies have identified several common Single Nucleotide Polymorphism (SNP) variants that increase disease susceptibility, independent studies of the role of copy number variation (CNV) in ALS have not yielded strong and replicable findings. The recent publication of the most comprehensive map of copy number variation across the genome provides the basis for a more robust assessment of the potential contribution of common variants to ALS compared to previous approaches which combined both CNV discovery and calling. We used a previously published likelihood ratio testing approach (CNVtools) to robustly genotype and test association at CNVs across the genome in a total of 671 ALS cases and ~600 controls genotyped on the Illumina 610 K platforms. This approach allows incorporation of the uncertainty of

the copy number calls into the association testing reflecting the inherent noisiness of copy number measurement. We used different methods of combining the information across the SNP probes within each CNV region to identify the optimal approach for our data. We took into account the potential for differential bias in CNV measurement between cases and controls, which could arise from DNA quality differences or differences in laboratory sample preparation procedures between cases and controls.

**47**

**Genome-wide Algorithm for Detecting CNV Associations with Diseases**

Yaji Xu (1), Bo Peng (1), Emily Y. Lu (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 DNA copy number variations (CNVs). Nonparametric and model-based statistical algorithms have been developed to detect CNVs from SNP data. However, these algorithms lack specificity to detect small CNVs due to the high false positive rate when calling CNVs based on the intensity values. Association tests based on detected CNVs therefore lack power even if the CNVs affecting disease risk are common. By combining an existing Hidden Markov Model (HMM) and the logistic regression model, we developed a new genome-wide algorithm to detect CNV associations with diseases. In our simulation studies, we showed that for large CNVs (♯SNPs$>=10$), although the association tests based on PennCNV calls gave more significant results (average $p$-value = 6.26e−25 when relative risk RR = 9.0), the new algorithm was able to capture the signals ($p$ = 3.12e−11 when RR = 9.0) as PennCNV did especially when the RR is high. However for small CNVs (♯SNPs$<10$), the new algorithm provided smaller average $p$-values in all the cases ($p$ = 7.54e−17 when RR = 9.0), and can capture the signals whereas PennCNV cannot ($p$ = 0.020 when RR = 9.0) even the RR is high. We conclude that the new algorithm is more sensitive and can be more powerful in detecting CNV associations with diseases than the existing HMM algorithm, especially when the CNV association signal is weak and a limited number of SNPs are located in the CNV.

**48**

**Detecting the Effect of Two Genes Involved in a Complex Disease in Family Based Studies**

Marie-Claude Babron (1), Michel Guilloud-Bataille (2), Emmanuelle Genin (1), Marie-Helene Dizier (1)
(1) Inserm U946
(2) Inserm U669

Not accounting for interaction in association analyses may reduce the power to detect the variants involved. Here, we investigate different designs to detect the effect of two disease variants using family-based association tests. Power is evaluated under different Gene-Gene interaction models, by simulating trio families typed for 300 SNPs, two of which causal. We define 4 different strategies: (S1) one-snp analysis of all SNPs, (S2) two-snp analysis of all possible SNPs pairs, (S3) lax preliminary selection of SNPs followed by two-snp analysis of all selected SNP pairs, (S4) stringent preliminary selection of SNPs, each being later paired with all the SNPs for two-snp analysis. Testing all SNP pairs (strategy S2) is never the best design, except when there is an inversion of the gene effect (flip-flop model). Testing individual SNPs (S1) is the most efficient when the two genes act multiplicatively (no interaction). Designs S3 and S4, with preliminary selection, are the most powerful for non-multiplicative models. Their respective powers depend on the level of symmetry of the model in terms of penetrance matrix and allele frequency. Since the true genetic model is unknown, we cannot conclude that one design outperforms another. Similar conclusions would be reached for a larger number of SNPs such as in a GWAS. Although a stronger multiple test correction would be needed, the power ranking of the different designs would remain unchanged.

**49**

**Multi-locus Analysis of Genome-wide Association Studies Using the Random Forest Algorithm: Application to a Study of melanoma**

Jennifer H. Barrett (1), Rosa Parisi (1), Mark M. Iles (1), D. Timothy Bishop (1)
(1) University of Leeds

In genome-wide association (GWA) studies usually each SNP is analysed separately; there is concern that this method has low power to detect the effect of SNPs that only have a strong effect on risk in combination (interacting SNPs). Random Forest (RF) is a classification algorithm from which importance measures can be derived, ranking variables by their contribution to classification accuracy. Since RF is a multi-locus method, it is hypothesised that these may be more successful at identifying interacting SNPs. We have investigated the properties of RF through simulation and applied RF to a GWA study of melanoma conducted by GenoMEL (Bishop DT et al, Nature Genetics 2009;41:920–925). The choice of RF tuning parameters (e.g. number of trees, terminal node size) affected performance. Considering a range of two-locus models, RF was found to have similar, sometimes lower, power than logistic regression of all pair-wise SNP interactions. The ranking of SNPs by RF was compared to that from single locus analysis, assuming an additive model, in the melanoma study. The SNPs showing association by single locus analysis that have since been replicated were also highly ranked by RF. Some SNPs were more highly ranked by RF because their association differed from an additive model. One of the more highly ranked SNPs (rs872071 in IRF4) has since been shown to be associated with melanoma (Duffy et al, Am J Hum Genet, in press); others will be investigated in phase 2 of the GenoMEL study.

**50**

**Rapid Testing of Gene-gene Interactions in Genome-wide Association Studies.**

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

One of the key challenges in performing genome-wide association (GWA) interaction studies is computational burden. To address this challenge, we propose a two-stage

gene-gene (G × G) interaction testing strategy. In the first stage, all pairs of SNPs are tested using a rapid interaction test. For dichotomous traits, this is equivalent to the PLINK "fast epistasis" approach. For a continuous trait, the distribution is dichotomised at the median, with half the sample treated as cases and half as controls. In the second stage, all pairs of SNPs with rapid G × G interaction *p*-values meeting a pre-defined significance threshold are carried forward for testing in a traditional, but more computationally demanding, generalised linear modelling (GLM) framework. Within this framework, we can assess the significance of interaction effects, adjusting for the marginal effects of SNPs, and allowing for potential non-genetic risk factors, which cannot be fully assessed in the first stage. We have undertaken a detailed simulation study to investigate the performance of the two-stage strategy in comparison to the full GLM applied to all pairs of SNPs. Over a range of models of interaction for quantitative trait association, our results highlight a minimal loss in power by employing a significance threshold of $P<10^{-4}$ in the rapid G × G interaction testing stage, but an expected reduction in computing time of more than 95% in a GWA study of 500,000 SNPs.

**51**

**Alternative Risk Cell Definitions Based on Ranking Improve Performance of Model-based Multifactor Dimensionality Reduction for Epistasis Detection**

Tom Cattaert (1), Jestinah M. Mahachie John (1), Francois Van Lishout (1), Kristel Van Steen (1)
(1) University of Liege

When searching for epistasis, parametric approaches have severe limitations. Alternatively, the non-parametric Multifactor Dimensionality Reduction method, MDR [1], can be applied. It handles the dimensionality problem by pooling multi-locus genotypes into two groups of risk: High Risk and Low Risk. However, MDR involves computationally intensive cross-validations because it is based on prediction. Recently, Calle et al. [2] proposed the Model-Based MDR (MB-MDR) method. Association tests are applied to identify risk cells and to test the final one-dimensional construct, and a third risk category, that of "No Evidence for Risk", is introduced. Using association tests allows multiple models to be proposed, no longer requires cross-validations, and flexibly deals with different outcome types.
We explore alternative definitions for risk cell identification, avoiding association testing at the cell risk assignment stage. In particular, multi-locus genotype cells are ranked according to their case to control ratios, hereby assuming that higher ranks are indicative for higher risk to disease. The three risk groups are now defined based on this ranking and the new one-dimensional construct is tested for association with the trait. We evaluate the impact of the newly introduced risk category definition on MB-MDR performance via extensive simulations.

References:
[1] Am J Hum Genet, 2001;69:138–147.
[2] Technical Report No. 24 2007. Department of Systems Biology, Universitat de Vic

**52**

**Using Biological Knowledge to Discover Higher Order Interactions in Genetic Association Studies**

Gary K. Chen (1), Duncan C. Thomas (1), Angela P. Presson (2)
(1) University of Southern California
(2) University of CA, Los Angeles

The recent successes of genome-wide association studies (GWAS) have revealed that many of the replicated findings have explained only a small fraction of the heritability of common diseases. One hypothesis that investigators have suggested is that higher-order interactions between SNPs or SNPs and environmental risk factors may account for some of this missing heritability. Searching for these interactions poses great statistical and computational challenges. We propose a novel method that addresses these challenges by incorporating external biological knowledge into a fully Bayesian analysis. The method is designed to be scalable for high-dimensional search spaces (where it supports interactions of any order) because priors that use such knowledge focus the search in regions that are more biologically plausible and avoid having to enumerate all possible interactions. We provide several examples based on simulated data demonstrating how external information can enhance power, specificity, and effect estimates in comparison to conventional approaches based on maximum likelihood estimates. We also apply the method to data from a GWAS for breast cancer, revealing a set of interactions enriched for the Gene Ontology terms growth, metabolic process, and biological regulation.

**53**

**A Two-stage Approach to Test for Gene-gene Interactions in Family Data Based on Within-family and between-Family information**

Lizzy De Lobel (1), Kristel Van Steen (2)
(1) Ugent
(2) University de Liege

The search for susceptibility loci in gene-gene interactions imposes a methodological and computational challenge for statisticians due to the large dimensionality inherent to the modelling of gene-gene interactions or epistasis. In an era where genome-wide scans have become relatively common, new powerful methods are required to handle the huge amount of feasible gene-gene interactions and to weed out the false positives and negatives from these results. One solution to the dimensionality problem is to reduce the data by preliminary screening of markers to select the best candidates for further analysis. Ideally, this screening step is statistically independent of the testing phase. To obtain two independent steps to test for associations in family data, we can split up the genotypic information in a between-family and within-family component as is done in the QTDT. Those two components are orthogonal so that one of the components can be used for screening and the other can be used for testing. The QTDT proposes a definition of these components for one locus. In our research, we define analogous components for gene-gene interactions and investigate the proporties of this screening technique in di erent types of simulations.

References:
[1] Abecasis GR, Cardon LR, Cookson WOC. 2000. A General Test of Asso-ciation for Quantitative Traits in Nuclear Families, American Journal of Human Genetics, no. (66):279–292.

**54**
**Detecting Gene-gene Interactions in Complex Diseases using Lasso Penalized Regression**
Sarah L. Keildson (1), Andrew P. Morris (1), Martin Farrall (1)
(1) Wellcome Trust Centre for Human Genetics

Coronary artery disease (CAD) is a complex genetic disorder that occurs when arteries become hard and narrow thereby restricting the heart's supply of blood. Since complex phenotypes are generally controlled by multiple genes that may interact with each other and the environment, unravelling the genetic architecture of these diseases is a continual challenge. Incorporating multi-locus analysis methods that allow for gene-gene interactions into genome-wide association studies (GWAS), however, may increase power to detect disease loci. Lasso penalization, a multi-locus selection method, was carried out on 34144 SNPs across 2100 candidate genes, to identify potential interaction effects that may contribute to disease association in 3146 CAD cases and 3352 controls. Lasso was used to select the 2000 most significant marginal SNPs and then again to examine all possible interactions among them. The first step identified previously reported main effect SNPs significantly associated with CAD. Significant interactions identified in stage two were analysed further, comparing the goodness of fit between models containing interaction terms and those with only main effects. In three cases, the interaction term significantly improved the fit of the main-effects model and the univariate *p*-values of the marginal SNPs were less significant than their interaction terms. This two-stage lasso selection may thus have the potential to detect gene-gene interactions in GWAS.

**55**
**WITHDRAWN**

**56**
**WITHDRAWN**

**57**
**A Mann-Whitney Based Whole Genome-wide Association Study Finds Significant Gene-gene Interaction for Type 2 Diabetes**
Qing Lu (1), Changshuai Wei (1), Chengyin Ye (1), Robert C. Elston (2)
(1) Michigan State University
(2) Case Western Reserve University

The potential importance of gene-gene interaction has long been recognized. However, identifying interactions has been a great challenge, especially when millions of genetic markers are involved. We here propose a Mann-Whitney based approach for whole genome-wide gene-gene interaction search. It extends the traditional univariate Mann-Whitney test to assess the joint association of multiple loci, considering all levels of possible interaction. Because only one overall significance test is conducted, it avoids the issue of the multiple testing. The approach adopts a computationally efficient algorithm, making whole genome-wide gene-gene interaction analysis feasible in a reasonable time on a high performance personal computer. We evaluated the approach using both simulation and a real data application. By applying the approach to 24 Type 2 diabetes (T2D) susceptibility genes, we identified a four-locus model strongly associated with T2D in the Wellcome Trust (WT) study (permutation $p$-value $< 0.001$), and replicated the finding in the Nurses' Health Study/Health Professionals Follow-up (NHS/HPFU) study ($p$-value = 3.03E-11). We also conducted a whole genome-wide gene-gene interaction search on nearly 500K loci. This approach identified four loci jointly associated with 2D ($p$-value = 1.29E-5) in the WT study, the significance of this association reaches 4.01E-6 in the NHS/HPFU study.

**58**
**Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data**
Jestinah Mahachie John (1), Francois Van Lishout (1), Kristel Van Steen (2)
(1) University of Liege
(2) University of Liege

Many common human diseases and traits are believed to be influenced by several genetic and environmental factors, each factor potentially having a modifying effect on the other. Understanding the effects of genes on the development of these complex diseases and traits in humans is a major aim of genetic epidemiology. It has been suggested by [1] on the Spanish Bladder Cancer study and [2] on the KORA study, Germany, that Model-Based Multifactor Dimensionality Reduction (MB-MDR) for dichotomous traits is a useful method for identifying high-order gene-gene interactions. From a theoretical point of view, [3] have shown the excellent power of MB-MDR for dichotomous traits and in the presence of noisy data. Although MB-MDR has been successfully applied to quantitative traits as well [2], the power of MB-MDR in identifying significant epistasis effects on a quantitative trait in the presence of noisy data has never been explored. The aim of our simulation study is to evaluate the power and type 1 error rates of Model-Based Multifactor Dimensionality Reduction for quantitative traits to detect gene-gene interactions in the presence of both error-free and noisy data. Considered sources of error are genotyping errors, missing genotypes, phenotypic mixtures and genetic heterogeneity. Technical Report No. 24 (2007), University of Vic 2. Allergy (2009), Early View 3. Ann Hum Genet (2010), In Press

**59**
**On a Test for Gene-gene Interaction**
Indranil Mukhopadhyay (1), Abhishek Pal Majumder (1), Partha Pratim Majumder (2)
(1) Indian Statistical Institute
(2) National Institute of Biomedical Genomics

In genetic association studies, a major challenge is to identify gene-gene interactions that may increase the disease risk. In the context of case-control study, the Multifactor-Dimensionality Reduction (MDR) method provides a robust and nonparametric approach to test whether a combination of polymorphic markers is associated with susceptibility to the disease. However, when the data are generated on a large number of SNPs leading to an astronomical number of SNP combinations, the method becomes computationally extremely heavy. We have developed a statistical test that will directly test for association of disease susceptibility with a number of genotype or allele combinations. This test is based on a

simple multinomial probability model and have shown that the asymptotic distribution of the test statistic under the null hypothesis is a mixture of chi-squares. This helps in obtaining the asymptotic power of the test without relying on simulations. For the procedure to be applicable, the total sample size, but not frequencies of individual genotype combinations, needs to be large. We have carried out simulation experiments to establish that our proposed test procedure can identify small differences in frequencies of genotype combinations between cases and controls with high statistical power, even for moderate sample sizes.

**60**

**Gene-gene Interaction Analysis Accounting for Family Structure: The Development of a New ASSOC Module in S.A.G.E.**

Junghyun Namkung (1), Daniel Baechle (1), Kevin Cartier (1), Robert Elston (1)
(1) Case Western Reserve University

Various types of multifactor dimensionality reduction have been implemented for the analysis of family data with the aim of identifying gene-gene interaction (G × G): PGMDR (Am J Hum Genet 2008(83), 457–467), MDR-phenomics (Am J Hum Genet 2007(81), 1251–1261) and FAM-MDR (Plos one 2010(5), e10304). PGMDR and MDR-phenomics use transmission disequilibrium test (TDT) type statistics, which only use genetic information conditional on parental genotypes, hence severely restricting the amount of available information used. FAM-MDR adopts a polygenic effect model to account for familial correlations. It applies the multi-locus genotype classification algorithm on the residuals from a regression model. Since, it performs many Wald type tests to determine risk groups for each genotype combination, it requires a large amount of computation time.
The ASSOC program in the S.A.G.E. package provides general association tests for both family data and case-control data, singly or simultaneously without conditioning on parental genotypes. In addition, it enables us to control for covariate effects easily. In this study, we propose and implement a new gene-gene interaction analysis method that can account for family structure, using the model fitting algorithm in ASSOC to compute individual score statistics. We show the performance of ASSOC G × G analysis option using simulated data and compare it with PGMDR and FAM-MDR.

**61**

**Exploring Gene-gene Interaction in Case-only Samples with Permutation Testing**

Ricardo Segurado (1)
(1) Trinity College Dublin

A case-only test for interaction consists of a simple test on a $3 \times 3$ or $2 \times 2$ table of genotype, or allele at one locus, versus genotype or allele at a second locus. Advantages of this method include a low type II error rate, the utility of obtaining an estimate of the interaction relative risk, and absence of a need to ascertain a control sample. However, multiplicity of tests in explorations of epistatic effects, as well as an unacceptably high type I error rate inherent to case-only samples in the presence of population stratification are problems. Here, an implementation of an allellic case-only test for interaction is presented which includes calculation of a gene-pair-wise statistical significance by MC permutations, used to correct for multiple tests while accounting for within-gene linkage disequilibrium, and for missingness in a consistent manner, and its properties explored in simulated and real data. Permutation-based strategies for adjusting for any correlation structure that is not of interest, and adjustments for cryptic population substructure using additional population membership information, are introduced.

**62**

**Knowledge-driven Multi-locus Analysis Reveals Gene-gene Interactions Influencing HDL Cholesterol Level in two Independent Biobanks**

Stephen D. Turner (1), Richard L. Berg (2), Dana C. Crawford (1), Joshua C. Denny (3), James G. Linneman (2), Catherine A. McCarty (2), Peggy L. Peissig (2), Luke V. Rasmussen (2), Dan M. Roden (4), Russell A. Wilke (5), Marylyn D. Ritchie (1)
(1) Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University Medical Center
(2) Biomedical Informatics Research Center, Marshfield Clinic Research Foundation
(3) Departments of Biomedical Informatics and Medicine, Vanderbilt University Medical Center
(4) Departments of Medicine and Pharmacology, Vanderbilt University Medical Center
(5) Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University

Cardiovascular disease (CVD) is the ♯1 cause of morbidity and mortality in industrialized nations. High density lipoprotein (HDL) particles may attenuate the development of CVD. Here we present a knowledge-driven multi-locus analysis of HDL cholesterol concentration using 3740 individuals from a GWAS in the Marshfield Personalized Medicine Project (PMRP). By incorporating information from biological pathways, protein families, and other data sources into our analysis, we identified several replicating gene-gene interactions that impact HDL level in the general population. Using 3740 individuals from the PMRP we identified 194 models with a significant interaction term $p$-value ($P < 0.01$) and a significant model fit ($P < 0.05$). Of these, 7 models replicated in 980 individuals in the Vanderbilt Genome Electronic Records (VGER) project with a significant interaction term ($P < 0.05$) and model fit ($P < 0.10$). Both the PMRP and VGER are part of the electronic Medical Records and GEnomics (eMERGE) network—a consortium of sites with DNA biobanks linked to medical records for genetic research. Several of these replicating models included interactions between genes known to be involved in HDL homeostasis, as well as novel interactions. The methodology and results presented here illustrate the utility of knowledge-based multi-locus analysis for the discovery of novel gene-gene interactions, and demonstrate possible unexplored mechanisms involved in HDL cholesterol homeostasis.

**63**

**A Detailed View on Model-Based Multifactor Dimensionality Reduction for Detecting Gene-gene Interactions in Case-control Data in the Absence and Presence of Noise**

Kristel Van Steen (1), Tom Cattaert (1), Malu L. Calle (2), Scott M. Dudek (3), Jestinah M. Mahachie John (1), Francois Van Lishout (1), Victor Urrea (2), Marylyn D. Ritchie (3)

(1) University of Liege
(2) University of Vic
(3) Vanderbilt University

Analyzing the combined effects of genes and/or environmental factors on the development of complex diseases is a great challenge from both the statistical and computational perspective, even using a relatively small number of genetic and non-genetic exposures. Several data mining methods have been proposed for interaction analysis, among them, the Multifactor Dimensionality Reduction Method (MDR), which has proven its utility in a variety of theoretical and practical settings. Model-Based Multifactor Dimensionality Reduction (MB-MDR), a relatively new MDR-based technique that is able to unify the best of both non-parametric and parametric worlds, was developed to address some of the remaining concerns that go along with an MDR-analysis. Whereas the true value of MB-MDR can only reveal itself by extensive applications of the method in a variety of real-life scenarios, here we investigate the empirical power of MB-MDR to detect gene-gene interactions in the absence of any noise and in the presence of genotyping error, missing data, phenocopies and genetic heterogeneity. For the considered simulation settings, we show that the power is generally higher for MB-MDR than for MDR, in particular in the presence of genetic heterogeneity, phenocopies and epistasis models involving markers with low minor allele frequencies.

**64**

**Software for Whole Genome-wide Gene-gene Interactions Analysis**

Changshuai Wei (1), Qing Lu (1)
(1) Department of Epidemiology, Michigan State Univ.

Following the recent interaction study on a small number of known loci, one of the natural next steps is genome-wide gene-gene interaction analysis. The genome-wide interaction study explores all genetic variants on the genome and holds great promise to uncover novel interactions. Such study requires intense computation and memory use, in particular, when considering high order interactions and thus traditional statistical tool may not fit such purpose. We here build computational efficient software to facilitate the whole genome wide gene-gene interaction analysis. The software implements two statistical genetic approaches that we previously developed. The first approach, a Mann-Whitney multi-locus approach, is designed to capture large effect loci and interactions, while the second one, a trees-assembling Mann-Whitney approach, is designed to detect small effect genetic variants and interactions. We optimize the C++ program for both approaches and make the whole genome wide gene-gene interaction feasible on a high performance personal computer. It takes about five days to analyze the whole genome data (up to 500,000 SNPs) using the Mann-Whitney multi-locus approach. For trees-assembling Mann-Whitney approach, we first implement a filter algorithm to preselect a set of loci and then conduct the analysis on the filtered dataset. The analysis of 10000 SNPs using the trees-assembling Mann-Whitney approach takes nearly 30 minutes on a high performance personal computer.

**65**

**Genome-wide Association Analysis of Gene × Gender Interaction of Neuroticism in the Netherlands population sample from NESDA and NTR registries**

Nagesh (Nash) R. Aragam (1), KeSheng Wang (1)
(1) East Tennessee State University

The reported heritability of neuroticism is equal or greater than heritability estimates for major depressive disorder (MDD). The gender differences for neuroticism has been found to moderate the prevalence of MDD in females. However, few genome-wide analyses of gender differences has been reported to date. A genome-wide association study of gene × gender (g × e) interactions for neuroticism as a endophenotype for MDD using a sample of 2748 individuals (902 males and 1846 females) with 437547 SNPs is presented here. Neuroticism as a quantitative trait was tested with linear regression using PLINK software, while g × e interaction was tested with a covariate gender file for each individual. We identified 34 neuroticism associated SNPs with $P<10^{-4}$. The best SNP was rs4806846 in TMPRSS9 gene ($p=7.79\times10^{-6}$) located at 19p13.3. Four other SNPs, rs220549 in GRIN2B, rs1046329 in SGCA, rs6757820 and rs4510687 showed strong associations with neuroticism. Haplotype analyses of above loci indicated stronger associations with neuroticism than single-marker analyses. In addition, we found 46 SNPs showing significant gene × gender interactions for neuroticism with $P<10^{-4}$. The best g × e SNPs were rs2430132 ($p=5.37\times10^{-6}$) located at 1q25 near FDPSL1; rs17674783 ($p=6.92\text{x}10^{-6}$) near SNTG1 and rs2612437 ($p=9.30\text{x}10^{-6}$) near MYOCD. Identifying specific genes associated with interactions of gender may provide further clues to the development of neurotic behaviors in men and women.

**66**

**Random Effect Joint Meta-Analysis of SNP and SNP by Environment Interaction Effect Estimates Obtained from Regression Models**

Han Chen (1), Alisa K. Manning (1), Josee Dupuis (1),
(1) Boston University

Joint meta-analysis of single nucleotide polymorphism (SNP) main effect and SNP by environment interaction effect estimates obtained from regression models has been proposed to identify novel risk loci that may not have been uncovered when ignoring gene by environment interactions. However, most joint meta-analyses use the fixed effect model to estimate the variability of regression estimates. When heterogeneity is present, spurious associations may be observed using the fixed effect model. The random effect model takes both within and between study variances into consideration. In the meta-analysis of only one parameter such as the SNP regression main effect, the random effect model has a larger variance estimate and a wider confidence interval for the effect size than the fixed effect model, and hence it yields fewer spurious associations due to heterogeneity. We propose a random effect approach to jointly meta-analyze SNP main effect and SNP by environment interaction effect estimated from regression models in the presence of heterogeneity. We develop a method of moments estimator for the between study covariance matrix to properly account for both within and between study variances. In our simulation, this estimator performs well in estimating the standard errors of SNP main effect and SNP by environment interaction effect estimates in the joint meta-analysis, and provides more adequate coverage of confidence intervals than the fixed effect model.

**67**

**Adaption of an Empirical Bayes Approach to Investigate Gene-environment Interactions in Large Consortia**

Rebecca Hein (1), Nilanjan Chatterjee (2), Garcia-Closas Montserrat (3), Chang-Claude Jenny (1)
(1) German Cancer Research Center (DKFZ)
(2) National Cancer Institute, NIH
(3) University of Cambridge

The empirical-Bayes (EB) procedure tests for interaction and exploits the G–E independence assumption but does not rely on this assumption (Mukherjee and Chatterjee, 2008). The method, which can closely maintain the desired type I error, can have increased power compared to the case-control approach.
The BCAC is an international consortium where investigators combine data from many studies to reliably assess genetic risks. G × E analyses will be employed to assess environmental modification of genetic risks and improve detection of susceptibility loci. BCAC involves population-based (pb) studies (providing control samples that are considered to be representative for the population from which the cases were drawn) as well as non-population-based (npb) studies with non- representative controls. While reliable estimates for genetic main effects are expected based on pb and npb studies, estimates of G × E effects may however be distorted using unrepresentative control samples.
To overcome this problem, one could restrict case-control analyses to pb studies, which would decrease the sample size or perform case-only analyses, which may yield false positive results due to G–E dependence. Alternatively, one could use an adapted EB approach that makes use of all available samples while aiming to maintain the prespecified significance level.
We perform simulations to estimate type I error and power of the different approaches and will present the results.

**68**

**A Bayesian Model Averaging Approach to Gene-Environment Interaction in a GWAS**

Dalin Li (1), David V. Conti (1)
(1) University of Southern California

Scan for G × E interaction had been mostly neglected in current GWAS. The conventional case-control design for G × E interaction suffers from low power. In contrast, the case-only analysis/design can be biased when its assumption of G–E independence is violated. Building upon previous work for combining the case-only and case-control analyses, we propose a model averaging approach for G × E interaction that leverages the fact that in a GWAS most of the SNPs are not associated with either the outcome nor with the E. We first show mathematically that when there is no G–E association or main effect of G, removing the corresponding terms in the log-linear model reduces the variance of the interaction estimate and thus increases the power to detect the G × E interaction. Then this method averages over the sub-models with and without the two terms, weighted by the posterior probability of each model. We show that in single-marker analysis this approach can be even more powerful than the case-only analysis when the G–E independence hypothesis holds. Furthermore the type I error rate is more robust than the case-only analysis when the assumption is violated. Increased power or increased robustness can be obtained with larger sample size. In genome-wide simulations, this approach has better overall performance with higher True Discovery Rates and lower False Discovery rates. An example data analysis is used to demonstrate the advantages of this approach.

**69**

**Genotype-Based Association Mapping of Complex Diseases: Gene-Environment Interactions with Multiple Genetic Markers and Measurement Errors in Environmental Exposures**

Iryna Lobach (1), Ruzong Fan (2), Raymond J. Carroll (2)
(1) New York University, School of Medicine
(2) Texas A&M University

With the advent of dense single nucleotide polymorphism genotyping, population-based association studies have become the major tool for identifying human disease genes and for fine gene mapping of complex traits. We develop a genotype-based approach for association analysis of case-control studies of gene-environment interactions in the case when environmental factors are measured with error and genotype data are available on multiple genetic markers. To directly use the observed genotype data, we propose two genotype-based models: genotype effect and additive effect models. The proposed risk functions can directly incorporate the observed genotype data while modeling the linkage disequilibrium information in the regression coefficients, thus eliminating the need to infer haplotype phase. Compared with the haplotype-based approach, the proposed estimating procedure can be much simpler and significantly faster. In addition, there is no potential risk due to haplotype phase estimation. Further, by fitting the proposed models, it is possible to analyze the risk alleles/variants of complex diseases, including their dominant or additive effects. To model measurement error, we adopt the pseudo-likelihood method by Lobach et al. (2008). Performance of the proposed method is examined using simulation experiments. An application of our method is illustrated using a population-based case-control study of association between calcium intake with the risk of colorectal adenoma development.

**70**

**Meta-analysis of Gene by Environment Interaction for Disease Outcomes**

Alisa K. Manning (1), Ching-Ti Liu (1), Josee Dupuis (1), L. Adrienne Cupples (1)
(1) Boston University

Interplay between genetic and environmental factors and their interactions contributes to the etiology of complex diseases. Because large sample sizes are needed to detect statistical gene by environment (G × E) interaction, it may be useful to meta-analyze aggregate results from multiple studies to assess the association between a disease and genetic loci (G) and account for possible interaction with an environmental factor (E). There are several sampling strategies and analytic methods to detect G × E interaction affecting risk of developing a disease. When G and E are independent, case-only designs have been found to be the

most efficient in detecting G × E interaction, as G × E interaction induces an association between E and G, producing an estimate of the OR of the G × E effect. For case-control samples, analytic methods include a 2-df joint test of the G and G × E logistic regression coefficients, and reducing the number of G × E tests by screening by either G or E effects. We investigate methods for meta-analyzing results obtained from different sampling strategies and/or analytic methods and present results of simulation studies where scenarios include differing distributions of E among the samples and differing proportions of case-only and case-control samples. This work shows that expanding the search beyond main effects meta-analysis to the meta-analysis of G × E interactions for disease outcomes may uncover additional genetic susceptibility loci contributing to disease risk.

**71**

**Sample Size Requirements to Detect Gene-Environment Interactions in Genome-wide Association Studies**

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

The standard analysis in a GWAS scans for genetic main effects and ignores the potentially useful information in the available environmental exposure data. Recently proposed methods to detect G × E interactions have shown increased power relative to traditional approaches. Among these, two-step analyses have been proposed to prioritize the large number of SNPs tested to highlight those likely to be involved in a G × E interaction. Murcray et al (2009) proposed screening on a test that models the G–E association induced by an interaction in the combined case-control sample. Alternatively, Kooperberg et al (2008) suggested screening on genetic marginal effects. In both methods, SNPs that pass the respective screening step at a pre-specified significance threshold are followed up with a formal test of interaction in the second step. We propose a hybrid method that combines these approaches by allocating a proportion of the experiment-wise significance level to each test. We show that the Murcray et al. approach is often the most efficient method, but that the hybrid approach is a powerful and robust method for nearly any underlying model. As an example, for a GWAS of 1 million SNPs including a single disease marker with minor allele frequency 0.15, and a binary exposure with prevalence 0.3, the Murcray, Kooperberg and hybrid methods are 1.90, 1.27, and 1.87 times as efficient, respectively, as the traditional case-control analysis to detect an interaction effect size of 2.0.

**72**

**Interaction Between Maternal and Offspring Genotypes at Different Loci**

Jeremie Nsengimana (1), Jennifer H. Barrett (1)
(1) Leeds Institute of Molecular Medicine, University of Leeds, UK

The interaction between maternal and offspring genotypes at two different loci may be more common than the often considered interaction at the same locus. In testis cancer, for example, maternal hormones have long been associated with the disease risk and the gene most associated with the disease in genomewide association studies (*KITLG*, see e.g. Rapley et al., Nat Genet 2009;41:807–810) is activated by estrogen. It can be hypothesized that maternal genes controlling estrogen level in the uterus interact with offspring *KITLG*. More generally, a locus acting through both maternal and offspring genomes might interact with a second locus acting solely through the mother or the offspring. We have extended Weinberg's log-linear method (Weinberg et al., Am J Hum Genet 1998;62:969–978) to the interaction between maternal and offspring genotypes at different loci. We are running simulations of family trios to evaluate this test and preliminary results indicate that our test is valid and has reasonable power. With 500 trios, an interaction relative risk (RR) of 2.0 (risk allele frequency 0.3 on both loci) is detected with 91% power ($a = 0.05$) while a dominant marginal RR of 1.5 is detected with power 55% for offspring genotypes and 66% for maternal genotypes. We will present the results from simulations of different scenarios and will discuss the impact of ignoring this type of interaction when it does exist.

**73**

**How Robust are Gene-Environment Interaction Methods to Deviations from Assumptions about the Interaction Functions**

Gang Shi (1), D.C. Rao (1)
(1) Washington University School of Medicine

Association analysis based on Gene-environment interactions (GEI) is attracting increasing interest for dissecting genetic architectures of complex traits. Our GEI methods assume age-dependent genetic effects using a Gaussian function in the variance components framework. We evaluated the robustness of our inference to deviations from the underlying functional forms of the interactions. We simulated a quantitative trait locus (QTL) whose effect depended on age through four functions: linear, exponential, logistic, and Gaussian. Linkage analysis with Gaussian function was conducted on data with four different types of underlying (true) interactions. We also carried out linkage analysis with the function that was actually used for simulation. We found that the logarithm of odds ratio (LOD) scores using Gaussian function are very close to those with correct functions, which suggests that testing GEI using Gaussian function is very robust. We also computed LOD scores based on a conventional variance component model without interactions, and found that modeling interaction with a Gaussian function that does not necessarily agree with the underlying interaction function can vastly improve statistical power. Therefore, the Gaussian function is not only appropriate for modeling genetic effects that have a peak at a certain age, but also for capturing general types of nonlinear interactions. Partly supported by grants GM28719, HL095054, and HL054473.

**74**

**Evaluating Associations and Interactions of Spontaneous Clearance of Hepatitis C Infection using Logic Regression**

Genevieve Wojcik (1), David Thomas (2), Priya Duggal (1)
(1) Johns Hopkins Bloomberg School of Public Health
(2) Johns Hopkins Medical Institutions

To elucidate the immune pathways involved with spontaneous clearance of Hepatitis C virus (HCV), a candidate

gene study was conducted on 1,130 markers in 1,788 individuals from 7 cohorts drawn from Europe and North America. Because these cohorts had high levels of heterogeneity in risk factors for HCV clearance, including ethnicity, HIV status, gender and route of transmission, there was a large amount of bias present. Principal components analysis (PCA) was used to adjust for population substructure, and subjects were then stratified by HIV status, gender and transmission. A logistic regression analysis using an additive model was run on each of the strata, resulting in a severe loss of power due to small sample sizes. To avoid this loss of power, logic regression was used to analyze the data. This method evaluates the effect of random Boolean expressions which can include any combination of the non-genetic covariates and markers. While controlling for the first two PCA vectors, the role of the three non-genetic covariates remained consistent and strong. There were also numerous interactions between markers that were both contingent upon as well as independent from these three non-genetic covariates. With logic regression it is possible to analyze highly heterogeneous data to elucidate possible interactions as well as main effects, while not decreasing the study's sample size through stratification.

**75**

**Adjusting for Covariates in Logistic Regression Models**

Chao Xing (1), Guan Xing (2)
(1) UT Southwestern Medical Center
(2) Bristol-Myers Squibb Company

A conventional wisdom in classic linear regression is that adjusting for covariates associated with the response variable can improve the precision of estimates by reducing the residual variance; however, covariate adjustment in logistic regression models always leads to a loss of precision. Nonetheless, this loss of precision does not always result in a loss of power. When the genetic and risk/preventive environmental factors are independent and do not have interaction effects on the disease, it is always more efficient to adjust for the predictive covariates. Recently, Kuo and Feingold (Genet Epidemiol 2010;34:246–253) compared the power of three logistic regression models to detect genetic effects, concluded that ''the most commonly used approach to handle covariates–modeling covariate main effects but not interaction–is almost never a good idea'', and recommended modeling only the genetic factors without covariate adjustment in genome-scanning. A person's genetic background is determined from birth, thus in most cases it is not unreasonable to assume it is independent of his/her subsequent environmental exposure, which has been a key assumption in some study designs to investigate gene-environmental interaction. If we assume that only a small proportion of genetic variants interact with the known environmental factors on disease susceptibility, then, contrary to the conclusion by K&F, we recommend adjusting for predictive covariates at the genome-scanning stage.

**76**

**One-step and Two-step Gene-set Analysis with Application to Alcohol Dependence Data**

Joanna M. Biernacka (1), Jennifer R. Johnson (1), Gregory D. Jenkins (1), Colin L. Colby (1), Brooke L. Fridley (1)
(1) Mayo Clinic

By assessing the evidence for association of a trait with a set of SNPs, gene-set analysis (GSA) incorporates knowledge regarding biological pathways. We considered two broad approaches to GSA of SNP data: The commonly used 1-step method in which all SNPs in a gene set are used without consideration of gene-level effects, and a 2-step method where SNPs in each gene are first used to evaluate association with the gene and then gene effects are aggregated to test association with the gene set. For the 1-step GSA we used Fisher's method to combine SNP $p$-values, while for the 2-step GSA we ran principal component analysis for genes followed by Fisher's $p$-value combination. Because of non-independence of SNPs, permutations must be used to assess pathway significance, particularly for the 1-step approach. In the 2-step approach, when effects of SNPs in a gene are modeled jointly, the effect of SNP dependence on the $p$-value combination step is reduced, allowing for a quick screening of pathway effects without permutations. Both types of approaches were used to test for association between KEGG pathways and alcohol dependence using data from the Study of Addiction: Genetics and Environment (SAGE). No significant pathway associations were detected after correction for multiple testing. With the 2-step approach the most significant pathway was ''Synthesis and degradation of ketone bodies'' (uncorrected $P < 0.001$; $p = 0.016$ with the 1-step approach).

**77**

**The Genome-Wide Annotation Repository (GWAR)**

William S. Bush (1)
(1) Center for Human Genetics Research, Vanderbilt University

The explosion of high-throughput data collection approaches for biological sciences and the analysis of their results has dramatically expanded the amount of information available for annotation of existing result sets and the generation of new hypotheses. Often, the results of an individual analysis is published in a stand-alone database or is simply released as flat-files, which limits their utility in other analyses. To address these issues, we constructed the Genome-Wide Annotation Repository (GWAR) to centrally house and inter-related analysis results and genomic annotations from multiple sources. These include SNP-based annotations, such as expression QTLs (eQTLs) and sites of selection, and gene-based annotations, such as KEGG or Reactome pathway membership or protein structure families. Currently, information within the repository can be used in a variety of ways to conduct analysis of genome-wide association data, including enrichment analysis using PARIS, and epistasis analysis using Biofilter and ATHENA. As new data sources are added to the repository, they become available for downstream analysis tools allowing novel re-analysis of existing GWAS data. Possible analyses include enrichment by protein family or transcription factor binding site rather than traditional pathways, or epistasis analysis among eQTLs for genes in a relevant biological process from the Gene Ontology (GO).

**78**

**Comparisons of Integrative Methods for Genomic Data using Sparse Canonical Correlation Analysis**

Prabhakar Chalise (1), Brooke L. Fridley (1), Leiwei Wang (2)
(1) Department of Health Sciences Research, Mayo Clinic College of Medicine
(2) Department of Molecular Pharmacology and Experimental Therapeutics

One integrative analysis approach to inspect the relationships between the data sets is based on canonical correlation analysis (CCA). However, when the number of variables far exceeds the number of subjects, such in the case of genomic studies, traditional CCA method is not applicable. To overcome this issue, sparce CCA (SCCA) for multiple data sets has been proposed incorporating a ''shrinkage'' penalty. We compared two existing methods for SCCA (Parkhomenko and Witten) using data from a pharmacogenomic study involving the gemcitabine drug pathway which consists of three data sets: Expression (31 variables), SNP (749 variables) and Cytotoxicity (8 variables). Analysis based on the selection of penalty terms as outlined by the authors resulted in non-sparse solutions, with 12–31 expression probe sets, 310–675 SNPs and 4–8 cytotoxicity phenotypes selected. This lack of a sparse solution is a result of the selection of the sparseness parameters based on the correlation and the trade-off between the maximum correlation and the sparsity of the variables. We next implemented a two step procedure for carrying out SCCA, by adding a further variable filtering using BIC function, as outlined by Zhou et al. Application of this additional filter resulted in a sparse-solution, with 1–2 expression probe sets, 1–5 SNPs, 1 cytotoxicity phenotype selected across the two methods. In addition, simulation studies are going on to assess the performances of the SCCA methods.

**79**

**New Candidate Loci for eGFRcrea from GWAS Leveraging Gene Annotation**

Daniel Chasman (1), on behalf of the CKDGen Consortium (2)
(1) Brigham and Women's Hospital, Boston MA 02215
(2) International collaboration

A recent genome-wide meta-analysis of glomerular filtration rate estimated from creatinine (eGFRcrea) from the CKDGen Consortium identified 24 relevant loci (Nat Gen 42:376), collectively accounting for only a small proportion (~1.4%) of the suspected heritability. The missing heritability may be explained by many, possibly hundreds, of additional common but weakly associated variants. To find additional associations for eGFRcrea, we first selected the top 25 or 200 genes that were functionally related to each previously reported gene according to GRAIL (PLoS Gen 5:e1000534). Second, we identified SNPs meeting gene-wide significance thresholds for eGFRcrea in these candidate genes using the Women's Genome Health Study (WGHS), a GWAS of 23,294 European women. Third, we replicated selected SNPs in a meta-analysis for eGFRcrea including 25 cohorts with >45,000 samples from the CKDGen Consortium. Choosing the top 25 genes, the method advanced 32 SNPs, all in different genes, to the replication stage, among which 26 (81%) had effect estimates consistent with the WGHS ($p = 0.0003$) and false discovery rate (FDR) of 0.07. For the top 200 genes, the method advanced 155 SNPs, 109 of which were consistent with the WGHS and had FDR 0.31. While our preliminary results require replication, they imply a large amount of undiscovered common variation associated with eGFRcrea, and emphasize the potential utility of integrating prior biological knowledge into genome-wide genetic analysis.

**80**

**Prioritizing SNPs for Functional Studies using a Bayesian Latent Variable Model**

Brooke L. Fridley (1), Ed Iversen (2), Ya-Yu Tsai (3), Ellen L. Goode (1), Tom Sellers (1)
(1) Mayo Clinic
(2) Duke University
(3) Moffitt Cancer Center

Numerous GWAS have identified general loci harboring phenotype-associated alleles. One difficult question facing researchers is how to prioritize SNPs for functional studies. Often, a list of the top *M* SNPs is determined based on the association *p*-value. However, complexities arise when multiple GWAS analyses are completed (e.g., for subgroups of subjects) and when biological knowledge is of interest. We propose a Bayesian latent variable model (BLVM) for incorporating observed ''features'' about a SNP to estimate a latent ''quality score'', with SNPs prioritize based on the posterior probability distribution of the quality score rankings. We illustrate the method using data from an ovarian cancer GWAS of 1815 cases (1070 serous subtype) and 1900 controls, using the following SNP features: *p*-value from analysis of all cases, *p*-value from analysis of serous subtype only, and minor allele frequency. On chromosome 20, the 1st-ranked SNP (ranked in top 5 markers 46.8%) from the BLVM ranked 2nd and 7$^{th}$ based on *p*-values from analyses of all cases and of serous subtype, respectively, and the 2nd (46.3%) ranked SNP from BLVM ranked 1st and 9$^{th}$ in analyses of all cases and serous subtype, respectively. However, the top SNP based on serous subtype analysis (ranked 197 for analysis of all cases), ranked 42nd (12.6%). As these results show, the BLVM is useful for integrating multiple SNP ''features'' to prioritize loci for functional studies.

**81**

**Testing for Weak Signal**

Peter J. Lipman (1), Michael Cho (2), Per Bakke (3), Amund Gulsvik (3), Xiangyang Kong (4), Sreekumar Pillai (4), Edwin Silverman (2), Christoph Lange (1)
(1) Harvard University
(2) Channing Laboratory, Brigham and Women's Hospital
(3) Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway; Institute of Medicine, Univ of Bergen, Norway
(4) GlaxoSmithKline

When a statistical analysis involves multiple tests, such as a Genome-Wide Association study (GWAS) that involves upwards of one million SNP-association tests, one must correct the nominal *p*-values to account for the multiple opportunities to achieve statistical significance. With con-

servative corrections (such as the widely-used Bonferroni correction), one often fails to reject the null hypothesis when the null hypothesis may in fact be false. Despite the great success of GWAS, the majority of loci remains undiscovered, but can be identified by large-scale follow-up studies. In this communication, we develop an overall statistical test that is applied to the set of most promising association tests from the original study. The overall test will conclude whether there is evidence that the null hypothesis is in fact false in the tested set of association tests, even though no individual test achieves a global significance level. That is, we create a test that examines if one or several (weak) signals exist in the data. The conclusions of the test are used to determine whether a follow-up study on the top hits is recommended. This single test is statistically straightforward and easily implemented. Using simulations, power is assessed under realistic scenarios of GWAS. The test can also be modified to help determine the number of top hits which should be further examined. We illustrate the method with an application to a chronic obstructive pulmonary disease GWAS dataset.

**82**

**Bayesian Path Analysis with Variable Selection for Integration of Multiple Genomic Data Types**

Steven Lund (1), Leiwei Wang (2), Brooke L. Fridley (2)
(1) Iowa State University
(2) Mayo Clinic

Drug response is most likely due to a complex relationship involving genetic variation, mRNA, etc. along with external environmental factors. Advances in technologies have made it common to collect measurements from multiple types of genomic data on a single set of samples. However, standard analysis methods examine the association of the phenotype with a single data type at a time. To overcome this deficiency, we propose a modeling approach which combines path analysis and stochastic search variable selection into a Bayesian model, allowing one to determine both direct and indirect effects on the phenotype. The model was applied to mRNA and SNP data from a pharmacogenomic study and demonstrates its ability to identify similar effects detected by the standard "one-at-a-time" approach. Results from a simulation study with 2 SNPs affect the phenotype as well as indirectly affecting phenotype through the mRNA expression of the corresponding genes, show the Bayesian model (one-at-a time analysis approach) detecting the direct SNP effects in 57% (14%) and 52% (19%) of the simulations, indirect SNP effects in 52% (14%) and 81% (19%) of simulations and direct mRNA effects in 52% (10%) and 76% (24%) of the simulations. In conclusion, joint analysis of multiple types of genomic data with the proposed model will generate novel hypotheses related to the relationship between genomic variation and complex phenotypes.

**83**

**Genomic, Transcriptomic, and Pathway-based Analysis of Carotid Intima-Media Thickness**

Phillip E. Melton (1), Joanne E. Curran (2), Melanie Carless (2), Matthew P. Johnson (2), Tom D. Dyer (2), Jean W. MacCluer (2), Eric K. Moses (2), Harald H. Goring (2), John Blangero (2), Laura Almasy (1)
(1) Southwest Foundation for Biomedical Research
(2) Southwest Foundation for Biomedical Research

Detection of subclinical atherosclerosis through measurement of carotid intima-media thickness (IMT) is an important tool for assessment of cardiovascular disease risk. Previous GWAS of IMT have not produced any significant associations. This study used GWAS results to investigate biological pathways influencing IMT phenotypes in 750 Mexican American participants from the San Antonio Family Heart Study. A GWAS (with Illumina's HumanHap 550 K BeadChip) was performed using internal and common carotid IMT near and far wall measurements under an additive measured genotype association model. As with past GWAS, no SNP was significant after correction for multiple testing. Genes within ranges 10 kb, 25 kb, 50 kb, 100 kb, 150 kb, and 250 kb upstream and downstream of marginally significant ($P < 0.01$) SNPs were entered into Ariadne Pathway Studio to identify relevant biological pathways and to investigate how the assignment of intergenic SNPs to genes might impact these results. The WNT signaling ($p = 1.04 \times 10^{-7}$) pathway remained significant after correction for multiple testing. We then investigated the relationship between lympochyte RNA expression data and IMT. The top gene was glutamate-cysteine ligase, catalytic subunit (*GCLC*), whose transcript is significantly correlated ($p = 7.8 \times 10^{-5}$) with internal carotid IMT. We conclude that through the use of transcript data and the investigation of biological pathways that we are able to detect genetic signals not identified by GWAS alone.

**84**

**Sharpening Pathway-Based Analyses using Hierarchical Clustering: Results of a Prostate Cancer Genome-Wide Association Study**

Idan Menashe (1), Dennis Maeder (1), Stephen Chanock (1), Philip S. Rosenberg (1), Nilanjan Chatterjee (1)
(1) Division of Cancer Epidemiology and Genetics, NCI, NIH

Pathway analyses compliment primary scans of genome-wide association studies (GWAS) to indentify overrepresentation of susceptibility loci in predefined gene-sets.
We applied two pathway analysis methods: gene-set enrichment analysis (GSEA), and adaptive rank-truncated product (ARTP) to the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer GWAS (1172 cases and 1157 control), using pathways with 10-100 genes from three publically available resources (BioCarta, KEGG, and PID). An artificial pathway composed of 31 genes associated with prostate cancer was used as a positive control. Pathways with FDR < 0.2 were considered noteworthy.
A high concordance was seen between results of GSEA and ARTP ($R = 0.72$; $P \sim 0$). Of the 445 pathways tested, only the "Action of Nitric Oxide in the Heart" pathway from BioCarta was noteworthy using ARTP ($p = 0.0002$; FDR = 0.09). The positive control pathway was ranked 2nd using ARTP ($p = 0.0027$; FDR = 0.60), but only 12th using GSEA ($p = 0.0182$; FDR = 0.54). Next, we used hierarchical clustering to merge pathways with similar gene content and reduce the number of gene-sets to 167. Post clustering, one gene-sets was noteworthy by both GSEA and ARTP ($p = 0.0008$; FDR = 0.13, and $p = 0.001$; FDR = 0.18 respectively), and three other gene-sets were noteworthy by ARTP only.
These results suggest that gene-set clustering may improve power of both GSEA and ARTP, and facilitate detection of pathways underlying prostate cancer predisposition.

**85**

**Pathway Genetic Analysis of Trait-Extreme Subjects Reveals Epistasis and Sex-by-gene Interactions in Essential hypertension.**

Pei-an Betty Shih (1), Atanas Kamburov (1), Jason H. Moore Moore (2), Brinda K. Rana Rana (1), Manjula Mahata Mahata (1), Sushil K. Mahata Mahata (1), Trey Ideker Ideker (1), Daniel T. O'Connor O'Connor (1)
(1) University of California, San Diego
(2) Dartmouth Medical School

While the GWAS have uncovered credible genetic loci, together these loci only explained less than 1% of phenotypic variance in hypertension. To probe the ''missing heritability'', this study undertook a pathway candidate gene approach to examine potential effect modifiers such as epistasis and gene-by-sex interactions. Using an extreme-phenotype case-control study design, we assessed gene-by-sex interactions and epistatic relationships of 49 hypertension candidate genes chosen from 4 pathways. The non-parametric MDR analysis revealed a two-locus model of *IL6* with *CHGA* to successfully predict cases, and significant gene-by-sex interactions for *COMT*, *IL6*, and *CHGA* (permutated $p$-value $< 0.001$). Parametric methods confirmed associations and demonstrated that BP residuals explained in this study population were increased by 4% of the total variance when accounting for the interaction effects.
*In cella* experiments confirmed the functional nature of epistasis (*IL6* and *CHGA*), while 2 Y-chromosomal markers verified sex-specific influence on genotype-phenotype correlations. Protein-protein interaction network and pathway analyses suggested roles for the candidate genes in expression of transcripts in hypertension, uncovering additional protein partners in pathway interactions, thus implicating novel pathways. Together, our data confirmed the important roles epistasis and gene-by-sex interactions play in heritable risk of hypertension.

**86**

**Cancer Associated SNPs are Enriched for Expression Quantitative Trait Loci**

Jennifer E. Below (1), Eric Gamazon (1), Sarah M. Larson (1), Nancy J. Cox (1)
(1) The University of Chicago

Reproducible trait associations from the National Human Genome Research Institute catalog of published genome-wide association studies are known to be enriched for expression quantitative trait loci (eQTLs) (Nicolae et al. PLoS Gen, 2010).
We have found that the proportion of eQTLs (assayed in HapMap CEU lymphoblastoid cell lines using the Affymetrix exon array) among reproducibly cancer-associated single nucleotide polymorphisms (SNPs) is similar to the proportion observed for the entire set of trait-associated SNPs (32% and 39%, respectively). Further, in simulations using frequency matched SNPs, we found that SNPs associated with all cancer phenotypes are enriched for cis-acting ($p$-value 0.00017), yet we found no enrichment in this set for trans-eQTLs. This proportion of eQTLs increases dramatically from 32% for all cancers to 47% when we restrict our analysis to cancers documented in the catalog that aggregate within families: chronic lymphocytic leukemia, breast cancer, ovarian cancer, and prostate cancer. 13% of these familial cancer-associated SNPs regulate 5 or more transcripts.
These observations suggest that, even more so than for complex traits in general, genetic risk factors for familial cancers will often affect phenotype through regulation of transcription.

**87**

**Genetics of X Chromosome Gene Expression in Human Monocytes—The Gutenberg Heart Study**

Raphaele Castagne (1), Tanja Zeller (2), Maxime Rotival (1), Silke Szymczak (3), Vinh Truong (1), Andreas Ziegler (3), Francois Cambien (1), Stefan Blankenberg (2), Laurence Tiret (1)
(1) INSERM-UMRS937, Pierre et Marie Curie University
(2) Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universitat Mainz
(3) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck

X-linked genes are tightly regulated by mechanisms of inactivation and dosage compensation to ensure balanced expression between males and females and between the X and the autosomes. The proportion of genes escaping these processes, as well as the influence of genetic variability on X-linked gene expression has not yet been explored in large-scale studies. We analyzed the expression profile of circulating monocytes (26,505 probes) in a sample of 1,467 subjects (49% of females). The proportion of probes expressed in monocytes was lower on the X than on autosomes (30.7% vs 42.6%). Among expressed transcripts, the median expression level was comparable on the X and on autosomes (7.94 vs 8.01, $p = 0.12$) indicating that the mechanism of dosage compensation of X-linked genes was globally achieved. There was a significant excess of female-biased transcripts on the X compared to autosomal transcripts (14.4% vs 5.5%, $P < 10^{-5}$) supporting the hypothesis that a fraction of X-linked genes escape inactivation in females. To investigate *cis* eQTLs, we analyzed 352,972 SNPs located at a distance $< 250$ kb of any probe. The proportion of *cis* eQTLs associated with an $R^2 > 5\%$ was lower for X-linked than for autosomal probes (13.4% vs 18.4% in males, 14.3% vs 18.7% in females) and among *cis* eQTLs, the median $R^2$ was also lower (8% vs 10%), suggesting that X-linked genes may be more tightly regulated than autosomal genes due to their fundamental role in development and differentiation.

**88**

**A Model to Estimate Alelic Imbalance using RNA-seq Data**

John P. Ferguson (1), Judy H. Cho (1), Dean Palejev (1), Hongyu Zhao (1)
(1) Yale University

Allelic imbalance refers to a situation where one allele from a pair of heterozygous genes is expressed at a higher level than the other. This phenomenon is observed in up to 20% of human genes [1] and is supsected to be implicated in the pathogenesis of many diseases [2]. Here we develop a novel Bayesian model that estimates allelic imbalance on an exon by exon basis from RNA-seq data. Unlike previous approaches, which have used a ''single snp'' approach to survey allelic imbalance ([2,3]), this model allows one to jointly use the information in several heterozygous snps within an exon, even when the phasing is unknown. This

feature significantly increases the power compared with a "snp by snp" analysis. The application of the model is demonstrated on both real and simulated RNA-seq data.

References:
[1] Serre D, et al. 2008. Differential Allelic Expression in the Human Genome: A Robust Approach To Identify Genetic and Epigenetic *Cis-Acting* Mechanisms Regulating Gene Expression. PLoS Genet e1000006.
[2] Heap GA, et al. 2010. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet 19:122–134.
[3] Degner JF, et al. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25:3207–3212.

**89**

**Deriving Individual Genotypic Barcodes from Gene Expression Data**

Ke Hao (1), Eric E. Schadt (1)
(1) Research Genetics, Rosetta Inpharmatics

**Backgrounds**. RNA profiling studies simultaneously capture the expression pattern of many genes. The expression levels are often under genetic control, characterized as expressional quantitative trait loci (eQTL). Large amount of mRNA expression data has been deposited into public databases, which may expose the identity of participants. **Methods**. Employing published eQTLs as the prior, we developed a Bayesian approach to predict genotype based on mRNA expression data. The predicted genotype on many eSNPs can reveal individual's identity. **Results**. Strong eQTLs are highly consistent across studies and tissue types (e.g. liver and adipose). Such high consistency allows us to use published eQTLs and predict eSNP genotypes of independent mRNA expression datasets. On empirical data, our method was shown to offer very high accuracy. In within-tissue scenario (eQTLs and testing data are on the same tissue type), we resolved 98% testing subjects' identity at p?1E-5 confidence level. In cross tissue scenario, we resolved 92% subjects' identify at p?1E-5. **Conclusions.** Individual's identity can be accurately resolved using RNA profiling data. We discuss a number of implications. For example, this novel method provides quality control against DNA-RNA pair mis-annotation in eQTL studies. And more importantly, mRNA expression datasets of public domain do not mask participant's identity.

**90**

**Genome-wide eQTL Interaction Analysis with INTERSNP**

Christine Herold (1,2), Eric R. Gamazon (1), Tim Becker (2,3), Nancy J. Cox (1)
(1) Department of Medicine, University of Chicago, Chicago, Illinois, United States of America
(2) Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany
(3) German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Interaction between genetic variants may partly explain the "missing heritability". Studies with SCAN (SNP and Copy number ANnotation) have revealed that SNPs associated with complex traits are more likely to be eQTLs than other SNPs chosen from high-density platforms. An understanding of epistasis at the level of the transcriptome should provide a powerful approach to determining whether epistasis plays a role in human complex traits. We present an approach to G × G interaction among eQTLs with INTERSNP. In order to overcome the computational constraints, SNPs are selected for joint analysis using a priori information. We tested only pairs of eQTLs, *cis* or *trans*, that show at least modest marginal association with transcripts. Additionally, we use pathway information to reduce the number of tests.
As an application, we considered the Hapmap eQTLs predicting the transcript level of *PTGER4* (reproducibly associated with Crohn's disease). Restricting to eQTLs located on the same chromosome of *PTGER4*, the best interaction result, without the use of a priori information, is the eQTL pair rs4869590 (a *cis* eQTL for *PTGER4*) and rs40172 (further downstream), showing a significant interaction after Bonferroni correction. The genome-wide interaction analysis also shows significant (Bonferroni-adjusted) results, including the eQTL pair rs2781575 and rs2328144. This study demonstrates the use of eQTL interaction analysis in elucidating the genetic architecture of complex traits.

**91**

**Adverse Effects of Nucleotide Excision Repair and Transcription Gene Abnormalities on Human Fetal and Placental Development**

Roxana Moslehi (1), A. Kumar (1), C. Signore (2), J.L. Mills (2), A. Dzutsev (3)
(1) School of Public Health, State University of New York (SUNY) at Albany, Rensselaer, NY
(2) National Institute of Child Health and Human Development (NICHD), NIH, Bethesda, MD
(3) National Cancer Institute (NCI), NIH, Frederick, MD

Our recent genetic epidemiologic investigation of gestational outcomes associated with abnormalities in trichothiodystrophy nucleotide excision repair (NER) and transcription genes, namely XPD(ERCC2), XPB(ERCC3), TTD-A(GFT2H5), and TTDN1(C7ORF11), revealed significantly increased risk of several severe gestational complications including preeclampsia in affected pregnancies where the fetus had two mutations, but not in unaffected pregnancies where the fetus was either heterozygote or had no mutations. To test our hypothesis that DNA repair/transcription genes are involved in normal placental development and decipher biologic mechanisms, we analyzed gene expression arrays of normal human tissues including placentas, placentas from pregnancies with preeclampsia and fibroblasts of patients with TTD and other DNA repair disorders. Comparison of gene signatures between normal and TTD-affected tissues revealed several relevant XPD-dependent pathways including progesterone metabolism and oxidative stress response pathways. The largest group of downregulated genes in preeclampsia belonged to RNA PolymeraseII-mediated pathways including subunits of transcription factorII(TFIIH) complex such as XPD, XPB and TTD-A; the key regulator of gene expression was GTF2E1(a component of TFIIA which modulates TFIIH). Our results indicate an important role for TTD NER/transcription gene products during normal human placental development and highlight the relevance of the exact genetic abnormality.

**92**

**Adaptive Linear Rank Tests for SNP-gene Expression Association Studies**

Silke Szymczak (1), Tanja Zeller (2), Maxime Rotival (3), Arne Schillert (1), Laurence Tiret (3), Stefan Blankenberg (2), Andreas Ziegler (1)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, Universitatsklinikum Schleswig-Holstein, Campus Lubeck
(2) Medizinische Klinik, Universitatsmedizin Mainz, Universitat Mainz
(3) INSERM UMRS937, Pierre and Marie Curie University and Medical School, Paris, France

Single nucleotide polymorphism-gene expression association studies (SNP-eQTL analysis) are performed to clarify the molecular function of genetic variants identified in genome-wide association (GWA) studies. The aim of these studies is to detect differences in location parameters of gene expressions given genotype. Expression data are often highly skewed or heavy-tailed. As a result, standard approaches such as the analysis of variance (ANOVA) or the Kruskal-Wallis test can have empirical type I error levels of up to 99% when the nominal type I error level is 5% (Szymczak et al. 2009 Stat Med 28:3581–3596). The median test is a valid alternative for comparison of medians if distributions differ in shape. However, it is less powerful than standard approaches in situations with symmetric distributions. A promising alternative are adaptive methods that estimate skewness and tail weight in the first step. These estimates are then used to select the most appropriate linear rank test statistic in the second step. We propose a novel adaptive test for the analysis of SNP-eQTL associations and demonstrate its validity in Monte-Carlo simulation studies. CADomics, a GWA study of coronary artery disease ($n = 5030$) in combination with global monocyte transcriptome data is used to compare identifed SNP-eQTL associations in two subgroups. We show that results of our new methods are more concordant than associations identified by standard approaches.

**93**

**Statistical Issues in Mapping QTLs for RNA-seq Data**

Wei Zheng (1), Debasish Raha (1), Michael Snyder (2), Hongyu Zhao (1)
(1) Yale University
(2) Stanford University

High throughput sequencing technology provides us unprecedented opportunities to study genotype-phenotype relationship. For example, we can map QTLs for gene expression variations (eQTL mapping) using mRNA sequencing in individuals genotyped at a dense set of markers. Compared to traditional eQTL mapping using microarray data, RNA-Seq data have many advantages, such as high resolution, low background, and ability to identify novel transcripts. Moreover, for genes with multiple isoforms, expression of each isoform can be estimated from RNA-Seq data, making it possible to map QTL not only for gene expression levels, but also for differential splicing. However, due to the unique features of this new data type, there are some statistical issues we need to take into consideration before we can make biologically meaningful conclusions. For example, read-mapping biases caused by SNP variations may confound with allelic imbalance in gene expression (Bioinformatics 2009, 25:3207–3212). Commonly used estimates of gene expression levels from RNA-Seq data (e.g. RPKM) are biased in terms of gene length, and potentially GC content and dinucleotide frequency, which may also affect QTL mapping. In this presentation, we discuss methods to incorporate different biases in RNA-Seq data processing using simulated and real data, and study the impact of different processing procedures on QTL mapping.

**94**

**Association Testing of Maternal CpG-Site-Specific Methylation and Congenital Heart Defects**

Stephen W. Erickson (1), Shimul Chowdhury (1), Mario A. Cleves (1), Stewart L. MacLeod (1), Weizhi Zhao (1), Ping Hu (1), Charlotte A. Hobbs (1)
(1) University of Arkansas for Medical Sciences

Congenital heart defects (CHDs) are the most common structural birth defects. The etiology of most CHDs is unknown but is thought to result from an interaction of multiple genetic, epigenetic, and environmental factors. Altered folate metabolism has been associated with CHDs and also with altered DNA methylation patterns. Evidence linking genetic and metabolic alterations in folate metabolism and CHDs exists, but the relationship between maternal DNA methylation and CHDs remains relatively unexplored. In a study built on the National Birth Defects Prevention Study (NBDPS), blood was collected from Arkansas NBDPS participants who delivered a singleton live birth with a non-syndromic CHD, and from Arkansas NBDPS mothers who had a live birth without a major defect. In 180 cases and 187 controls, proportion of methylation was interrogated at over 27,000 CpG sites in over 14,000 genes using the Infinium HumanMethylation27 BeadChip. Lifestyle information was available for all subjects to perform covariate adjustments. We have found associations between site-specific methylation and CHD in multiple genes, including GDF3, EGFR, MAP4K5, and UGDH, which have previously been identified as critical in heart development. Here, we focus on statistical methodology issues, including transformation of methylation ratios, model building, variable selection, adjustment for multiple testing, and distinguishing true associations from potentially spurious ones due to between-chip variation.

**95**

**Association between Parental Age and DNA Methylation Patterns in Newborn Umbilical Cord Blood**

Julia Krushkal (1), Frances A. Tylavsky (1), Ronald M. Adkins (2)
(1) Department of Preventive Medicine, the University of Tennessee Health Science Center
(2) Department of Pediatrics and Children's Foundation Research Center of Memphis, the University of Tennessee Health Science Center

In recent years, industrialized nations have witnessed a dramatic increase in the average age of women giving birth. While epidemiological studies show an increased incidence of pregnancy complications, chromosomal abnormalities, and certain disorders such as cancer and neurocognitive diseases in the offspring of older parents, the molecular mechanisms of this increase are yet to be

fully understood. Decreases in the DNA methylation levels in adult tissues with age have been well documented. However, little is known about the influence of parental age on DNA methylation in the offspring. We assayed DNA methylation patterns in umbilical cord blood at 27,578 CpG sites genome-wide in 168 newborns and related them to numerous parental and newborn characteristics. A generally negative genome-wide correlation was observed between parental age and DNA methylation of autosomal probes in the newborn. Methylation of 144 CpG probes in 142 genes was significantly correlated with maternal age at the genome-wide level. A weaker correlation with paternal age was also present. Many hypomethylated genes are involved in cancer-related processes, which may be predisposing the newborn to increased cancer risk. Additional genes are involved in mesodermal development, neurological regulation, glucose/carbohydrate metabolism, nucleocytoplasmic transport, and transcriptional regulation. We believe this is the first demonstration of an effect of parental age on DNA methylation in the next generation.

**96**

**Addressing Genomic Imprinting in a Family-based Genome-wide Association Study**

Andre Scherag (1), Ivonne Jarick (2), Carolin Putter (1), Anke Hinney (3), Bernhard Horsthemke (4), Karl-Heinz Jockel (1), Johannes Hebebrand (3)
(1) Institute for Medical Informatics, Biometry and Epidemiology; Essen; Germany
(2) Institute of Medical Biometry and Epidemiology; Marburg; Germany
(3) Department of Child and Adolescent Psychiatry and Psychotherapy; Essen; Germany
(4) Institute of Human Genetics; Essen; Germany

The results of genome-wide association studies (GWAS) have had a dramatic impact on our understanding of the inheritance patterns of common variants for common complex disorders. However, most of the statistical models applied in the GWAS analyses were relatively simple offering the option for meta-analyses which are currently performed at a global scale for various phenotypes. In parallele there is a tendency to re-assess the given GWAS data sets using more complex statistical models. We demonstrate the usefulness of Bayesian methods for a genome-wide imprinting analysis of 424 German nuclear families with extremely obese offspring based on Affymetrix Genome-Wide Human SNP Arrary 6.0. Genomic imprinting is an epigenetic process in which the copy of a gene inherited from one parent is expressed at a significantly lower level than the copy from the other parent. We will discuss challenges and implications of our proceeding and demonstrate the impact of the priori choice on the observed results. Finally, we will discussion discrepancies between the Frequentist and the Bayesian analyses.

**97**

**Power and FDR Control of Univariate Tests for DNA Methylation Microrarrays**

Timothy J. Triche (1), Peter W. Laird (2), Kimberly D. Siegmund (1)
(1) Department of Preventive Medicine, Keck School of Medicine, University of Southern California
(2) USC Epigenome Center, Norris Cancer Center, Keck School of Medicine, University of Southern California

DNA methylation microarrays present a high-throughput, low-cost option for epigenetic profiling. On the popular Illumina platform, DNA methylation is measured at each locus using a beta value, the proportion of total signal intensity due to methylated probe intensity. In a collection of independent samples, this quantity often results in an asymmetric, unimodal distribution of values whose variance is a function of the mean. In a simulation study, we compare the performance of the Mann-Whitney, Welch, and 2-sample *T* tests across a variety of balanced and unbalanced sample sizes. Data for over 22,000 loci are simulated using beta distributions, with estimates of the mean and variance from 48 tissue samples of colorectal cancer and adjacent normal tissue. *p*-values are adjusted for multiple testing using the Benjamini-Hochberg approach and a 5% false discovery rate (FDR). We find that the Mann-Whitney test achieves the nominal FDR for sample sizes as low as 30 while attaining the highest power. The Welch *T*-test is the least suited to unequal sample sizes of moderate number, resulting in a false discovery rate far in excess of the nominal value. The inflated FDR is more severe for the skewed beta distributions observed in DNA methylation assays than for the same means and variances simulated under Gaussian distributions. Future work will evaluate the effectiveness of analyses using a generalized linear model with binomial variance and beta regression.

**98**

**Evaluation of Public Control Data and Case-control Ratios for Genetic Association Studies**

Indra Adrianto (1), Christopher J. Lessard (2), Adam Adler (1), Kenneth M. Kaufman (3), Kathy L. Moser (2), Courtney Gray-McGuire (1)
(1) Oklahoma Medical Research Foundation
(2) Oklahoma Medical Research Foundation; University of Oklahoma Health Sciences Center
(3) Oklahoma Medical Research Foundation; University of Oklahoma Health Sciences Center; Oklahoma City VA Medical Center

In genetic association studies, using public control data has become a common practice to improve statistical power and reduce costs. Previous studies have suggested matching up to 5 controls per case. In order to assess public control data and case-control ratios, we genotyped 212 cases with an autoimmune disease and 62 controls of European ancestry, then merged them with the Illumina iControl Hap550v1 and Hap550v3 (3,172 controls) data. We also removed 1,729 iControl samples $<30$ years old to match our study data. There are 320,371 common SNPs in the merged data. We matched from 1 control up to 7 controls per case based on identity-by-state allele sharing. Association analysis was performed using logistic regression adjusted for sex and the first 4 principal components. The results show association signals start to exceed the genome-wide significance threshold ($P<5\times10^{-8}$) when the case-control ratio is 1:5. The strongest association signals were achieved when the case-control ratio is 1:6 with 11 SNPs exceeding the threshold. These SNPs are located in the major histocompatibility complex (*MHC*) region on chromosome 6. When we used all available data without any adjustment, we found $>80$ SNPs in the *MHC* region exceeding the threshold. However, these

strong signals are artificially inflated due to the Northern and Southern European substructure in the data. The results suggest that determining the optimal choice of controls is necessary in genetic association studies.

**99**

**Association Between Mammographic Breast Density and a Chromosome 7p Marker is Confounded by Allele Frequency Variation Across Populations**

C.M.T. Greenwood (1), A.D. Paterson (1), L. Linton (2), I.L. Andrulis (3), C. Apicella (4), A. Dimitromanolakis (1), J.L. Hopper (4), E.M. John (5), V. Kriukov (2), L.J. Martin (2), R. Parekh (1), A. Salleh (2), E. Samiltchuk (1), E. Satariano (6), M. Southey (4), J.M. Rommens (1), N.F. Boyd (2)
(1) The Hospital for Sick Children, Toronto, ON
(2) Campbell Family Institute for Breast Cancer Research, Ontario Cancer Institute, Toronto, ON
(3) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON
(4) The University of Melbourne, Melbourne, Australia
(5) Cancer Prevention Institute of California, Fremont, CA; Stanford Cancer Center, Stanford, CA.
(6) Cancer Prevention Institute of California, Fremont, CA

Mammographic density (MD), determined by mammography, is known to be highly heritable and a strong risk factor for breast cancer. In the context of a genome-wide linkage study, epidemiological and mammographic data were assembled for 1415 small families from several sources. Our primary phenotype was formed by fitting a linear model to the square root of percent MD, adjusting for age at mammogram, number of live births, menopausal status, weight, height, weight squared and hormone replacement therapy. We tested for within-family association using QTDT at 5680 SNPs from the Illumina Infinium II Human Linkage-12 panel.
The strongest signal from the orthogonal test was at rs723149 on chromosome 7p (? = 0.265 (allele A), s.e. 0.065, $p = 5e{-}5$). Analysis of this marker in a replication set of 1105 unrelated women from Ontario, showed no evidence of association with MD ($p = 0.88$). However, the allele frequency varies substantially, so that allele A is rare in African populations but the major allele in some other populations. For a subset of 316 women in the replication set with very high or low MD, it was possible to infer population substructure from additional genotyping. A consistent trend for association (? = 0.41, s.e. = 0.2712, $p = 0.13$) was seen among 256 of these women who were inferred to form a single population cluster. This marker may be associated with breast density, but interpretation is confounded by ethnicity.

**100**

**WITHDRAWN**

**101**

**Effect of Confounding and Heterogeneity on GWAS among Admixed Populations**

Jinghua Liu (1), Frank Gilliland (1), William J. Gauderman (1), David V. Conti (1)
(1) University of Southern California

There is increasing demand for genome-wide association studies to be performed on diverse populations, such as Hispanics. However, since Hispanics consist of admixed individuals, these studies lead to challenges such as confounding due to population stratification and heterogeneity due to differential LD between parental ancestries. Here, we investigate three models to test the association in admixed individuals. The first model tests the marginal effect of a SNP; the second model tests the marginal effect of individual local ancestry akin to admixture mapping; the last model contains both the main effect of a SNP, local ancestry, and the interaction between them. This model has the potential to capture effects due to SNP variation, effects due to differential local ancestry, and heterogeneity in SNP effects due to differential LD. Through simulations, we investigate the performance of these models under various scenarios of LD and allele frequencies. In this context, we also investigate the efficiency of adjusting for individual global ancestry to control for confounding due to population stratification. Finally, we apply these models to the University of Southern California's Children's Health Study. We estimate individual global and local ancestries through STRUCTURE and HAPMIX and perform genome-wide scans of all three models. We discuss areas of interest and how they differ in the genetic effect, local ancestry, and the effect of heterogeneity due to differential LD.

**102**

**Theoretical Formulation of Principal Components Analysis: Application in Inferring Admixture Proportions**

Jianzhong Ma (1), Christopher I. Amos (1)
(1) Department of Epidemiology, U.T. M.D. Anderson Cancer Center

Principal components analysis (PCA) is now widely used in genetic epidemiologic studies for detecting and correcting for population stratification. To provide a theoretical guidance for interpreting the pattern of principal component (PC) scatter plot, we have established a theoretical formulation of PCA by treating individual samples from different populations as features and genetic markers as samples. Here, we extend our theory to the case of population admixture. A reduced eigen-equation in terms of the variance-covariance parameters of the random vector of allele frequencies is derived and can be solved numerically to yield eigenvectors that are the axes of variation required for differentiating the populations. We theoretically show that there exists an asymptotically stable pattern of PC scatter plot as the sample size increases. The asymptotic form of the eigen-equation can be used to explain the special PC scatter patter when admixed individuals are included in PCA. Based on our theory, we propose a couple of estimators for the admixture proportions of individuals with an arbitrary number of parental populations. We also show that population stratification can be corrected for by removing the top PCs that differentiate parental populations for markers of which the allele frequencies of the admixed population can be derived from those of the parental populations using the admixture proportions.

**103**

**Tracing the Population Origin of Non-European Chromosomal Segments Identified by the Method of Rare Heterozygotes and Homozygotes (RHH) in Admixed Subjects of European Caucasian Descent**

Ralph McGinnis (1), William McLaren (2)
(1) Wellcome Trust Sanger Institute
(2) European Bioinformatics Institute

The method of rare heterozygotes and homozygotes (RHH) can visualize genome mosaicism in admixed European Caucasians with African or Asian ancestry (McGinnis et al, Human Molecular Genetics, in press). This raises the methodological issue of determining the exact non-European population from which a visualized chromosomal segment of admixture arises. This determination is complicated since typically only one of the subject's two extended haplotypes spanning the segment derives from the non-European population. To identify segment origin in admixed Caucasians of the 1958 British Birth Cohort (58BBC) genotyped on the Illumina 550K (Illm550K) chip, we pursued several strategies. The most promising has been to derive the "European" haplotype spanning a subject's admixed segment by "zeroing" the subject's heterozygous Illm550 K SNPs in the segment and imputing the zeroed genotypes using (a) the subject's homozygous Illm550 K SNPs plus (b) genotypes at all Illm550 K SNPs in the segment from ~1500 unadmixed 58BBC subjects. The resulting combination of imputed and called homozygous genotypes define the European hapotype which in turn yields the non-European haplotype by subtracting the imputed allele from each zeroed heterozygote. By comparing the non-European haplotype against haplotypes or genotypes of 51 populations in the Human Genome Diversity Panel typed by Illm550 K, we have achieved partial success in identifying population of origin and will illustrate our results.

**104**

**Effective Population Sizes of Current Human Populations**

Leeyoung Park (1)
(1) Yonsei University

In order to accurately estimate the effective population size ($N_e$) of the current human population, two new approaches, which were modified from the previous methods, were employed in this study. One is based on the linkage disequilibrium (LD) from completely unlinked loci between different chromosomes, and another is based on the deviation from the Hardy-Weinberg Equilibrium (HWE). When random mating is assumed, genetic drifts in population naturally induce the linkage disequilibrium between chromosomes and the deviation from HWE. The latter provides information on the $N_e$ of current population, and the former provides the same when the $N_e$ is constant. If $N_e$ fluctuates, the $N_e$ of the previous generation can be derived by applying the $N_e$ estimates from HWE to the estimates from LD between chromosomes. Using HapMap Phase III data, the estimates were varied from 358 to 10,437, depending on populations and time. The $N_e$ appeared to fluctuate as it provided different estimates from each of the two methods. The $N_e$ estimates of current and previous generations were found to agree well with the overall increment observed in recent human populations.

**105**

**Population-based Association Models Adjusting for Family and Population Structure**

Gina M. Peloso (1), Josee Dupuis (1), Kathryn L. Lunetta (1)
(1) Department of Biostatistics, Boston University School of Public Health

When a sample is composed of families, population structure (PS) in genetic association analyses can be addressed by performing family-based tests that condition on parental genotypes or their sufficient statistics. These statistics are immune to biases due to PS, but are known to have low power, particularly for unselected samples. Alternatively, several linear mixed effects (LME) models and score tests have been proposed to evaluate the association between a genetic marker and outcome controlling for both population and family structure. Some of the models correct for family structure in the variance and adjust for estimates of ancestry as fixed effects, while others correct for both family and population structure in the variance using estimated kinship or identity by state matrices from genome-wide genotype data. We compare the type I error, power, and computation time of several models that account for family structure with continuous outcomes using a range of family sizes and population structures, including 2 and 3 generation families with admixed and discrete PS. We find that for discrete PS, LME models accounting for family structure as a random effect and population structure as a random or fixed effect appear to have the best power, while maintaining appropriate Type I error. However, LME accounting for both family and population structure as random effects requires considerably longer computing time than other methods.

**106**

**Interrogating Local Population Structure for Fine Mapping in Genome-wide Association Studies**

Huaizhen Qin (1), Nathan Morris (1), Sunjung Kang (1), Mingyao Li (2), Bamidele Tayo (3), Helen Lyon (4), Joel Hirschhorn (5), Richard S. Cooper (3), Xiaofeng Zhu (1)
(1) Department of Biostatistics and Epidemiology, Case Western Reserve University
(2) Department of Biostatistics & Epidemiology, University of Pennsylvania School of Medicine
(3) Department of Preventive Medicine and Epidemiology, Loyola University
(4) Department of Genetics, Harvard Medical School
(5) Department of Genetics, Harvard Medical School; Programs in Medical and Population Genetics, Broad Institute of Harvard and MIT

Adjustment for population structure is necessary to avoid bias in genetic association studies of susceptibility variants for complex diseases. Population structure may differ from one genomic region to another due to the variability of individual ancestry associated with migration, random genetic drift or natural selection. Current association methods for correcting population stratification usually involve adjustment of global ancestry between study subjects. We suggest interrogating local population structure for fine mapping to more accurately locate true casual genes by adjusting local ancestry. By extensive simulations on genome-wide data sets, we show that adjusting global ancestry may lead to false positives when local population structure is an important confounding factor. In contrast, adjusting local ancestry can effectively prevent false positives due to local population structure and thus can improve fine mapping for disease gene localization. We applied the local and global adjustments to the analysis of data sets from three genome-wide association studies, including European Americans, African Americans, and Nigerians. Both European Americans and African Americans demonstrate greater variability in local ancestry than Nigerians. Adjusting local ancestry successfully

eliminated the known spurious association between SNPs in the LCT gene and height due to Northern and Southern European population structure in European Americans.

**107**

**Haplotype Structure Analysis of Isolated and Stratified Populations via the Standardized Maximum Distance Measure for Haplotype Variety**

Cyril Rakovski (1), Michelle Creek (1)
(1) Department of Mathematics and Computer Science, Schmid College Science, Chapman University, Orange, CA 92866, USA

We undertake a study to assess the level of haplotype variety in isolated and stratified populations and implement a model for detecting and quantifying the differences in haplotype structures of these populations. We propose a novel measure for haplotype variety—the Standardized Maximum Distance statistic (SMD) that determines the degree of departure from independent transmissions of the alleles at given loci through a standardized version of the maximum of the absolute values of the differences of all haplotype frequencies and the product of the corresponding allele frequencies. For the analysis of isolated populations we use the haplotype data from release III of the HapMap database. However, the SMD statistic is biased in small sample data scenarios such as the HapMap. Thus, we implement a nested simulation study to estimate and remove such bias; we also contrast the performance of our method to a novel use of a classical MCMC-based contingency table goodness-of-fit test. Lastly, we compare the haplotype structures of these populations through a generalized linear model. Our results show that all population groups have significantly different adjusted average SMD values. Our findings demonstrate the existence of significant differences in haplotype variety between human populations and corroborate previous research based on alternative measures of haplotype variety.

**108**

**Stratification-Score-Based Matching Outperforms Other Matching Approaches when Controlling for Confounding**

Glen A. Satten (1), Michael P. Epstein (2), Richard Duncan (2), Alaine Broadaway (2), Andrew S. Allen (3)
(1) CDC
(2) Emory University
(3) Duke University

Proper control of confounding due to population stratification is crucial for analysis of genetic association studies. Control of confounding by fine matching of case to control participants based on genetic ancestry is increasingly popular. Genetic ancestry can typically be summarized by a small number of genomic variables (e.g., principal components) that are significant predictors of genomic variability. However, it is hard to match study participants on many continuous variables. Existing approaches thus form matches using a scalar measure that combines the significant components. However, use of ancestry components that do not predict disease can lead to inaccurate matches, and hence to improper control of confounding. Here we propose matching on the stratification score, which is the probability of disease given genomic variables [AJHG 80:921–930]. When matching on the stratification score, case participants are matched to control participants who have a similar risk of disease, given genetic ancestry information. Using a genome-wide case-control study of schizophrenia among African-Americans, we show confounding is resolved by our approach but not by other matching procedures (GEM, spectral-GEM and GSM). We also use simulated data to compare approaches. [GEM = GEnetic Matching, AJHG 82:453–463; spectral-GEM, Genet Epidemiol 34:51–59; GSM = Genetic Similarity score Matching, Genet Epidemiol 33:508–517].

**109**

**Factor Analysis of Population Structure and Admixture**

Daniel Shriner (1)
(1) National Human Genome Research Institute

Principal components analysis of genetic data is used to avoid inflation in type I error rates in association testing due to population stratification by covariate adjustment using the top eigenvectors and to estimate group membership independent of self-reported or ethnic identities. Factor analysis was developed to identify which principal components should be retained. I compare an established technique from factor analysis, Velicer's minimum average partial test, to the most widely used implementation of principal components analysis in genome-wide association analysis, EIGENSOFT. By computer simulation based on coalescent theory, EIGENSOFT is shown to systematically overestimate the number of significant principal components. Furthermore, this overestimation is larger for samples of admixed individuals than for samples of unadmixed individuals. Overestimating the number of significant principal components can potentially lead to a loss of power in association testing by adjusting for unnecessary covariates and may lead to incorrect inferences about group differentiation. Velicer's minimum average partial test is shown to have both smaller bias and smaller variance, often with a mean squared error of zero, in estimating the number of significant principal components. Velicer's minimum average partial test is implemented in a computationally efficient algorithm named FACTORSTRAT and is suitable for genome-wide genotype data with or without population labels.

**110**

**Alternative Approach to Control for Population Structure in Analysis of Genome-wide Association Data**

Fengyu Zhang (1)
(1) National Instituites of Health

Population stratification has been one of major conerns in analysis of genome-wide association data using case-control design. Because it is difficult to match the genetic backgound between cases and controls at the stage of study design, one has to turn to regression approach to control for population structure while assessing the SNP-genotype association. People often use principle component analysis (PCA) to identify several principal components or PCA-correlated SNPs as covariates to control for population structure. However, this approach may face several issues such as SNP linkage disequilibrium or SNP selections. We developed an mutlivariate

approach to perform genomic matching. We demonstated that this approach is not sensistive to SNP LD structure as the PCA-based approach, and allow us to undertake fine mapping of association such as haploytpe analysis while controling for potential population structure.

**111**
**Haplotype-mining in GWAS: Application to Data from the Genetic Epidemiology of CLL consortium**
Ryan Abo (1), Nicola J. Camp (1), Genetic Epidemiology of CLL (2)
(1) University of Utah
(2) Consortium

Missing heritability in common diseases may be due to rare risk variants. Common SNPs inadequately tag rare variation, and therefore single SNP analyses in genomewide association studies (GWAS) have low power to detect rare risk variants. Haplotype analyses hold potential for greater power to detect these rare risk variants. Here we present a two-stage strategy to perform haplotype analyses in GWAS data. In the first step, genomic regions with significant clustering of only nominal single marker *p*-values are identified (LHiSA, Guedj et al., 2006). For the second step, we apply a haplotype-mining method (hapConstructor, Abo et al., 2008) in these specific regions. We applied our two-stage strategy to GWAS data for familial chronic lymphocytic leukemia (CLL) from the Genetic Epidemiology of CLL (GEC) consortium. The resource contained 102 familial CLL cases and 296 controls genotyped on 830,255 SNPs across the genome. We identified 59 regions that exhibited significant clustering of significance values and had at least 8 contiguous SNPs with only nominal *p*-values ($0.05 < P < 0.005$). We performed haplotype-mining in these regions and identified six regions with significant associated haplotypes ($P < 1.0 \times 10^{-3}$). These associated haplotypes attained significances up to three orders of magnitude more significant than that of the single SNP results. These findings are hypothesis-generating due to the data-mining approach and replication will be required.

**112**
**Overlapping Haplotype Association Analysis via Penalized Logistic Regression**
Kristin L. Ayers (1), Heather J. Cordell (1)
(1) Institute of Human Genetics, Newcastle University

It has been suggested that testing for association between conserved haplotypes and a phenotype may capture untyped causal variants in weak linkage disequilibrium with nearby typed markers. We propose a sliding window approach which uses inferred overlapping haplotypes as variables in penalized logistic regression. We investigate a penalty with four separate components: (1) a standard lasso penalty for shrinking the size of coefficients while performing model selection, (2) a group lasso that encourages all haplotypes in a window to be included or excluded from the model, (3) an allele sharing penalty that encourages haplotypes with similar alleles to have similar coefficients, and (4) a variation of the fused lasso which encourages overlapping haplotypes with shared alleles to have similar coefficients. The penalized likelihood is maximized with a cyclic coordinate ascent algorithm, allowing quick coefficient estimation for a large number of markers. We compare our method to single marker analysis and competing haplotype methods on a variety of simulated case control data sets and a real data set.

**113**
**A Common APOBEC3H Haplotype Associates with Decreased HIV-1 Sequence Editing and Lower HIV-1 RNA Set-Point In Early Untreated HIV-1 Infection**
Jason D. Barbour (1), Pierre-Antoine Gourraud (1), Jon Woo (1), Ali Karaouni (1), Timothy Schmidt (1), Gerald Spotts (1), Jorge R. Oksenberg (1), Frederick M. Hecht (1), Teri J. Liegler (1)
(1) University of California San Francisco

**Introduction:** We mapped genetic variation of the human APOBEC3 (A3) locus in a cohort of treatment naive, recently HIV-1 infected adults to measures of sequence editing activity in the integrated HIV-1 DNA genome, and HIV-1 clinical markers.
**Methods:** We examined single nucleotide polymorphisms (SNP) in the APOBEC3 locus on chromosome 22, paired to population sequences of pro-viral HIV-1 *vif* of peripheral blood mononuclear cells (PBMC), from 96 recently HIV-1 infected treatment naive adults.
**Results:** We found evidence for the existence of an APOBEC3H linkage disequilibrium (LD) block associated with variation in GA→AA, or APOBEC3F signature, sequence changes in pro-viral HIV-1 vif sequence. We identified a common 5 position risk haplotype telomeric to APOBEC3H (A3Hrh). These positions were in high LD ($D' = 1$; r? = 0.98) to a previously described A3H "RED" haplotype containing a variant (E121) with enhanced susceptibility to HIV-1 Vif (Zhen et al., 2009 [1]). Homozygote carriers of the A3Hrh had lower HIV-1 RNA levels over time during early, untreated HIV-1 infection, ($p = 0.018$ and 0.015 mixed effects model) and lower GA→AA (A3F) sequence editing on pro-viral HIV-1 *vif* sequence ($p = 0.01$).
**Discussion:** Our results suggest genetic variants of A3H that do not bear the E121 mutation controlling Vif susceptibility may exert a steady GA→AA sequence editing pattern upon HIV-1, contribute to viral diversification *in vivo*, and associate with elevated HIV-1 RNA levels.

**114**
**Simulation of Phased Genotype Data with Pre-specified LD Coefficients Barhdadi A12, Dube MP12. 1 Montreal Heart Institute, 2 Faculte de Medecine, Universite de Montreal**
Amina Barhdadi (1), Marie_Pierre Dube (1)
(1) Montreal Heart Institute Universite de Montreal

Simulation of phased genotypes data is often undertaken to validate novel statistical methods in genetic association studies. Simulation of SNP polymorphisms with realistic linkage disequilibrium (LD) pattern is important for the success of new statistical methods in association studies. We use a mathematical approach that generates random vectors with prescribed marginals and correlation matrix to simulate binary genetic markers with prespecified allele frequencies and LD pattern. In this approach, SNP genotypes are modeled as a multivariate random variable with known marginal distributions and pairwise correlation as measured by $r^2$. Our simulation algorithm consists

firstly of generating disease locus genotypes and then the other SNPs genotypes are simulated using a moving-window algorithm to avoid working with large proportion of LD matrices. The local LD structure of HapMap data or any other available LD data is used as input correlation matrix. LD coefficients as estimated from HapMap data are compared to those estimated from the simulated data. The accuracy of the reproducibility of the LD structure depends on the sample size, the size of the sliding window and the amount of LD. Our simulation algorithm is very simple and fast to implement compared to the existing simulating programs. We anticipate that it can be useful as a tool for comparisons and performance evaluation of statistical methods designed to detect association between genotypes and human diseases.

**115**

**Power of Different Measures of Similarity for Haplotype Sharing Analysis to Detect Rare Disease Variants in Association Tests**

Lars Beckmann (1)
(1) German Cancer Research Center

It has been hypothesized that haplotype sharing measures applied in association studies might have improved power to detect the effects of rare variants compared to single point test. In the simulation study presented here, I applied the general approach of Mantel statistics to correlate genetic and phenotype similarity. The genetic similarity between haplotypes was estimated using four different measures: (1) the number of shared intervals, (2) the physical length of the shared intervals, (3) the genetic length of the shared intervals in centimorgans, (4) the genetic length of the shared intervals in linkage disequilibrium units (LDU). The simulated data sets were based on three haplotype distributions, reflecting different linkage disequilibrium patterns. A log-additive disease model was simulated at a single binary disease locus. To increase the complexity of the models, lower and upper limits for the number of haplotypes carrying the disease allele as well as a threshold for the total relative frequency of the risk haplotypes were introduced.
In conclusion, the genetic similarity measured by LDUs showed higher power than the alternative measures, although pinpointing to a specific genomic region was hampered by dependencies between the test statistics. Haplotype sharing analysis showed higher power when the disease locus was not included into the analysis. In scenarios in which the disease locus was included, the single point chi-square was the most powerful approach.

**116**

**Development of a HapMap r27 Translation Program to Extract User-Friendly Output**

Anthony M. D'Amelio Jr (1), Carol J. Etzel (1)
(1) U.T. M.D. Anderson Cancer Center

In recent genetic studies, the International HapMap project has been a tremendous tool in finding single nucleotide polymorphisms (SNPs) that could be used in association analysis, as well as having a well-validated database of individuals with which to conduct analysis. In the newest version of HapMap (release 27), 11 different populations have full genetic data with hundreds of thousands of SNPs available. However, with this new version comes new formatting issues which makes genetic applications of HapMap data difficult. For example, the newest version of Haploview produces error statements when trying to open HapMap files. Therefore, I created a HapMap translation program (HTP) in Matlab that converts the HapMap information into more user-friendly output in 4 steps. For example, suppose a user wants to extract information on SNPs rs9405048 and rs9262152, then 1) the user inputs the base pair information for both SNPs into HapMap, and extracts the genetic information (Chr6:30778260.30788905); 2) the user inputs that extracted genotype file from HapMap, the population being studied, and the list of SNPs to be analyzed into Matlab; 3) the HTP automatically removes the initial HapMap formatting (23 initial "**words**") and extracts only the SNP alleles initially specified by the user and 4) the HTP converts the SNP data into the same numerical format that is used in Haploview. The resulting data are now ready for haplotype analysis.

**117**

**Tagsnps Genotypes and Haplotypes of NOS3 Gene and Hypertension Risk in a Brazilian Case-Control Study**

Marcelo R. Luizon (1), Valeria C. Sandrim (2), Tatiane C. Izidoro-Toledo (1), Eduardo B. Coelho (1), Jose Eduardo Tanus-Santos (1)
(1) University of Sao Paulo, FMRP/USP, Ribeirao Preto-SP, Brazil
(2) Santa Casa de Belo Horizonte, Belo Horizonte-MG, Brazil

NOS3 haplotypes formed by functional SNPs (T-786C/rs2070744 and Glu298Asp/rs1799983) and a variable number of tandem repeats in intron 4 have been associated with hypertension. However, other polymorphisms may also contribute to these findings. We evaluated the association of NOS3 tagSNPs genotypes and haplotypes with hypertension in a case-control study of hypertensives (81 whites and 53 blacks) and normotensives (83 whites and 43 blacks). TagSNPs rs3918226, rs3918188, rs743506, and rs7830 were selected at SeattleSNPs database (MAF $>0.10$) in order to capture $>60\%$ of NOS3 variability. Multifactor dimensionality reduction method was used to determine which genotypes are most predictive of risk. Two best combinations of genotypes included rs3918226 and rs3918188 (cross-validation consistency = 10 and precision $>56\%$; $P<0.01$). Haplotype-specific score statistics were estimated by Haplo.stats. The haplotype "CCbGCGA" was more common in normotensives than hypertensives, both when considered the whole sample (3% vs. zero, respectively; $p=0.05$) and only white individuals (6% vs. zero, respectively; $p=0.007$). The haplotype "TCbGAGC" was more commonly found in the whole sample of hypertensives than in the normotensives (8% vs. zero, respectively; $p=0.02$). These results suggest one eNOS haplotype associated with a protective effect against hypertension despite of the ethnic label, and one eNOS haplotype conferring susceptibility to hypertension. Financial Agency: FAPESP.

**118**

**Lessons We Have Learned from Genome-wide Haplotype Analysis (GWHA)**

Manuel Mattheisen (1), Tim Becker (2), Christine Herold (3), Stefanie Nowak (4), Jessica Becker (5), Ruth Herberz (5), Stefan Herms (5), Kerstin Ludwig (5), Heiko Reutter (6), Markus M. Nothen (7), Per Hoffmann (7), Elisabeth Mangold (4), Michael Knapp (8)
(1) Institute of Human Genetics and Institute for Medical Biometry, Informatics, and Epidemiology, University of Bonn, Bonn, Germany
(2) German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
(3) Department of Medicine, University of Chicago, Chicago, Illinois, United States of America
(4) Institute of Human Genetics, University of Bonn, Bonn, Germany
(5) Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany
(6) Institute of Human Genetics and Department of Neonatology, Children's Hospital, University of Bonn, Bonn, Germany
(7) Institute of Human Genetics and Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany
(8) Institute for Medical Biometry, Informatics, and Epidemiology, University of Bonn, Bonn, Germany

We present our experience in a full genome-wide haplotype analysis (GWHA) from a case-control study of nonsyndromic cleft lip with or without cleft palate (NSCL/P; 1,717 individuals and 521,288 SNPs[1]). In particular we focussed on two main issues: selection of strategy for a GWHA and pitfalls in computation and interpretation of results.
Different strategies and analytical tools for GWHA are available today. To name only two, one has to decide whether to include each combination of a predefined number of SNPs and distance (e.g. all combinations of 2 SNPs with an utmost distance of 100 KB between), or each haplotype derived by haplotype-block estimation (e.g. based on consideration of the CEU Phase 2 HapMap sample). After careful evaluation of different tools and strategies, we argue in favour of the first strategy. We implemented a GWHA framework in the pre-existing software INTERSNP, since there was no tool available that filled all our needs.
During the evaluation and the final GWHA for the NSCL/P sample major pitfalls were identified. For some of the tools restrictions on their usage became apparent. In particular, it turned out that especially for combinations with a greater number of SNPs, haplotype frequency estimation was not accurate in some cases. In order to allow for an efficient screening of our results we developed a framework for the evaluation and interpretation of GWHA results.

References:
[1] Mangold, Ludwig, Birnbaum, et al. 2010. Nat Genet 42:24–26.

**119**

**Power of Single-Marker versus Multi-Marker Tests of Association**

Xuefeng Wang (1), Robert C. Elston (1), Daniel J. Schaid (2), Nathan J. Morris (1)
(1) Case Western Reserve University
(2) Mayo Clinic

Recently, Kim et al. (2010) derived single-marker and multi-marker statistics to compare cases and controls according to allele frequencies, Hardy-Weinberg Disequilibrium, and composite linkage disequilibrium (LD) scores. They found that power can differ for single-marker versus multi-marker tests, depending on the LD pattern among the measured markers. We explore the power differences of the most extreme single-marker test versus a multi-marker test numerically, computing the asymptotic power for both approaches. Our strategy is to focus directly on the correlation patterns for the measured SNPs and the unmeasured causal SNP, because the power depends only on the correlation patterns—different simulation parameters could give the same correlation pattern, and hence the same power. We set the Type-I error rate to 0.05, the causal allele frequency to 0.2, the causal allele relative risk to 1.2, with 1,000 cases and 1,000 controls. For the ranges of correlations we simulated, the difference in power between the single-locus tests and the multivariate test ranged from $-0.80$ (favoring the multivariate test) to 0.10 (favoring the single-locus tests). Because the relative loss in power due to multiple degrees of freedom is much smaller when tests are performed at smaller significance levels, further results will be presented involving tests at lower significance levels, together with a comparison of multi-marker tests that include or do not include marker cross-products.

**120**

**A New Haplotype Similarity Test for Genetic Association of Quantitative Traits in GWAS Studies**

Wei W. Yang (1), Chi C. Gu (1)
(1) Washington University School of Medicine

Published genome-wide association studies (GWAS) almost exclusively rely on single-SNP tests. Many methods are available to analyze aggregate effects of clusters of SNPs in proximity. But most of them are suitable for candidate gene study and often require separate treatment of quantitative from discrete traits. We propose a novel method for haplotype similarity analysis that properly accounts for undesirable correlations in similarity measures. Specialized algorithms were developed to factorize extremely large matrices so that memory and computational burden required by the restricted maximum likelihood (REML) estimation are reduced to square of the sample size; and to perform fast estimation of haplotype similarities. These are implemented in a C program called HSim. The validity of the method was confirmed by accurately recovered model coefficients in simulation studies. Wald statistics test for non-zero coefficients also conformed to standard normal distributions under null hypothesis. The power of the test is then evaluated with simulated genotype data. Finally, HSim was applied to a real GWAS data in 9,315 individuals to analyze genome-wide association of mean arterial blood pressure. It took 4 days and 15 hours to finish 33,815 SNPs on chromosome 22 on a single CPU (full results to be updated). In summary, the new method provides an effective and practical utility for identifying important haplotype in association with binary or quantitative traits in GWAS studies.

**121**

**Accuracy and Computational Efficiency of a Graphical Modeling Approach to Linkage Disequilibrium Estimation**

Haley J. Abel (1), Alun Thomas (1)
(1) University of Utah

We develop recent work on using graphical models for linkage disequilibrium to provide efficient programs for model fitting, phasing and imputation of missing data in large data sets. Two important features contribute to the computational efficiency: the separation of the model fitting and phasing-imputation processes into different programs, and holding in memory only the data within a moving window of loci during model fitting. Optimal parameter values were chosen by cross validation to maximize the probability of correctly imputing masked genotypes. The best accuracy obtained is slightly below that had from the Beagle program, and our programs are slower for small data sets. However, for large data sets they are considerably more efficient. For a reference set of n individuals genotyped at m markers the time and storage required for fitting a graphical model are O(nm) and O(n+m) respectively. To impute the phases and missing data on n individuals using an already fitted graphical model requires O(nm) time and O(m) storage. Note that while the time for fitting and imputation are both O(nm) the imputation process is considerably faster, thus, once a model is estimated from a reference data set, the marginal cost of phasing and imputing further samples is very low. Existing methods, including Beagle, combine the estimation and imputation stages in implementations requiring time of $O(n^2m)$ and storage of O(nm).

**122**
**A Rank-based Association Test that Incorporates Uncertainty in Imputed SNPs**
Elif Fidan Acar (1), Lei Sun (1)
(1) University of Toronto

Research on imputation-based genetic association studies has mostly focused on evaluation and improvement of genotype imputation accuracy. While many imputation algorithms have been proposed, novel statistical testing strategies of the imputed genotypes remain largely unexplored. The popular dosage approach, using the expected number of copies of the risk allele to define the genotype variable, imposes a restrictive additive model on the association analyses; while the best-guess approach, identifying the genotype with the highest posterior imputation probability, fails to incorporate imputation uncertainty in the testing procedure. A robust, powerful association test statistic is developed by generalizing the Kruskal-Wallis test to account for imputation uncertainty. Using the posterior probability of each genotype group, we construct a nonparametric test statistic and derive its asymptotic null distribution. The extended statistic uses all available information and requires no assumption on the data distribution. Simulation and application studies show that the proposed test outperforms its parametric predecessors. Depending on the specific model assumptions, the gain of the power could be over 50%, while the loss of power is negligible when the simulated data are tailored for a best-guess or dosage model. With its ease of implementation, this generalized Kruskal-Wallis test could shed new light on current genetic association studies of imputed SNPs.

**123**
**Reference Samples in Imputation and its Implications in Association Analysis Results**
Mariza de Andrade (1), Martha E. Matsumoto (1), Sooraj Maharjan (1), Elizabeth J. Atkinson (1), Sharon L.R. Kardia (1)
(1) Mayo Clinic

Imputation of untyped genetic markers has been shown to increase genome coverage but there are still questions about which reference samples to use and how much the reference influences association analysis results. First we used four reference panels to evaluate the overall quality and accuracy of the imputed markers. Second, we compared association analysis results using known markers and imputed versions of the same markers. To compare the four reference panels, we used two populations from the GENOA cohort: 1385 European Ancestry (EA) sibships and 1535 African Ancestry (AA) sibships. We used MACH software for the imputation and four reference panels all pulled from the HapMap Phase II data: 1) 60 CEU samples; 2) 60 YRI samples; 3) combination of the 1st and 2nd panels; 4) combination of panel 3 with 90 JPT/CHB samples. For each study population we applied a two-step procedure specified in MACH. We evaluated the imputation by examining the quality measures and the accuracy of masked genotypes. To compare the association analysis results, we first evaluated the familial relationship implied by the genotype data and performed association analysis using the known genotypes and using the imputed genotypes of selected known markers. Our preliminary results show that panel 1 is a good reference for the GENOA EA sibships, panel 3 is a good reference for the GENOA AA sibships, and the association analysis results are not affected if a good reference panel was used.

**124**
**Predicting Highly Polymorphic Alleles Using Unphased and Flanking Single Nucleotide Polymorphisms**
Sue Li (1), Hongwei Wang (1), Anjane Smith (1), Bo Zhang (1), Gary Schoch (1), Daniel Geraghty (1), John Hansen (1), Lue Ping Zhao (1)
(1) Fred Hutchinson Cancer Research Center

Recent advances in SNP array technologies have enabled genomewide association studies of many complex traits. To facilitate interpretations and establish biological basis, it could be advantageous to infer alleles of functional genes when SNP discoveries are within or nearby genes. Leslie et al. (2008) have proposed an IBD-based method for predicting human leukocyte antigen genes with SNP data with satisfactory accuracy on the order of 97%. Building upon their success, we introduce a complementary method for predicting highly polymorphic alleles using *unphased SNP data* as training data set, since relatively few phased gene alleles and SNP alleles are gathered in typical GWAS. Due to abundance of unphased SNP data sets, the method is readily applicable to relatively large population studies, in which multiple copies of relatively uncommon alleles could be observed. Applying it to HLA genes in a cohort of 630 healthy individuals as a training set, we constructed predictive models for HLA-A, B, C, DRB1 and DQB1. Then, we performed a validation study with

another cohort of 630 individuals, and the predictive models achieve predictive accuracies for HLA alleles defined at intermediate or high resolution ranging as high as (100%, 97%) for HLA-A, (98%, 96%) for B, (98%, 98%) for C, (97%, 96%) for DRB1, and (98%, 95%) for DQB1, respectively.

**125**

**Comparison of Tagging and Imputation for HLA Allele Prediction and Association Testing**

Judong Shen (1), Stephen Leslie (2), Silviu-Alin Bacanu (1), John C. Whittaker (3), Gil McVean (2), Matthew R. Nelson (1)
(1) Genetics, R&D, GlaxoSmithKline, Research Triangle Park, NC, USA
(2) Department of Statistics, University of Oxford, Oxford Ox1 3TG, UK
(3) Genetic, R&D, GlaxoSmithKline, Harlow, UK

The major histocompatibility complex includes the genes that encode the human leukocyte antigens (HLA) involved in adaptive immune response. Making inferences about HLA associations using SNPs offers a promising alternative to HLA typing. Both tagging and imputation methods have been proposed. However, there remains considerable uncertainty regarding the reliability with which methods can infer HLA alleles from SNPs and the extent to which power to detect HLA associations is eroded when imputation or tagging is used. We compare the ability of single-SNP tagging, single-SNP imputation and multi-SNP hidden Markov model (HMM) imputation to predict high resolution alleles of *HLA-A*, *B*, *C*, *DQA1*, *DQB1* and *DRB1* in subjects of European background and the potential power loss in a case-control association testing framework. We find that multi-SNP HMM imputation generally performs better on both common and rare HLA alleles than single-SNP methods. As expected, rare alleles are more difficult to predict by all three methods, requiring standard methods of high resolution typing when such are of pivotal research interest. For some HLA alleles, lower tag SNP predictive values may be offset by relatively high HMM imputation no-call rates, suggesting that a mixed approach to analysis may be most powerful. Overall, we conclude that when sufficient training data are available, exploratory HLA analysis can be carried out using HMM imputation on SNP genotypes with little loss of power.

**126**

**Genotype Imputation African Americans**

Yan V. Sun (1), Wei Zhao (1), Sharon L.R. Kardia (1)
(1) University of Michigan

Meta-analysis of multiple genome-wide association studies using different genotyping platforms is often the only option for achieving sufficient sample sizes. The known linkage disequilibrium (LD) structure for HapMap populations and the sufficiently dense SNP genotyping platforms make imputing unmeasured SNPs based on LD structure accurate and feasible in Caucasians. However, the imputation performance in African American (AA) populations can be suboptimal because their LD profile is different from either HapMap African or Caucasian populations. To investigate the imputation performance in AAs, we used MACH to impute HapMap SNPs using 550,326 SNPs measured by Affymetrix 6.0 chip on 1,263 AAs, and compared the genotypes of 637 high-quality candidate gene SNPs measured by TaqMan assay with the imputation results. The overall concordance rate between the imputed (MACH) and the measured (TaqMan) genotypes was 89.7%. In contrast, there was 97.8% concordance between measured genotypes from Affymetrix 6.0 vs. TaqMan assay. Of the imputed SNPs 49.8% have concordance rate higher than 95%. An imputation quality score (rsq in MACH) of 0.9 is needed to achieve concordance rate higher than 95%. These results suggested that the meta-analysis results of imputed genotypes in AAs have to be carefully examined and validated because they are prone to imputation errors.

**127**

**How Confident are We About the Results from a Genome-wide Association Study Using Imputed Data in Leiden Longevity Study?**

Hae-Won Uh (1), Stefan Bohringer (1), Jeanine J. Houwing-Duistermaat (1)
(1) Leiden University Medical Center

Genotype imputation has become an essential tool in the analysis of genome-wide association scan. The success of the meta-analyses using imputed genotypes can be dramatic. Several issues, however, arise when applying these methods in practice. Many papers discuss on imputation accuracy of various software, in which positive findings were reported. In the Leiden Longevity Study (LLS) an affected sibling pair—control design was considered. The outcome of interest was exceptional longevity. One half of each 450 long-lived sibling pair was genotyped with Affymetrix 500 K array; after quality check 350 K SNPs were available. The remaining half and the control population were genotyped with Illumina 650 K array. MACH was used for imputation to fill the gaps in the Affymetrix array. The 100 top-ranking SNPs from the association analysis were selected to be genotyped with the Sequenom MassArray platform for both of the sibling pairs. Since only a part of each data set was imputed, we developed an efficiency measure that reflects impact on the tests due to uncertainty caused by imputation. We found that the results from imputed data and the replication did not match as well as expected. This might be caused by combining different arrays, and/or imputation of either cases or controls only. Morever, a more stringent threshold of Rsq greater than 0.30 seems appropriate. We recommend that results from GWAS with imputed data should be cautiously interpreted.

**128**

**Meta-analysis vs GWAS in Lung Cancer Studies**

Kwangmi Ahn (1), Carla J. Gallagher (1), Steven S. An (2), Joshua E. Muscat (1)
(1) Penn State College of Medicine
(2) Johns Hopkins University

The sequencing of the human genome has made it possible to identify an informative set of more than one million single nucleotide polymorphisms (SNPs). Genome Wide Association Studies (GWAS) have now become a standard method for the discovery and replication of new candidate loci or genes. Since GWAS use tag SNPS, they cannot be directly compared to previous candidate genes. In addition, many

associations with candidate genes have not been replicated. This study compared the results from a GWAS study of lung cancer to candidate genes that have been analyzed in meta-analysis. The goals were to determine the consistency of findings across all study designs. We studied the population of white Caucasian drawn from a European GWAS of lung cancer. We included SNPs in Linkage Disequilibrium with 10 candidate genes that were part of meta-analysis (MA). Finally, we have 14 SNPs in MA and 23 SNPs in the GWAS. We found that 5 SNPs from 4 genes in MA are significant statistically and 3 SNPs from one gene in the GWAS are. The mixed results indicate that GWAS findings are sometimes inconsistent with the results from MA. These differences may be due to the quality of studies, appropriate control of covariates and other factors. The findings indicate that for certain genes, their role in the development of lung cancer is unclear and that GWAS findings are possibly not the gold standard in etiologic research.

**129**

**Genome Wide Association of Healthy Ageing**

Nora Franceschini (1), Gil Atzmon (1), Ellen Demerath (1), Melissa E. Garcia (1), Robert Kaplan (1), Thomas Kocher (1), Maris Kuninggas (1), Kathryn L. Lunetta (1), Joanne M. Murabito (1), Anne B. Newman (1), Toshiko Tanaka (1), Alexander Teumer (1), Henning Tiemeier (1), Greg Tranah (1), Cornelia M. van Duijn (1), Stefan Walters (1)
(1) The CHARGE Aging and Longevity Working Group

Although longevity and healthy ageing show moderate heritability (20–40%), studies struggle to identify causal genes. Methods: Meta-analysis of 9 genome wide association (GWA) studies (~25 000 participants, ~8000 deaths) of the CHARGE consortium (11 years follow-up) and two survival outcomes: mortality and time to incident disease or mortality. GWA results were integrated with pathway and network analysis. Results: 18 independent single nucleotide polymorphisms (SNPs) passed the threshold ($P<10^{-5}$) for mortality and 10 for time to incident disease or mortality. Several associations were in/near genes highly expressed in the brain (*HECW2*, *HIP1*, *BIN2*, *GRIA1*), genes related to neural development/function (*KCNQ4*, *LMO4*, *GRIA1*, *NETO1*, *PTK2*) and ageing traits such as autophagy (*ATG4C*), cancer (*ATG4C, HIP1, HECW2, VWA5A*), amyotrophic lateral sclerosis (*GRIN2B*), mitochondrial depletion syndrome (*SUCLA2*), muscular dystrophy (*COL6A3*), Alzheimer's (*LMO4*) and Werner syndrome (*WRNIP1*). Pathway analysis identified an enrichment of genes involved in cell communication, and in cellular and neural development/function. Conclusions: Although our analysis did not reveal significant GWA findings for ageing, some of the top associated genes (related to neural regulation) were represented by more than one locus reinforcing the accuracy of the findings. This implicates neural tissue and brain as key factors in regulating ageing and achieving longevity.

**130**

**Genetic Loci for BMI and BMI Change in the Transition from Adolescence to Young Adulthood**

Mariaelisa Graff (1), Penny Gordon-Larsen (1), Charles C. White (2), Julius Ngwa (2), Najaf Amin (3), Tonu Esko (4), Paul Scheet (5), Claudia Schurmann (6), Alexander Teumer (6), Caroline S. Fox (7), Lu Qi (8), Rob M. van Dam (9), David Strachen (10), Andres Metspalu (3), Cornelia M. van Duijn (4), David Schlessinger (11), Henry Voelzke (6), Kari E. North (12), Sonja Berndt (13), L.A. Cupples (14)
(1) Department of Nutrition, Carolina Population Center, University of North Carolina, Chapel Hill
(2) Department of Biostatistics, Boston University School of Public Health, Boston University Medical Campus
(3) Estonian Genome Center, Institute of Molecular and Cell Biology, University of Tartu; Estonian Biocenter, Tartu
(4) Department of Epidemiology, Erasmus Medical Center, Rotterdam
(5) Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston
(6) Institute for Community Medicine, Ernst-Moritz-Arndt-Universitat Greifswald
(7) Framingham Heart Study, Framingham
(8) Department of Nutrition, Harvard School of Public Health; Channing Laboratory, Brigham and Women's Hospital, Boston
(9) Department of Nutrition, Harvard School of Public Health; Yong Loo Lin School of Medicine, National University of Singapore
(10) Division of Community Health Sciences, St. George's, University of London
(11) Gerontology Research Center, National Institute on Aging, Baltimore
(12) Department of Epidemiology, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill
(13) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda
(14) Department of Biostatistics, Boston University School of Public Health; Framingham Heart Study, Framingham

The period between late adolescence and young adulthood is a high risk period for weight gain. We sought to identify loci influencing BMI in late adolescence and change in BMI from adolescence to young adulthood. We performed a meta-analysis of 9 genome-wide association studies in Caucasian men and women that included observations of BMI level between ages 16–21 y ($N = 13530$) and BMI change between ages 16 and 35 y ($N = 6626$). Each study imputed ~2.5 million SNPs in HapMap, tested for association assuming an additive genetic model, stratified by gender and/or case-control status, accounting for age, center, principal components, and relatedness. We used multiple BMI observations at different ages to perform a growth curve analysis for BMI change. Inverse normal transformed residuals were used for both outcomes. The meta-analysis was conducted using the inverse variance weighted method. For BMI level, we identified 5 independent loci ($P<5.0\times10^{-8}$) in men and women combined near TNNI3K ($p=1.9\times10^{-12}$), TMEM18 ($p=3.0\times10^{-11}$), MC4R ($p=3.4\times10^{-09}$), FTO ($p=2.2\times10^{-09}$), and PMAIP1 ($p=4.2\times10^{-09}$); 2 in men only (near PMAIP1, $p=5.5\times10^{-10}$ and MC4R, $p=3.2\times10^{-08}$); and 1 in women only (near TNNI3K, $p=9.6\times10^{-09}$) We found no significant loci for BMI change. In summary, TNNI3K and PMAIP1 are novel hits; expression of genetic effects may relate to particular periods of the life course and vary by gender.

**131**

**A Powerful SNP-by-sex Interaction is Revealed in Numerous Novel Loci for Body Fat Distribution: Results of a Meta-Analysis of 77,000 Individuals**

Iris M. Heid (1), GIANT Consortium (2)
(1) Regensburg University Medical Center
(2) Genetic Investigation of ANThropometric Traits Consortium

Body fat distribution exhibits substantial differences between men and women. Waist-hip ratio (WHR) is an established measure of body fat distribution and a predictor of metabolic health consequences independent of overall adiposity. WHR is heritable, but few genetic variants influencing this trait have been identified. We conducted a sex-combined meta-analysis of 32 genome-wide association studies (up to 77,167 participants), following up 16 loci strongly associated with WHR adjusted for body mass index in additional 30 studies (up to 113,636 subjects). We identified 14 loci influencing WHR, including 13 novel loci in or near genes, e.g. *VEGFA*, *GRB14*, *and ADAMTS9* ($P$ from $1.9 \times 10^{-9}$ to $1.8 \times 10^{-40}$), and the known signal at *LYPLAL1*. Seven of the 14 WHR loci exhibited marked sexual dimorphism, all with a stronger effect on WHR in women than men ($P$ for sex difference from $1.9 \times 10^{-3}$ to $1.2 \times 10^{-13}$). Our data not only provide evidence for multiple loci that modulate body fat distribution, independent of overall adiposity, but also a proof-of-principle of powerful gene-by-sex interactions. These results call for more analyses separately for men and women and power computations illustrate where such an approach could gain over sex-combined approches.

**132**

**Genetics of Coronary Artery Disease: Results from the CARDIOGRAM Meta-analysis**

Inke R. Konig (1), Jeanette Erdmann (2), John R. Thompson (3), Michael Preuss (1), Devin Absher (4), Themistocles L. Assimes (5), Stefan Blankenberg (6), Eric Boerwinkle (7), Li Chen (8), Adrienne Cupples (9), Alistair S. Hall (10), Eran Halperin (11), Christian Hengstenberg (12), Hilma Holm (13), Reijo Laaksonen (14), Mingyao Li (15), Winfried Marz (16), Ruth McPherson (8), Kiran Musunuru (17), Christopher P. Nelson (3), Mary S. Burnett (18), Stephen E. Epstein (18), Christopher J. O'Donnell (19), Thomas Quertermous (5), Daniel J. Rader (20), Robert Roberts (8), Arne Schillert (1), Alexandre F. Stewart (8), Gudmar Thorleifsson (13), Unnur Thorsteinsdottir (13), Benjamin F. Voight (21), George A. Wells (8), Andreas Ziegler (1), Sekar Kathiresan (17), Muredach P. Reilly (20), Nilesh J. Samani (22), Heribert Schunkert (2)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, Lubeck, Germany
(2) Medizinische Klinik II, Universitat zu Lubeck, Lubeck, Germany
(3) Department of Health Sciences, University of Leicester, Leicester, UK
(4) Hudson Alpha Institute, Huntsville, Alabama, USA
(5) Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
(6) Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universitat Mainz, Universitatsmedizin, Mainz, Germany
(7) University of Texas Health Science Center, Human Genetics Center and Institute of Molecular Medicine, Houston, TX, USA
(8) The John & Jennifer Ruddy Canadian Cardiovascular Genetics Centre, University of Ottawa, Ottawa, Canada
(9) Department of Biostatistics and Epidemiology, Boston University, USA
(10) LIGHT Research Institute, Faculty of Medicine and Health, University of Leeds, Leeds, UK
(11) The Blavatnik School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Te
(12) Klinik und Poliklinik fur Innere Medizin II, Universitat Regensburg, Regensburg, Germany
(13) deCODE Genetics, 101 Reykjavik, Iceland
(14) Science Center, Tampere University Hospital, Tampere, Finland
(15) Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA
(16) Synlab Center of Laboratory Diagnostics Heidelberg, Heidelberg, Germany
(17) Cardiovascular Research Center and Cardiology Division, Massachusetts General Hospital, Boston, MA, USA
(18) Cardiovascular Research Institute, MedStar Research Institute, Washington Hospital Center, Washington, DC, USA
(19) National Heart, Lung and Blood Institute, Framingham Heart Study, Framingham, MA and Cardiology Division, Massachusetts General
(20) The Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
(21) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA
(22) Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK

Coronary artery disease has a strong heritability that is poorly characterised. Genome-wide studies provide a powerful approach to identify genetic variants associated with risk of common diseases.
In the consortium Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM), we combined data from 14 genome-wide association studies to compare allele frequencies of more than 2 million single nucleotide polymorphisms in 22,233 coronary artery disease patients and 64,762 controls, all of European ancestry. Promising variants were analyzed for replication in up to 49,343 additional individuals. Novel loci confirmed to be associated with risk of coronary artery disease were further investigated for correlation with coronary risk factors in population-based samples. We established association with 13 novel chromosomal regions, thus doubling the number of known loci for coronary artery disease. In this presentation, major results from this study will be shown. Associations will be presented in main and important subgroups. Conclusions regarding the genetics of coronary artery disease will be drawn.

**133**

**Impact of One-carbon Metabolism-related Gene Polymorphisms on the Risk of Colorectal Cancer: A Meta-analysis**

Zahra Montazeri (1), Julian Little (1), Mahmood Tazari (2)
(1) University of Ottawa
(2) Carleton University

Although colorectal cancer (CRC) is the second most common cause of cancer death in North America, it is a

highly treatable form of cancer when detected in its early stages. Observational studies have suggested that higher red cell folate levels and higher reported intakes of dietary folate and supplemental folic acid are associated with reduced risk of CRC. Different studies show that functional polymorphisms in genes encoding one-carbon metabolism enzymes, *MTHF, RMTR, MTRR*, CBS, and TS, influence folate metabolism and thus might be suspected of having an impact on the risks of CRC. As part of the development of a field synopsis of the population impact of genetic variation on CRC, we are using meta-analytic methods to investigate the impact of genetic variation in one-carbon metabolism on CRC. In compare with the other studies, current analysis is more comprehensive and includes meta-analyses of variants of other genes thought to affect one-carbon metabolism for which evidence has been accumulating. An extensive search has been conducted using several electronic databases up to April 2010. The results are extracted from studies eligible for inclusion. In addition to standard tests, possible reasons for heterogeneity such as adequacy of addressing population stratification will be examined. The possible occurrence of publication bias is also being examined by formal testing. Some proposed enhancements of statistical methods of meta-analysis will be examined.

**134**

**Critical Issues in the Meta-analysis of Genome-wide Association Studies (GWAS) with Small Number of Cases: The CKDGen Consortium**

Cristian Pattaro (1), Christian Fuchsberger (2), Alexander Teumer (3), Ming-Huei Chen (4), Carsten A. Boger (5), Matthias Olden (5), Xiaoyi Gao (6), Afshin Parsa (7), Qiong Yang (4), Daniel Taliun (1), Jeffrey R. O'Connell (7), Aaron Isaacs (8), Daniel I. Chasman (9), Iris M. Heid (5), W.H. Linda Kao (10), Caroline S. Fox (11), Anna Kottgen (12), on behalf of the CKDGen Consortium (13)
(1) Institute of Genetic Medicine, European Academy of Bolzano (EURAC), Italy
(2) University of Michigan, Ann Arbor, USA
(3) Greifswald University, Germany
(4) Boston University, USA
(5) Regensburg University, Germany
(6) Washington University, St. Louis, USA
(7) University of Maryland, Baltimore, USA
(8) Erasmus University, Rotterdam, Netherlands
(9) Harvard University, Boston, USA
(10) Johns Hopkins University, Baltimore, USA
(11) Framingham Heart Study and Harvard University, Boston, USA
(12) Freiburg University, Germany and Johns Hopkins University, USA
(13)

To uncover genetic loci associated with severe forms of common chronic diseases, meta-analyses are often applied to GWAS of binary traits based on clinically relevant thresholds to quantitative biomarkers. Due to small case numbers, extreme thresholds can produce spuriously significant results that are difficult to control. We explored this issue in an inverse-variance weighted fixed-effects meta-analysis of 11 GWAS of ~2.5 million SNPs on chronic kidney disease stage 3b (CKD3b) with cases and controls defined by glomerular filtration rate $<45$ and $>60\,ml/min/1.73\,m^2$, respectively. To limit the risk of false positives, results were filtered on combinations of minor allele frequency (MAF, 1 to 10%), between-study heterogeneity ($I^2$ statistic, 100 to 25%) and % of studies with the SNP available (PS, 25 to 75%). Cases were 25 to 199 per study (total = 1061), controls 637 to 21,940 (50,729). Despite the inflation factor ? = 1.060, we observed an excess of small *p*-values with 116 independent SNPs had p?5e−06 vs 5 expected for a million tests. MAF filtering had a bigger impact than $I^2$ and PS on removing excessively low *p*-values. ? = 1.016 was minimum for MAF?10%, $I^2$?25% and PS?25%, identifying 4 loci suggestively associated with CKD3b (p?5e−06). Meta-analyses of binary trait GWAS with small case numbers require careful selection of SNPs for replication. While our results await external replication, MAF filtering seems to be the most effective way to reduce the risk of false discoveries.

**135**

**Synthesis-View: Visualization and Interpretation of SNP Association Results for Multi-cohort, Multi-phenotype Data and Meta-analysis**

Sarah Pendergrass (1), Scott M. Dudek (1), Dana C. Crawford (1), Marylyn D. Ritchie (1)
(1) Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville TN, USA

GWAS findings are being further investigated for replication and characterization, both in the populations in which the initial GWAS findings were discovered (such as European-Americans) as well as in new cohorts and populations. To increase sample size, meta-analysis is often used to combine results from multiple research sites. Multiple independent and correlated phenotypic measurements may be included in the analyses, such as cardiovascular disease and related biomarkers (lipids, inflammation, etc). Visualization of these data, including the integration of *p*-values or other metrics is integral to being able to interpret as well as share the complex and multi-layered results of these follow-up studies. The software "Synthesis-View" has been developed to visually synthesize multiple pieces of information of interest from these studies with the flexibility to perform multiple types of data comparisons. Through the use of stacked data-tracks information on SNP genomic location, presence of the SNP in a specific study or analysis, as well as related information such as genetic effect size and summary phenotype information, is plotted according to user preference. Synthesis-View provides a way to understand in greater depth the relationships between SNPs, strata, sample size, and phenotypes. Synthesis-View is freely available for non-commercial research institutions, for full details see https://chgr.mc.vanderbilt.edu/synthesisview.

**136**

**Genome-wide Association Study Meta-analysis Identifies Novel Susceptibility Loci for Severe Diabetic Retinopathy.**

Anna Tikhomirov (1), Michael A. Grassi (1), Sudha Ramalingam (2), Jennifer E. Below (1), Nancy J. Cox (1), Dan L. Nicolae (1)
(1) The University of Chicago
(2) PSG Institute of Medical Sciences and Research

Diabetic retinopathy is a leading cause of blindness. The genome-wide association study (GWAS) is a powerful tool

for identifying genetic variation affecting predisposition to common, complex diseases like diabetic retinopathy as it allows for a systematic survey of the entire genome without a priori knowledge of potential susceptibility alleles. We performed association tests for the severe manifestations of diabetic retinopathy in two large, type I diabetic cohorts ascertained from the Genetics of Kidney in Diabetes (GoKinD) and the Epidemiology of Diabetes Intervention and Control Trial (EDIC) studies. A combined total of 1995 subjects (281 cases, 1714 controls) were studied. The meta-analysis identified a significant intergenic single nucleotide polymorphism (SNP), rs227455, on chromosome 6 ($p$-value $1.59 \times 10^{-7}$) that is an eQTL (expression quantitative trait locus) for BRCC3, a component of the BRCA1/BRCA2 complex. Copy number polymorphism analysis yielded a significant association at rs10521145 ($p$-value $3.43 \times 10^{-6}$) in the intron of CCDC101, a histone acetyltransferase. rs10521145 is in perfect linkage disequilibrium (LD) with the copy number region CNVR6685.1 on chromosome 16 at 28.5 Mb, a gain/loss site that contains the sulfotransferase genes SULT1A1 and SULT1A2. In summary, we identified novel genetic loci that are associated with the sight threatening complications of diabetic retinopathy.

**137**

**How to Detect Genetic Loci in the Presence of Heterogeneity between Men and Women? Methodological Issues and Results from Gender-specific Meta-analyses for Human Anthropometric Measures: The GIANT Consortium**

Thomas W. Winkler on behalf of the GIANT (Genetic Investigation of ANthropmetric Traits) Consortium (1)
(1) Department of Epidemiology and Preventive Medicine, Regensburg University Medical Center

Recently, a few sexually dimorphic variants associated with human anthropometric traits were identified through sex-specific analyses following up genome-wide significant hits from a gender-combined meta-analysis. To unveil further variants, we conducted gender-specific meta-analyses in 58 genome-wide association studies (up to 60,587 men, 73,138 women). We analyzed the association between ~2.8 million SNPs and 9 traits: height, weight, BMI, waist and hip circumference and the ratio of both with and without adjustment for BMI. Differential gender effects can be the result of SNPs with (1) marked association in both genders, but opposite effect directions (OED); or (2) marked association in only one gender with concordant effect directions (CED). Our strategies to detect CED were: (1) Select SNPs that are significant in one gender, but not in the sex-combined analysis and try to replicate either the sex-specific association or the gender-heterogeneity; (2) Select SNPs with low (<1E-3) gender-combined $P$ values and test these for heterogeneity. To find OED, we identify SNPs with significant heterogeneity in a genome-wide scan, and confirm them by replication. We report the most stringent (FDR<1%) stage I associations that will be tested in replication samples. For waist: SNPs in genes SMOC2, PDZRN4 were found to show strong sex-specific effect. Although replication is in progress, we seem to have strong evidence for several novel gender-specific associations.

**138**

**Meta Analysis of GWAS Studies with Overlapping Individuals**

Siyan Xu (1), Emelia J. Benjamin (2), Kathryn Lunetta (1)
(1) Biostatistics Department, Boston University School of Public Health
(2) Department of Medicine, BU School of Medicine, Department of Epidemiology, BU School of Public Health

Occasionally some individuals will be included in multiple GWAS that are meta-analyzed for a trait. Overlapping samples may occur when the studies draw subjects from the same small population, particularly if a disease is uncommon. When one or more individuals are present in multiple GWAS, the GWAS SNP tests are correlated. If the GWAS are combined via traditional meta-analysis, the test statistics will be inflated due to ignored correlation. Recently, Lin and Sullivan (The American Journal of Human Genetics 2009;85:862–872) proposed a method for taking into account known overlap in subjects in meta-analysis of case-control studies. However, due to privacy concerns, it is sometimes not possible to identify the specific individuals that contribute to multiple studies. We show through simulation and theoretical calculations that one can estimate the correlation between studies by calculating the correlation between the study regression estimates or test statistics. We find that the correlation in regression coefficients for two samples for a logistic model with a binary trait or a linear model with a quantitative trait is approximately equal to the proportion of overlapping individuals in the two samples. We then account for the study correlations in the meta analysis using optimal weights. We show that incorporating between-study correlation into the meta-analysis in this manner is more powerful than using genomic control and yields correct type I error.

**139**

**Combinations of Newly Confirmed Glioma-Associated Loci Link Regions on Chromosomes 1 and 9 to Increased Disease Risk**

Tun-Hsiang Yang (1), Mark Kon (2), Jui-Hung Hung (1), Charles DeLisi (3)
(1) Bioinformatics Program, Boston University
(2) Bioinformatics Program; Department of Mathematics and Statistics, Boston University
(3) Bioinformatics Program; Department of Biomedical Engineering, Boston University

Among the problems plaguing efforts to associate genes with disease is the small percentage of common genes found when studies using different populations are compared. This report demonstrates general methods that considerably improve recovery of reliable discriminators between glioma and normal phenotypes. In particular we find the following. (1) A meta-analysis of single nucleotide polymorphism (SNP) data from The Cancer Genome Atlas (TCGA), and the Adult Glioma Study identifies 15 SNPs which represent 6 genomic regions—5q15.33, 9q21.3, 11q13.4, 1p21.2, 3q26.2 and 7p15.3—four of them not previously reported. Eight genes known to be cancer-associated are included in, or are in strong linkage disequilibrium with, one or more of these SNPs. The importance of this increase rests in part on the non-linear increase in the number of significant SNP combinations,

and the resulting ability to more reliably infer processes that underlie disease etiology. The relative risk associated with SNP pairs and triplets is briefly discussed, the highest odds ratios for SNP combinations occurred on chromosomes 1 and 9. (2) Some SNPs that are not invariant across populations are in reproducible pathways, suggesting that they affect the same biological process, and that population discordance can be partially resolved by evaluating processes rather than genes.

**140**

**Are large Sample Sizes Always Necessary in Genome-wide Association Studies?**

Emmanuelle Bouzigon (1), Florent Monier (1), Eve Corda (2), Florence Demenais (1)
(1) INSERM, U946
(2) CEPH

A recent genome-wide association study (GWAS) of a quantitative phenotype, applied to less than 200 subjects taken from the extremes of the trait distribution, reported 3 loci at $P<10^{-15}$ (Menzel et al. Nat Genet 2007;39: 1197–1199). To investigate further the advantages of this sampling design, we compared the outcomes of a GWAS for lung function measurements conducted in 530 subjects by examining either the full distribution of the trait or extreme discordant values (25–75% percentiles). Analyses were performed using linear regression under an additive genetic model. Each SNP effect was tested by the Wald test. Simulations have shown that the type I error of the latter test was not inflated by extreme discordant sampling designs. We found that the correlation coefficient of $-\log_{10}(p\text{-values})$ associated with the Wald tests applied to the full sampling and extreme sampling design was 0.92 for the 513,000 SNPs analyzed. Among SNPs with allele frequency greater than 0.15 and with P?$10^{-4}$, 80% of SNPs detected by the full sampling design were detected by the extreme sampling design. Eight of these SNPs had P?$10^{-5}$ under either one of these designs and three of them were identical. The extreme discordant sampling design appears to retain the power of the full sampling design and can be a cost-effective strategy for GWAS of quantitative traits. Non-parametric and exact tests that can be used at smaller allele frequencies will be further examined.
Funded by European Commission (GABRIEL)

**141**

**A Comparison of Power and Sample Size in Methods of Testing for Association on the X Chromosome**

Geraldine M. Clarke (1), Andrew P. Morris (1)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford

Testing for genotype-phenotype association on the X chromosomes poses special problems. Specifically, at most single nucleotide polymorphisms, females contribute two alleles whereas males contribute only one. Moreover, X-inactivation ensures that in females only one of these copies is activated. In this situation, the assumption that the effect of one copy of a variant allele on a phenotype will be the same in males as in females may not be realistic. Various methods for detecting genetic association with loci on the X chromosome that find a way to combine evidence from both sexes and deal with X-inactivation have been proposed including: (1) treat males as homozygous females taking no account of the X chromosome; (2) as for (1) and also adjusting for sex as a covariate; (3) meta-analysing the results of male and female specific tests of association. Here, we compare the power and sample size, as well as type I error inflation, of these different population-based methods of disease-gene association on the X chromosome.

**142**

**GSI: Genotype Sample Investigation**

Claus T. Ekstr?m (1)
(1) Univeristy of Copenhagen

Genome-wide association studies (GWAS) have rapidly become a standard method for identification of disease genes and genes influencing quantitative traits. The large number of SNPs tested mean that very large samples are required in order to detect an appreciable fraction of the truly risk causing variants and that data needs to be carefully screened to remove potential errors. GWAS often employs independent individuals which may make error identification more problematic than for other gene-identification approaches, where related individuals are available. With GSI: we use known associations between genotypic markers and phenotypes to identify potential sample mix-ups. A well-known check is to see if the observed genotypes match the sex of the individual. However, less stringent associations between known genotypes and quantitative or qualitative phenotypes can be combined to identify potential sample mix-ups. We show how a few known independent associations can be combined to flag misidentifications, discuss power issues, discuss how related individuals can be used to improve the method, and apply the method to a larger Danish study.

**143**

**Exploring the Protocol Used to Quantify Biological Analytes at UK Biobank and the Influence of that Protocol on the Power of Genetic Association Studies**

Amadou Gaye (1), Paul Burton (1), Tim Peakman (2)
(1) University of Leicester
(2) UK Biobank

UK Biobank processes biological samples within 36 hours from collection. A certain proportion of the variation between samples may be due to difference in processing time. This can bias the results of association studies using biological analytes as measures of environmental determinants of disease.
The contribution of delayed processing to the overall variation between subjects was determined by variance component analysis. The impact of delayed processing on the power of a case-control study was estimated. An error can occur when samples are drawn manually to carry out assays; the influence of such sample misclassification on power was investigated.
15 of the 47 analytes have a contribution of processing delay ? 10% of the overall variation between subjects. For those analytes a sample size increase between 12% and 159% is required to compensate for the power loss due to delayed processing. The sample size required has to be multiplied by 1.24 and 1.11 to compensate for the power drop due to the respective misclassification rates of 10% and 5%.

If cases and controls for an analysis are derived from different studies, it is critical that care is taken to ensure that the difference between protocols is adjusted for. It is important to specify a limited delay in processing for analytes that are very sensitive to delayed assay. In the presence of well measured data the bias arising from manual misclassification of samples may have a high impact on power.

**144**

**ESPRESSO: A Simulation Platform for Realistic Power Analysis and Sample Size Estimation**

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

One of the main limitations of genetic association studies in the investigation of common diseases is the lack of power. The aim of this work is to build up a simulation platform that enables the design of studies that take full account of the impact of key determinants of power. The simulation is an unmatched case-control. Genetic variants can be modelled as binary or additive and environmental exposure can be modelled as binary or quantitative. Unlike in conventional approaches, assessment errors in both exposure and outcome, impact of unmeasured aetiological determinants as well as heterogeneity in disease are included as parameters in the simulation to allow for realistic sample size estimation. The data is analysed by unconditional logistic regression. A standalone simulation platform has been developed. It enables the impact of the key power determining factors to be studied and to be taken into full account in designing large scale biobanks and association studies. This program allows one to realistically calculate the sample size required to study a disease under several models.
Along with power and sample size calculations ESPRESSO can also be used to answer relevant scientific questions that have been identified as being critical for biobanks and association studies. It has been used recently to determine the impact of a UK Biobank lab protocol on the power of case-control studies; this project has been extremely successful.

**145**

**A Nonparametric Approach to Population Based Association Tests**

Sharon M. Lutz (1) Nan Laird (1), Christoph Lange (1)
(1) Harvard School of Public Health

In population-based genetic association studies, the standard approach is to model the phenotype of interest as a function of the offspring genotype. In order for this approach to be valid, the distributional assumptions about the phenotype have to be correct, including the specification of potential ascertainment conditions that were used for the recruitment of the study. Here we propose an alternative approach to population-based association analysis. We develop a general framework of conditional score-tests that treat the genetic information as the random variable and condition upon the phenotypic information. This approach is robust against phenotypic heterogeneity. However, the power of the approach can be increased by incorporating a plausible model for the phenotype into the test statistic. Based on theoretical considerations and on simulation studies, we show that the new approach is robust against mis-specification of phenotype assumption and, at the same time, achieves the same power level as standard genetic association tests for population-based designs. For studies with ascertainment conditions for quantitative traits, our simulation studies also illustrate that, although the proposed tests do not require any adjustment for the ascertainment conditions, their power levels are much higher than those of standard approaches.

**146**

**Quality Control Pipeline for Genome-Wide Association Studies in the eMERGE Network: Comparing Single Site QC to a Merged QC Approach**

Marylyn Ritchie (1), Loren Armstong (2), Yuki Bradford (1), Chris Carlson (3), Dana Crawford (1), Andrew Crenshaw (4), Mariza de Andrade (5), Kim Doheny (6), Jonathan Haines (1), Geoff Hayes (2), Gail Jarvik (3), Lan Jiang (1), Hua Ling (6), Iftikhar Kullo (5), Rongling Li (7), Teri Manolio (7), Martha Matsumoto (5), Cathy McCarty (8), Andrew McDavid (3), Daniel Mirel (4), Lana Olson (1), Justin Paschall (9), Elizabeth Pugh (6), Luke Rasmussen (8), Russ Wilke (1), Rebecca Zuvich (1), Stephen Turner (1)
(1) Vanderbilt University Center for Human Genetics Research
(2) Northwestern University, Chicago, IL
(3) University of Washington/Group Health Cooperative, Seattle, WA
(4) Broad Institute of MIT and Harvard, Cambridge, MA
(5) Mayo Clinic, Rochester, MN
(6) Center for Inherited Disease Research (CIDR), Johns Hopkins University, Baltimore, MD
(7) National Human Genome Research Institute (NHGRI), Bethesda, MD
(8) Marshfield Clinic, Marshfield, WI
(9) National Center for Biotechnology Information (NCBI)

Genome-wide association studies (GWAS) are being conducted at an unprecedented rate in disease-based cohorts and have increased our understanding of the pathophysiology of complex disease. Regardless of context, the practical utility of this information will ultimately depend upon the quality of the original data. It has been established that quality control (QC) procedures for GWAS are computationally intensive, operationally challenging, and under constant evolution. What has not yet been explored in detail are the challenges that emerge when multiple GWAS datasets, are merged for downstream GWAS analysis; a scenario that is likely to increase in frequency with the advent of dbGaP. The genomics workgroup of the NHGRI funded electronic Medical Records and Genomics (eMERGE) network has spent a considerable amount of effort developing strategies for QC of these data. We compare the characteristics of various QC measures between each of the five eMERGE sites and the merged dataset. Here we enumerate some of the challenges in QC of merged GWAS datasets, including merging data from 2 genotyping centers and other errors inherent from merging ~17000 samples GWAS data, and describe the approaches that the eMERGE network is using to guarantee quality assurance in GWAS data, thereby minimizing potential bias and error in GWAS results. Finally, we describe the best practices that we have decided upon and discuss areas of ongoing and future research.

**147**
**Power of Genome-wide Search Strategies for Binary Trait Loci**
Zheyang Wu (1), Hongyu Zhao (2)
(1) WPI
(2) Yale

For more fruitful discoveries of genetic variants associated with diseases in the genome-wide association studies, it is important to know whether joint analysis of multiple markers is more powerful than the commonly used single-marker analysis. We provide analytical power calculations for various methods to detect binary trait loci: the marginal search, the exhaustive search, the forward search, and a two-stage screening search. In the context of binary traits, we define genetic models based on disease odds for joint genotypes and derive asymptotic distributions of score tests in logistic model fitting. Our statistical framework takes into account LD, random genotypes, and correlations among test statistics. We derive the analytical results under two power definitions: the power of finding all the associated markers and the power of finding at least one associated marker, and two types of error controls: the discovery number control and the Bonferroni type I error rate control. After demonstrating the accuracy of our analytical results by simulations, we apply them in a broad genetic space to investigate the relative performances of different search methods under different power and error control definitions. The relative performances for binary trait have both similarities and differences from those in quantitative trait study. Our analytical study provides rapid computation as well as insights into the statistical mechanism of capturing genetic signals.

**148**
**Power Gains using Phenotyped but Ungenotyped Relatives in Genetic Association Studies**
Wei V. Zhuang (1), J.M. Murabito (2), K.L. Lunetta (1)
(1) Boston University School of Public Health
(2) Boston University School of Medicine

In some long-term longitudinal studies such as the Framingham Heart Study, there are individuals with rich phenotype data who died before providing DNA for genetic studies. Thus, the individuals have no genotype data but have phenotypic data, and often genotypes of relatives such as offspring and spouses are available. Visscher and Duffy (Genet Epid 2006;30:30–36) explored the power increase due to the inclusion of such individuals in a genetic association test for a quantitative trait in the case of a single genotyped SNP in a single type of relative. Chen and Abecasis (Am J Hum Genet 2007;81:913–926) showed that the power to detect a SNP association with a quantitative variable may be increased by imputing the SNP genotypes for individuals genotyped on a sparse genotype set using their relatives genotyped on a dense set of genotypes. We investigate the power gain when we impute genotypes for phenotyped relatives of genotyped individuals in genetic association studies of quantitative and dichotomous traits. We also examine which factors influence the change in the power and type I error. We verify our theoretical results with simulations. We show that while the expected power when ungenotyped individuals are included increases, the power for any one study may be increased or decreased. The heritability of the trait and the concordance of the trait between the genotyped individuals and their ungenotyped relatives are important factors in power improvement.

**149**
**WITHDRAWN**

**150**
**Retaining Power: Is it Possible to Simply and Effectively Adjust for Multiple Comparisons in a Candidate Gene Region?**
Audrey E. Hendricks (1), Richard H. Myers (2), Kathryn L. Lunetta (1)
(1) Biostatistics, Boston University School of Public Health
(2) Neurology, Boston University Medical School

Widely accepted significance thresholds for genome-wide scans that account for multiple testing exist, but there is less agreement on the methods to control for multiple testing in a candidate gene region. The Bonferroni correction is simple and widely used, but is often overly conservative due to the high correlation among gene region SNPs. Permutation, another common solution, provides a type I error that asymptotically approaches the chosen significance level. However, permutation is computationally intensive and is, thus, not often used for high-throughput analyses or simulations. An intriguing option is to calculate the effective number of independent SNPs in a gene region, and to use this value in the Bonferroni correction. This solution is less conservative than the simple Bonferroni correction and is computationally simple. We evaluate four methods for calculating the effective number of independent SNPs: Cheverud (2001), Li and Ji (2005), Gao et al. (2008), and Galwey (2009). We compare the type I error rates of these methods over simulation scenarios where we vary the underlying LD structure, region size, and sample size. We find that all methods are sensitive in varying degrees to changes in sample size and LD structure. Overall, Galwey's method appears to be anticonservative, Cheverud's method appears to be overly conservative, and both Li and Ji's as well as Gao et al.'s methods produce type I errors close to the chosen significance level.

**151**
**Don't Let LD Bring You Down—A Fast Permutation Test Algorithm for Powerful Genome-wide Association Testing**
Roman Pahl (1), Helmut Schafer (1)
(1) Institute of Medical Biometry and Epidemiology

In genome-wide association studies (GWAS) examining hundreds of thousands of single nucleotide polymorphism (SNP) markers, the potentially high number of false positive findings requires statistical correction for multiple testing, for which permutation tests are considered the gold standard, because they not only provide unbiased genome-wide type I error control but also the highest statistical power. On the other hand, permutation testing is computationally very intensive, especially with large-scale data sets of modern GWAS. In recent years, the computational problem has been circumvented by using approx-

imations to permutation tests, which, however, may be biased. Here we present a novel permutation test algorithm one or more orders of magnitude faster than existing implementations, enabling efficient permutation testing on a genome-wide scale. Our algorithm does not rely on any kind of approximation and hence produces unbiased results identical to a standard permutation test. The speed-up in parts is achieved by exploiting correlations between SNPs due to linkage disequilibrium, so that our algorithm shows a particularly effective performance when analyzing high density marker sets. The basic underlying algorithms as well as the freely available software are presented along with runtime results for typical applications on a genome-wide scale.

**152**

**Latent Class Model with Familial Dependence to Address Heterogeneity in Complex Diseases: Adapting the Approach to Family-based Association Studies**

Alexandre Bureau (1), Jordie Croteau (2), Aurelie Labbe (3), Chantal Merette (1)
(1) Universite Laval
(2) Centre de recherche UL - Robert-Giffard
(3) McGill University

Clinical diagnoses of complex diseases may often encompass multiple genetically heterogeneous disorders. Latent class (LC) analysis can be applied to measurements related to the diagnosis, such as detailed symptoms, to define more homogeneous disease sub-types, influenced by a smaller number of genes that will thus be more easily detectable. My collaborators and I have previously developed a LC model allowing dependence between the latent disease class status of relatives within families. Our strategy to incorporate the posterior probability of class membership of each subject in parametric linkage analysis is not directly transferable to association tests. Under the framework of family-based association tests (FBAT), we propose to make the contribution of a subject to the FBAT statistic proportional to his posterior class membership probability. Simulations show a modest but robust power advantage compared to assigning each subject to his most probable class and to the analysis of the disease diagnosis. We examined the association to LCs formed using autism spectrum disorder (ASD) symptoms in eight regions previously reported associated to autism in families from the Autism Genetics Research Exchange. The analysis using the posterior probability of membership to a LC detected an association in the JARID2 gene as significant as that for ASD ($p = 3 \times 10^{-5}$). The larger effect size for the LC than for ASD (odds ratio = 2.17 vs. 1.55) suggests a greater genetic homogeneity.

**153**

**Linkage Analysis of Carotid Intima-Media Thickness in the Jackson Heart Study**

Sarah G. Buxbaum (1), Lynette Ekunwe (1), Ervin R. Fox (2), Greg Evans (3), Daniel F. Sarpong (1)
(1) Jackson State University
(2) University of Mississippi
(3) Wake Forest University

The Jackson Heart Study (JHS) is a longitudinal study of 5301 African Americans in the Jackson, MS metropolitan area. Nested within the JHS is a family study. 374 autosomal microsatellite markers were used to conduct a linkage analysis of carotid intima-media thickness (cIMT), a measure of subclinical atherosclerosis, using data from 1091 full sibs and 305 half sibs. Model free linkage analysis was conducted in the SIBPAL program in the S.A.G.E. package. Two cIMT measures were analyzed: sum45_1, an estimate of average far wall cIMT of the right and left common, bifurcation and internal carotid arteries, and CCA45_1, an estimate of average right and left common carotid far wall cIMT. Both measures were estimated to be 29% heritable based on full sib correlations in FCOR in S.A.G.E. Particularly strong evidence for linkage was found on chromosome 19p13. A multipoint analysis of CCA45_1 gave a LOD score of 5.7 and a singlepoint LOD of 5.1 at the same marker. At the same locus, sum45_1 also had a peak with a multipoint LOD score of 2.5 and a singlepoint LOD of 2.9. Chromosome 3 also had significant linkage with sum45_1 (LOD = 4.3) at 108 cM, but not with CCA45_1. Suggestive linkage was found on chromosome 12, but only with CCA45_1 (LOD = 2.4). Singlepoint and multipoint LOD scores of 4 and higher were found on chromosome 18 using CCA45_1, while the highest LOD for sum45_1 on chromosome 18 was 2.63 at the p arm telomere.

**154**

**Testing and Estimating Within-family Association for Dichotomous Traits**

Wei-Min Chen (1), Ani Manichaikul (1), Stephen S. Rich (1)
(1) Center for Public Health Genomics, University of Virginia

A large proportion of family-based association studies involve phenotype data that were either previously collected for affected-sib-pair linkage analysis, or newly collected using a parent-offspring trio design. In either case, there is little phenotypic variation across families, and it is sufficient to examine association that only exists within families. Not only is the statistical power to identify genetic variants maintained, testing the within-family association also gives the advantage of eliminating confounding factors such as population substructure and genetic heterogeneity across families. We propose an association test under a quasi-likelihood framework to test and estimate within-family association in general pedigrees. Our extensive computer simulations show our proposed test consistently outperforms existing association methods that test within-family association, including the recently published GDT method, while being as powerful as association tests that examine total association when the pedigree structure is the same across families. Our proposed methods also provide accurate estimation of association effect size, a major improvement over existing methods. The analysis a large-scale T1D linkage dataset consisting of >10,000 individuals from the T1DGC shows that our proposed method is able to identify more previously identified variants than any existing association methods (including GDT and MQLS).

**155**

**Genome-Wide Linkage Scan of Mismatch-Repair Proficient Colorectal Cancer Families**

Mine S. Cicek (1), Brooke L. Fridley (1), Daniel J. Serie (1), William R. Bamlet (1), Julie M. Cunningham (1), Noralane

M. Lindor (1), Stephen N. Thibodeau (1), John Potter (2), Ellen L. Goode (1)
(1) Mayo Clinic, Rochester, MN
(2) Fred Hutchinson Cancer Research Center, Seattle, WA

A substantial proportion of familial colorectal cancer (CRC) is not a consequence of mismatch repair (MMR) mutations, supporting the existence of additional loci. To gain further understanding of MMR-proficient CRC, we conducted a genome-wide linkage scan in 548 families with no identified germline MMR mutations, including 293 white families with no evidence of defective MMR. We studied families from the Colon CFR, a multi-site NCI-supported consortium, and from Newfoundland, defining proficient MMR (pMMR) families as those with no (MSS) or low (MSI-L) microsatellite instability and presence of complete immunohistochemical expression for the following MMR genes: *MLH1*, *MSH2*, *MSH6*, and *PMS2*. Mismatch-repair proficient family members ($N = 1{,}458$) with an average family size of 5.0 genotyped individuals (2.2 affected and 2.8 unaffected), were assessed using the Affymetrix 10 k 2.0 Array or the Illumina Linkage Panel 12 (10,091 pooled SNPs with $r^2 < 0.1$) and analyzed using MERLIN. Assuming a dominant mode of inheritance, our results provide evidence for linkage among pMMR families on chromosome 12 (HLOD = 3.01, 285.15 cM, near SNP rs952093). In addition, results suggest multiple regions with modest evidence of linkage ($1.5 < \text{HLOD} < 3$) in specific pMMR subsets, defined by number of affected cases per family (2, 3, 4+), mean age of genotyped affected cases ($<50$, 50–59, $>60$), microsatellite instability (MSS, MSI-L), and ascertainment method (population-based, clinic-based).

**156**

**Fine Mapping of a Locus (TST2) Controlling the Extent of Antimycobacterial Immunity in an Area Hyperendemic for Tuberculosis**

aurelie cobat (1), alexandre alcais (1)
(1) INSERM U980 - University Paris Descartes, Paris, France

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 extent of TST reactivity, i.e. the intensity of T-cell-mediated delayed type hypersensitivity (DTH) to tuberculin varies greatly. Recently, a genome-wide linkage search for loci impacting on TST reactivity in a panel of 128 families including 350 siblings from an area hyperendemic for tuberculosis identified a major locus (*TST2*), on chromosome 5 ($P < 10^{-5}$) that controls the extent of TST reactivity (Cobat et al, 2009). We report here on a genocentric fine mapping of the *TST2* region in the same cohort as the one used for linkage analysis. One hundred and thirteen SNPs were genotyped to capture genetic variation within the 13 known genes in the *TST2* region. Consistent with experimental data in mice, we identified SLC6A3, encoding the dopamine transporter DAT1, as the most promising gene for further studies. In addition, we used this dataset to compare the performance of several methods developed in the context of familial association studies of quantitative traits such as the those implemented in the FBAT software and approaches relying on the estimating equations framework.

Reference:
[1] Cobat et al. 2009. J Exp Med 206(12):2583–2591.

**157**

**Exploratory Association Analysis in a Subset of Finnish Prostate Cancer Families Linked to 8q24**

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, Baltimore, MD
(2) Institute of Medical Technology, University of Tampere and Tampere University Hospital, Tampere, Finland
(3) National Human Genome Research Institute, National Institutes of Health, Baltimore, MD/Fox Chase Cancer Center, Philadelphia, PA
(4) Department of Urology, Tampere University Hospital, University of Tampere, Tampere, Finland

Prostate Cancer (PRCA) is a leading cause of cancer death among men. Familial inheritance of PRCA has been demonstrated. Our genome-wide linkage analysis in 69 Finnish hereditary PRCA (HPC) families found significant linkage on 2q37.2 and 17q21–22. Ordered subset analysis (OSA) conditioned on non-parametric linkage to these loci to detect other loci linked to HPC in subsets of families showed a linkage peak on 8q24.22–q24.3 (OSA LOD = 3.195, ?LOD = 2.963, $p = 0.02$) in 15 families weakly linked to 2q37. Linkage and association to 8q24 in HPC has been reported in many populations. We conducted a case-control association study using 27 SNPs within the 1-LOD drop region of maxLOD scores in the 8q24 region. We selected one HPC affected individual from each of the 15 families as cases. Unrelated married-in individuals from all 69 Finnish families served as controls. Using PLINK, we tested for marker association of HPC among cases and controls using chi-square and logistic regression. Association results revealed 3 nominally significant SNPs within this 1-LOD drop region: rs1398296 ($p = 0.012$), rs4246828 ($p = 0.045$) and rs7386971 ($p = 0.081$). These SNPs were also nominally significant using logistic regression, $p = 0.012$, $p = 0.035$ and $p = 0.07$, respectively. These SNPs remained significant after permutation tests but not after correction for multiple testing. Future meta-analyses of the 8q24 region may increase the power to detect markers significantly associated with HPC.

**158**

**Investigating the Genetic Susceptibility of Congenital Polycythemia through Shared Genomic Segment and Linkage Analysis in a Unique Extended Family**

Karen Curtin (1), Nicola J. Camp (1), Felipe R. Lorenzo (2), Sabina Swierczek (2), Josef T. Prchal (2)
(1) University of Utah School of Medicine, Genetic Epidemiology
(2) University of Utah School of Medicine, Hematology

Using shared genomic segment (SGS) and linkage analysis, we studied an extended family with polyclonal congenital polycythemia that appears to follow an autosomal dominant pattern of inheritance that cannot be explained by known susceptibility variants. Of 26 phenotyped members, 5 affected and 11 unaffected subjects were

genotyped on an Affy 500 k platform.After quality control 317,612 SNPs were used in SGS to determine regions of possible excessive genomic sharing in genotyped cases. A pruned set of 8394 SNPs selected to maximize information content (heterozygosity?0.3) and minimize linkage disequilibrium (spacing?100 kb, $r^2$?0.16) were assessed in parametric multipoint linkage using a Markov chain Monte Carlo method (MCLINK) which allows for fully informative multilocus analysis on extended pedigrees. SGS identified genomic regions shared by all cases at chromosomes 18 (1528 SNPs potentially shared); 14 (1001 SNPs); 16 (818 SNPs); and 3 (815 SNPs). The chr.14 region (65,182,986–73,386,753 bp) contains a subregion of 74 consecutive SNPs at which all cases are homozygous. Linkage identified a peak LOD score of 3.41 on chr.16 which is contained within the SGS region (77,507,755–81,049,285 bp). SGS and multipoint linkage indicate four regions of interest that may contain a de novo locus in this family; in particular, an 8Mb region on chr.14 which may contain a deletion or segment of uniparental disomy and a 3.5 Mb region on chr.16 identified by both analysis techniques.

**159**

**Combining Multi-marker Association Tests for Quantitative Traits in the Family-based Association Study**

yilin dai (1), jianping dong (1), renfang jiang (1), ling Guo (1)
(1) Michigan Technology University

We proposed a new multi-marker test for a family-based association study of candidate region detection or a genome-wide scan using automatic weighted sum of two association tests (FBAT_WS). One of them estimates the genetic effect from both within-families and between-families variation. Another is from between-families variation. The weights are computed automatically based on the estimation of the population stratification existing in the family data. Due to population stratification and linkage disequilibrium which cause the bias of the estimate, the permutation procedure is employed and descried for this situation. One distinct advantage of family-based study is robustness to the population stratification. If only consider the estimate of the genetic effect from the within-families variation to avoid population stratification, we may lose the useful information from between-families variation. Therefore, we propose a powerful method to capture more important information from multiple loci in the family-based study while maintaining the robustness to the population stratification. In the simulation study, we examine the type-I error and compare the power with other FBAT tests under different scenarios. It has the correct type-I error rate when applied to data with population stratification and more powerful in our numerical study.

**160**

**Linkage and Association Studies Revealed a Locus for Obesity-related Quantitative Traits on Chromosome 1q43 in Caribbean Hispanics**

Chuanhui Dong (1), Ashley Beecham (2), Susan Slifer (2), Clinton B. Wright (1), Susan H. Blanton (2), Tatjana Rundek (1), Ralph L. Sacco (1)
(1) Deptartment of Neurology, University of Miami
(2) John P Hussman Institute for Human Genomics, University of Miami

While obesity is more prevalent in Hispanics than whites in the United States, little is known about the genetic etiology of the related traits in this population. We performed genome-wide linkage analyses in 1,390 subjects from 100 Caribbean Hispanic families on six obesity-related quantitative traits: body mass index, body weight, waist circumference, waist-to-hip ratio, abdominal and average triceps skinfold thickness, after adjusting for significant demographic and lifestyle factors. We then carried out a follow-up association analysis of top linkage peak in an independent community-based Hispanic subcohort ($N = 652$) from the Northern Manhattan Study. Linkage analyses identified one region on chromosome 1q43 showing the strongest evidence for body weight (maximal LOD = 2.69, $p = 0.0002$) and LOD scores $> 1.50$ for all investigated traits, except waist-to-hip ratio. We also detected four regions with LOD scores ? 2.00 for an obesity-related trait on chromosome 9q33–34, 14q32, 16p12, and 16q22–24. In the follow-up association analysis of 3,394 SNPs across 10 Mb within the flanking region of our linkage peak on chromosome 1q43, we further detected several SNPs and haplotypes in *CHRM3, RGS7, PLDS, KIF26B, SMYD3, EDARADD* and *CREM2* genes associated with obesity-related quantitative traits (p?0.001). Our results suggest that multiple genetic loci, particularly those on 1q43 region, may contribute to the variations in obesity-related quantitative traits in Caribbean Hispanics.

**161**

**Ordered Subset Analysis Identifies Loci Influencing Lung Cancer Risk on Chromosomes 6q and 12q**

Shenying Fang (1), Susan M. Pinney (2), Joan E. Bailey-Wilson (3), Mariza A. de Andrade (4), Yafang Li (1), Ming You (5), Ann G. Schwartz (6), Ping Yang (4), Marshall W. Anderson (2), Christopher I. Amos (1)
(1) Univeristy of Texas MD Anderson Cancer Center
(2) University of Cincinnati
(3) National Human Genome Research Institute
(4) Mayo Clinic College of Medicine
(5) Washington University
(6) Karmanos Cancer Institute

Genetic heterogeneity in the familial lung cancer linkage study has not been sufficiently evaluated with respect to family-specific susceptibility. 93 families were collected by the familial lung cancer recruitment sites of the Genetic Epidemiology of Lung Cancer Confsortium. We estimated linkage scores for each family by the Markov Chain Monte Carlo procedure using SimWalk2 software. We used ordered subset analysis (OSA) to identify genetically more homogenous families by ordering families based on a disease-associated covariate. Permutation testing was used to determine the significance of the linkage in optimal OSA subsets. A genome-wide screen for lung cancer loci identified strong evidence for linkage to 6q23–25 and nearly significant evidence for linkage to 12q24 using OSA, with peak LOD scores of 4.19 and 2.79, respectively. We found other chromosomes also suggestive for linkages, including 5q31–q33, 14q11, and 16q24. Analyses were also

conducted ranking the family level of risk for several other cancers and identified subsets of families with significantly increased LOD scores on chromosomes 1q23, 2p11, 6q23–25, 13p12, and 17p11 ($P < 0.05$). Our OSA results support 6q as a lung cancer susceptibility locus and provide nearly significant evidence for disease linkage on12q24. Validation studies using large sample size are needed to verify the presence of several other chromosomal regions suggestive of an increased risk for lung cancer and/or other cancers.

**162**

**The Statistical Equivalent of the Binary TDT for Quantitative Traits: Univariate And Multivariate Models**

Saurabh Ghosh (1), Tanushree Haldar (1)
(1) Human Genetics Unit, Indian Statistical Institute

The classical Transmission Disequilibrium Test (TDT) for binary traits circumvents the problem of population stratification as it tests for allelic association in the presence of linkage. Since clinical end-point traits are often defined by quantitative precursors, it may be a more prudent strategy to analyze the quantitative phenotypes without dichotomizing them into binary traits. Although some methods have been developed for testing transmission disequilibrium in the context of quantitative traits, these are not direct extensions of the classical TDT. We propose a simple logistic regression based test that can be analytically shown to be statistically equivalent to the TDT for binary traits, and hence is not susceptible to the presence of population stratification in the data. The proposed method can be easily extended to incorporate multivariate phenotypes. We perform Monte-Carlo simulations under a wide spectrum of genetic models and probability distributions of the quantitative trait values to evaluate the power of the proposed procedure and compare with the FBAT approach with identical data. We find that our method yields more power than FBAT if we suitably incorporate trios with both parents heterozygous in our likelihood as well as for multivariate phenotypes based on principal components. We apply our method to analyze externalizing symptoms, an alcoholism related endophenotype from the Collaborative Study on the Genetics Of Alcoholism (COGA) project.

**163**

**Tests of Association for Family Data: Tiled Regression with Generalized Estimation Equations**

Yoonhee Kim (1), Cristina Justice (1), Heejong Sung (1), Juanliang Cai (1), Alexa J.M. Sorant (1), Dana Behneman (1), Mera Krishnan (1), Nancy H. Miller (2), Alexander F. Wilson (1)
(1) Genometrics Section, IDRB, NHGRI, National Institutes of Health(2) The Children's Hospital, University of Colorado

Tiled regression is the use of stepwise and multiple regression methods in predefined segments of the genome, defined by hotspot blocks, to identify independent genetic variants responsible for the variation or susceptibility in quantitative and qualitative traits, respectively. Multiple and stepwise regression methods are used to test for associations on the sequence variants in each tile to select the independent markers within tile. Higher order stepwise regressions are then used to identify significant variant across tiles, chromosomes and the entire genome. In this study, the tiled regression framework is extended to allow for family data. Generalized estimating equations (GEE) were incorporated to the tiled regression framework in order to account for familial correlations within families. This approach is illustrated with familial idiopathic scoliosis (FIS) data [Miller et al., 2005]. The method classified 1732 SNPs in candidate regions located on chromosomes 1, 8, and 9 into 622 tiles. The data were comprised of 704 sibships from 187 families. The final linear model from these three candidate regions to predict the FIS curvature is following: Max Curvature = −4.03∗rs2929460 (Chr8)+3.61∗rs10100733 (Chr8)+3.41∗rs7870176 (chr9)+ 4.90∗rs7038954 (Chr9).

**164**

**Shared Genomic Segment Analysis: The Power to Succeed Where GWAS Fails**

Stacey Knight (1), Ryan P. Abo (1), Haley J. Abel (1), Alun Thomas (1), Nicola J. Camp (1)
(1) University of Utah

Shared genomic segment (SGS) analysis is a method that uses dense genotyping in extended high-risk pedigrees to identify regions of sharing between distantly related cases. Here, we illustrate the power of this approach to identify dominant rare risk variants. We simulated extended high-risk pedigrees and high-density genotype data. Twelve genetic models were considered, based on disease prevalence (1% or 0.5%), minor allele frequency (MAF = 0.005, 0.0005, or 0.00005), and dominant penetrance (0.2 or 0.5). Pedigrees were categorized by their significant excess of disease ($P < 0.01$, $< 0.001$ and $< 0.0001$), and had at least 15 total meioses between all cases. Across these scenarios, the power for a single pedigree ranged widely. Nonetheless, 10 or fewer pedigrees was sufficient to gain excellent overall power ($> 80\%$) for the majority (70%) of the models considered. Power increased with increases in penetrance, disease MAF, and the excess of disease in the pedigree. Power for a single pedigree was best for models with higher penetrance and disease MAF (40.7%–66.1%), or the highest risk pedigrees and higher penetrance (31.4%–86.9%); conditions under which few pedigrees ($< 6$) would be sufficient for excellent overall power. For all scenarios genomewide association studies (GWAS) would have negligible power ($N = 30{,}000$; power = 10%). In conclusion, SGS analysis is a powerful method for detecting dominant rarer risk variants and offers a valuable complement to GWAS for gene-mapping.

**165**

**Comparison of Family Based Methods For Genome-Wide Association Study In the Framingham Eye Study**

Sean Lacey (1), Jacki Buros (1), Kathryn Lunetta (1), Adrienne Cupples (1), Lindsay A. Farrer (1), Gyungah Jun (1)
(1) Boston University

In genome-wide association studies (GWAS) of family data, there are several options available to test association. However, the relative merits of these methods have not been compared for rare variants with minor allele frequency (MAF) between 0.01 and 0.1 under different genetic models. We investigated three different family-based methods for GWAS: the two-level Haseman-Elston

method, generalized estimating equations (GEE), and the generalized disequilibrium test (GDT). We conducted association analyses for each method with a quantitative trait of pupil size in the Framingham Eye Study. After cleaning for quality control including call rate >98%, and Hardy-Weinberg $P > 10 \times 10^{-6}$, a sample of 1499 individuals in 588 families and about 346 K SNPs (Affy 550 k) were available for analysis. We detected multiple genome-wide significant SNPs with $P < 5 \times 10^{-8}$ with low MAF (<0.01) under the additive genetic model using the GEE method but not using the RELPAL and GDT methods. Inflation factor using the GEE method was slightly increased with inclusion of rare SNPs; but in general, there was no significant inflation.

**166-WITHDRAWN**
**Linkage Study of Prostate Cancer in High-risk African American Families from Louisiana.**
Elisa M. Ledet (1), Joan E. Bailey-Wilson (2), Diptasri M. Mandal (1)
(1) Department of Genetics, Louisiana State University Health Science Center- New Orleans, LA
(2) NHGRI, NIH, Baltimore, MD

Prostate cancer is a complex multi-allelic disease and the most common malignancy in men throughout the world. In the U.S., a lifetime risk of mortality from prostate cancer is 3% for white men and 4% for African-American men. Previously, we performed a genome-wide linkage scan on three families using microsatellite markers and identified a region on chromosome 22q13. We have since continued recruitment and accrued 20 large high-risk African-American families with at least 3 affected individuals; 28 large high-risk Caucasian families have also been recruited. An Infinium II HumanLinkage-12 panel (Illumina, Inc.) with 6,090 SNP markers was performed on 180 DNA samples, including 15 African American families and 4 Caucasian families. This panel is optimized for linkage detection and SNPs are distributed on every chromosome with an average gap of 441 Kb and 0.58 cM. Three samples were discarded from analysis due to poor performance and a total of 177 samples imported into BeadStudio version 3.3.7. Quality control and pruning of released SNPs was performed using PLINK. Linkage analysis is ongoing on the African-American cohort with Merlin and Genehunter-Plus. We intend to identify any markers associated with prostate cancer in high-risk African-American families and document any correlation between clinical features, such as prostate specific antigen (PSA), "age of onset" and/or Gleason score.

**167**
**Efficient Haplotype Reconstruction for Pedigree Data with Zero Individual Genotype Mismatches**
Qing Li (1), M. Daniele Fallin (2), Joan E. Bailey-Wilson (1)
(1) Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, USA
(2) Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, USA

Haplotype phase imputation is useful for most haplotype based association tests and for missing genotype imputation. However, phase imputation in extended pedigrees is not trivial. For large haplotype blocks, the number of possible diplotype configurations grows exponentially with the number of markers in the block. There are several haplotype phase estimation or reconstruction algorithms, and a few recent methods handle extended pedigree data. However, haplotypes with rare frequencies are often left out in reconstruction. In application, this can result in misclassification of imputed genotypes. These erroneous genotype assignments are considered particular to individuals, and are usually ignored. Based on population haplotype frequencies, we developed an efficient algorithm to impute phases (and therefore genotypes) for haplotype blocks of up to 8 SNPs, in trios as well as in extended pedigrees. To ensure no individual genotype assignment errors, we consider all possible haplotypes, even those with extremely rare frequencies. We evaluated our phase imputation method using simulated data with masked genotypes within pedigrees. We examined the misclassification proportions of imputed genotypes when rare haplotypes are left out, and compared our methods with other methods: PHASE, HAPLORE, PedPhase, and PhyloPed.

**168**
**Generalized Method in Candidate Gene Analysis for Nonnormally Distributed Quantitative Traits using Family Trios**
Jingky Lozano (1)
(1) Georg-August-Universitaet Goettingen, Germany

The occurrence of nonnormally distributed quantitative traits in candidate gene analysis poses difficulties for statistical methods that are sensitive to distributional assumptions. In family-based designs, several approaches exist to deal with nonnormally distributed traits. In this study, a *Generalized Quantitative Transmission Disequilibrium Test* (GQTDT) for detecting genetic main effects and epistasis using quantitative traits and family trios is proposed. The method is based on the *Generalized Additive Model for Location, Scale and Shape* (GAMLSS) (Rigby and Stasinopoulus, Appl Stat, 2005) which allows not only the mean but all parameters of the conditional distribution of the quantitative trait to be included in the model as parametric and/or additive smooth nonparametric functions of the explanatory variables and/or random effect terms. To determine the power and Type I error of the method in detecting epistatic and genetic main effects, family trios with genotypes of two unlinked biallelic loci were simulated. Quantitative traits influenced by one or two loci and epistasis were created. The traits were simulated either as normally distributed or skewed to the right. Using the appropriate distribution or link, the GQTDT results show reasonable Type I errors, good increase in power in detecting genetic main effects and slightly increased power in detecting epistasis compared to traditional regression method.

**169**
**Contribution of Family Studies in the Era of Genome Wide Association and Sequencing Studies**
Kathleen Ries Merikangas (1), Kung-Yee Liang (2), Robert C. Elston (3)
(1) National Institute of Mental Health
(2) Johns Hopkins Bloomberg School of Public Health
(3) Case Western Reserve University

The successful identification of genes underlying complex disorders using the genome wide association (GWA) approach in large samples of cases and controls has led to the identification of genes for several complex disorders. This approach has been particularly successful for diseases that have valid phenotypic definitions, a priori evidence for familial recurrence, and some knowledge of underlying biologic pathways. However, progress in identifying and replicating genes for some complex disorders, particularly neuropsychiatric disorders, has been less successful. There are several limitations of the GWA approach including small effects of variants on disease risk, single nucleotide polymorphisms may be remote from the actual genes underlying complex disorders, and confounding due to population stratification. This presentation will describe the value of family studies in obtaining indirect evidence for genetic transmission in terms of risk estimation, elucidation of genetic architecture, and dissection of complex phenotypes. Specific issues in selection of study designs and analyses for detecting genes will be described. We conclude that concepts of population genetics will be even more critical with advances in molecular genetics, and that family and pedigree studies will play a major role in development of disease risk profiles by integrating emerging findings on genetic and environmental risk factors.

**170**

**A New Framework for Structural Equation Models (SEM) in Family Data**

Nathan Morris (1), Robert Elston (1), Cathy Stein (1)
(1) Case Western Reserve

Most diseases of interest to modern genetic epidemiologists are complex both in their etiology and measurement. That is, they result from a complicated interplay of various environmental and genetic factors, and they are subject to fuzzy, noisy and often multidimensional disease definitions. Such traits call for a modeling approach that accounts for both the causal relationships between variables and the errors associated with the measurement of these variables. One such modeling paradigm is known as Structural Equation Modeling (SEM). The classical formulation of SEM involves two sets of equations: the measurement equations which account for measurement error and the structural equations which account for causal relationships among the variables. In this work we show how both these equations may be used in general pedigree structures. Furthermore, our work allows the statistical modeler to separate the process of modeling familial correlation from the process of developing an SEM. This separation is facilitated by the use of Kronecker notation. Our framework may be used to compare causal models in family data without genetic marker data. It also allows for a nearly endless array of genetic association and/or linkage tests. We will briefly present a robust limited information method of model fitting. We will also show some simulation results to demonstrate the statistical properties of our approach to fitting. We are currently implementing our framework in an R package.

**171**

**Quantitative Trait Loci for Brain Natriuretic Peptides Concentration among African Americans at the Jackson Heart Study**

Solomon K. Musani (1), Sarah G. Buxbaum (2), Ramachandran S. Vasan (3), Aurelian Bidulescu (4), Herman A. Taylor (5), Ervin R. Fox (1)
(1) University of Mississippi
(2) Jackson State University
(3) Boston University
(4) Morehouse University
(5) University of Mississippi; Jackson State University

1,300 Jackson Heart Study participants (mean age 55?12 years, 63% women) underwent both routine echocardiography and testing for serum Brain Natriuretic Peptide concentration (BNP), a biomarker correlated with severity of heart failure. Sex specific multivariable models and Haseman-Elston regression were used to estimate heritability. We also performed multipoint marker linkage analysis using data from 374 autosomal microsatellite markers typed in 1,300 participants belonging to 264 families. Age and echocardiographic variables accounted for 45% and 33% variation in log BNP concentration in men and women, respectively. After multivariable adjustment of log BNP, $h^2$ was 24%. Multipoint linkage analysis revealed regions of significant linkage to log BNP on chromosome 5q15–q21 (LOD score 3.73), and suggestive linkage on chromosomes 7q36 (LOD score 2.22) and distal 16p (LOD score 2.78). The significant QTL on chromosome 5q15–q21 is located near *Proprotein convertase subtilin/kexin type 1* (*PCSK1*) gene that has been reported to be associated with body mass index. A SNP at the 5′ end of *PCSK1* at 5q15 was significantly associated with log BNP with additive effect, nominal $P < 0.0001$ in men, and $p = 0.002$ after further adjustment for left ventricular size.

**172**

**A Major Locus Predisposing to HHV-8 Infection in Children Maps to Chromosome 3p22 in an African Population.**

Vincent Pedergnana (1), Antoine Gessain (2), Patricia Tortevoye (2), Delphine Bacq-Daian (2), Anne Boland (2), Laurent Abel (1), Sabine Plancoulaine (1)
(1) Laboratoire de Genetique Humaine des Maladies infectieuses, Institut National de la Sante et de la Recherche Medicale U980
(2) Unite d'Epidemiologie et Physiopathologie des Virus Oncogenes, Institut Pasteur

Infection by human herpesvirus-8 (HHV-8), the aetiological agent of Kaposi's sarcoma, was shown to exhibit strong familial aggregation in endemic countries. To further investigate the role of host genetic factors, we studied 40 families (608 subjects; 1–88 yrs old) living in an isolated area of Cameroon. HHV-8 infection status (HHV-8+/HHV-8-) was determined by specific serological assays. Global HHV-8 seroprevalence was 60%, raising from 32% under 10 years up to a plateau around 60% after 15 years. We first performed a segregation analysis that provided strong evidence for a recessive gene controlling predisposition to HHV-8 infection. This gene is predicted to have a major effect during childhood, with almost all homozygous predisposed subjects (~7% of the population) being infected by age 10, a genetic model similar to our previous analysis conducted in another population of African origin with a lower HHV-8 seroprevalence (12%).* We then conducted a model-based linkage analysis using the 20 most informative families (i.e. with at least one HHV-8+

child <10 yrs old). The 289 subjects of these families were genotyped using the Illumina linkage IVb panel (6089 SNPs), and a single region of chromosome 3p22 was significantly linked to HHV-8 infection (LOD = 3.81). Our study provides the first evidence that HHV-8 infection in children from endemic areas has a strong genetic basis that involves at least one major recessive locus.
*Plancoulaine et al., JID, 2003;187:1944–1950.

**173**
**Sample Size Calculations for the Transmission Disequilibrium Test (TDT) in the Presence of Population Stratification**
Ronnie A. Sebro (1)
(1) Institute for Human Genetics, University of California, San Francisco

The Transmission Disequilibrium Test (TDT) is a test for association in the presence of linkage, which compares the rate of transmission of each allele from a heterozygous parent to an affected offspring. The TDT has gained popularity because it preserves the Type I error rate in the presence of population stratification. However, population stratification results in a decrease in the number of heterozygous parents compared to that calculated assuming Hardy-Weinberg Equilibrium (HWE). Current sample size calculations for the TDT assume HWE, however no guidelines for sample size calculations exist for the TDT in the presence of population stratification. Sample size calculations for the TDT depend on the genetic mode of inheritance (MOI) and the distribution of the informative mating types for the TDT. Both parents cannot be considered independent in the presence of population stratification, resulting in a change in the relative proportion of the informative mating types. These findings result in an altered mating type distribution compared to that expected assuming HWE. We show how the distribution of the informative mating types can be calculated in the presence of population stratification, and use the method described by Knapp (1999) for TDT sample size calculations, to calculate the true sample sizes. The true sample sizes required to achieve a pre-specified level of power tend to be larger, but can be smaller than sample sizes calculated assuming HWE.

**174**
**Genetic Heterogeneity of Prostate Cancer Susceptibility in Finland: Evidence for Several Novel Loci and Replication of HPCX1 and HPC10 Loci**
Claire L. Simpson (1), Cheryl D. Cropp (1), Tiina Wahlfors (2), Asha George (3), Nati Ha (2), Teuvo L.J. 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
(3) National Human Genome Research Institute, National Institutes of Health and Fox Chase Cancer Center
(4) Department of Urology, Tampere University Hospital, University of Tampere

Prostate cancer is the most common cancer in US men. We reported a linkage scan in 69 Finnish Hereditary Prostate Cancer (HPC) families that replicated the HPC9 locus on 17q21–q22 and identified a locus on 2q37. Here we used ordered subset analysis, conditioned on non-parametric linkage to these loci to detect other loci linked to HPC in subsets of families. OSA using parametric LOD scores is ongoing. Significant linkage to a 5-cM interval with a peak OSA nonparametric allele-sharing LOD score of 4.876 on Xq26-q27 (?LOD = 3.19, empirical $p = 0.009$ for ?LOD) was observed in a subset of 41 families weakly linked to 2q37, overlapping the HPCX1 locus. Other linked loci were 12q21–q23 in 17 families unlinked to 2q37, and 8q24 in 15 families weakly linked to 2q37, overlapping the HPC10 locus. Significant linkage to Xq25 was observed in a subset of 41 families most strongly linked to 17, with a peak OSA LOD score of 3.54 (?LOD = 1.48, $p = 0.04$), 10cM away from the peak found by conditioning for linkage to 2q37. Other strongly linked loci were 3q26–q27 and 12q14–q21 in families unlinked to 17. We conditioned on maximum family NPL scores for 2q37 and 17q21–q22. Two peaks (12q21 and 6q15) were also found by conditioning on just one of the loci, but 2 others were novel: 18q12–q12 and 22q11.1–q11.21, which is close to HPC6. Using OSA allows us to find additional loci linked to HPC in subsets of families, and underlines the complex genetic heterogeneity of HPC even in highly aggregated families.

**175**
**Effects of Measured Susceptibility Genes on Cancer Risk in Family Studies**
Chih-Chieh Wu (1), Louise C. Strong (2), Sanjay Shete (1)
(1) Dept Epidemiology, M. D. Anderson Cancer Center
(2) Dept Genetics, M. D. Anderson Cancer Center
Numerous family studies have been performed to assess the associations between cancer incidence and genetic and non-genetic risk factors and to quantitatively evaluate the cancer risk attributable to these factors. However, mathematical models that account for a measured hereditary susceptibility gene have not been fully explored in family studies. In this report, we proposed statistical approaches to precisely model a measured susceptibility gene fitted to family data and simultaneously determine the combined effects of individual risk factors and their interactions. Our approaches are structured for age-specific risk models based on Cox proportional hazards regression methods. They are useful for analyses of families and extended pedigrees in which measured risk genotypes are segregated within the family and are robust even when the genotypes are available only in some members of a family. We exemplified these methods by analyzing 6 extended pedigrees ascertained through soft-tissue sarcoma patients with p53 germ-line mutations. Our analyses showed that germ-line p53 mutations and sex had significant interaction effects on cancer risk. Our proposed methods in family studies are accurate and robust for assessing age-specific cancer risk attributable to a measured hereditary susceptibility gene, providing valuable inferences for genetic counseling and clinical management.

**176**
**Evaluation of Effects of Genetic polymorphisms and Age Trajectories of Physiological Indices on survival**
Konstantin G. Arbeev (1), Svetlana V. Ukraintseva (1), Liubov S. Arbeeva (1), Igor Akushevich (1), Alexander M. Kulminski (1), Deqing Wu (1), Anatoliy I. Yashin (1)
(1) Duke University

We evaluated the effects of genetic polymorphisms and age trajectories of different physiological indices on survival using the extended version of the stochastic process model of aging aimed at analyses of genetic and non-genetic subsamples of longitudinal data. We applied the model to data on mortality and longitudinal measurements of physiological indices collected for participants of the original cohort of the Framingham Heart Study (26 biennial exams) and data on the APOE (carriers vs. non-carriers of the e4 allele) and ACE D/I polymorphisms collected for a subsample of such individuals. We estimated different aging-related characteristics, such as aging-related decline in stress resistance (associated with the narrowing of the U-shape of mortality risk as a function of a physiological index) and adaptive capacity, mean "allostatic" states (the trajectories of the indices that organisms are forced to follow by the process of allostatic adaptation), and physiological "norms" (the age-specific values of indices minimizing mortality risk), for carriers of different alleles/genotypes. The results show that such characteristics differ for carriers of alleles (genotypes) of both sexes. This indicates the possibility of genetic effects on respective aging-related mechanisms and such differences may contribute to the observed patterns of allele-(genotype-) specific mortality rates for both sexes.

**177**

**Association of Candidate SNPs with Melanoma Susceptibility in Australian Adults**

Gemma Cadby (1), Sarah V. Ward (1), Judith M. Cole (2), Michael Millward (3), Lyle J. Palmer (1)
(1) Centre for Genetic Epidemiology and Biostatistics, University of Western Australia
(2) Western Australian Melanoma Advisory Service
(3) School of Medicine and Pharmacology, University of Western Australia

Melanoma is the most aggressive form of skin cancer and also one of the most preventable cancers. Genetic epidemiological research has led to some understanding of the major environmental and genetic risk factors for melanoma, however no studies of genetic variants and melanoma risk to date have been conducted in Western Australian adults. A subset of 800 Caucasian subjects from the Western Australian Melanoma Health Study, and 1366 controls from the Busselton Health Study, were genotyped at 18 SNPs. These SNPs were chosen as they had been identified as melanoma-risk SNPs in earlier studies. Logistic regression analysis, adjusted for age, sex, BMI and age:sex interactions, identified 9 SNPs (modelled codominantly) which were significantly associated with melanoma susceptibility. After adjustment for multiple testing using the FDR, four SNPs remained significantly associated with melanoma. These were rs258322 ($q = 1 \times 10^{-4}$) in MC1R, rs1393350 ($q = 1 \times 10^{-2}$) in TYR, rs12203592 ($q = 1 \times 10^{-5}$) in IRF4, and rs1011970 ($q = 4 \times 10^{-2}$) in MTAP. The MC1R, TYR and IRF4 genes have all been associated with pigmentation traits, while MTAP has been associated with naevi counts. This study has replicated associations between pigmentation and naevi genetic variants and melanoma susceptibility in a sample of Western Australian adults.

**178**

**Male-Specific Risk Alleles of Autism Spectrum Disorders in a Genome-wide Association Analysis**

Shun-Chiao Chang (1), David L. Pauls (2), Christoph Lange (1), Jessica Lasky-Su (3), Mark Daly (2), Susan L. Santangelo (2)
(1) Harvard School of Public Health
(2) Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School
(3) Channing Laboratories, Brigham and Women's Hospital and Harvard Medical School

Autism spectrum disorders (ASD) are highly heritable but heterogeneous. For reasons not yet understood, ASD are four times more common in males than in females. Previous studies of ASD employing a sex-spliting strategy have identified genetic variants specific to families in which all affecteds are male (male-only; MO), suggesting a sex-specific genetic heterogeneity in ASD. We evaluated the genomewide sex-specific genetic architecture of ASD in 674 families (2,423 subjects) of European ancestry from the Autism Genetic Resource Exchange (AGRE) in this study. 60% of the multiplex European families were classified as MO while 40% contained at least one affected female (female-containing; FC). A total of 341,849 SNPs passing quality control were analyzed. We did not observe genomewide significant association to ASD in the full sample analysis. However, in the MO family set analysis, we identified a novel genomewide significant association with ASD in introns of the *XG* gene in the pseudoautosomal boundary on Xp22.3 and Yp11.31 (PAB1) based on Bonferroni correction, of predominantly paternal origin. In addition, 6 markers that reside within a 550-kb intergenic region on 13q33.3 between the *MYO16* and *IRS2* genes showed suggestive association with ASD in the MO families. None of these markers appeared to be associated with ASD in the FC families (all $P > 0.05$). Our results suggest that the PAB1 and an intergenic region on 13q33.3 may harbor male-specific genetic variants for ASD.

**179**

**Uncovering Genetic Modifiers of the Cardiac Phenotype in 22q11.2 Deletion Syndrome**

Stephanie L. Ciosek (1), Tracy Busse (2), Elizabeth Goldmuntz (3), Donna McDonald-McGinn (2), Elaine Zackai (3), Beverly S. Emanuel (3), Deborah A. Driscoll (1), Sulagna C. Saitta (3), Marcella Devoto (3)
(1) University of Pennsylvania School of Medicine
(2) The Children's Hospital of Philadelphia
(3) The Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine

The 22q11.2 deletion syndrome (22qDS) affects approximately 1 in 4000 live births and is the most frequently occurring microdeletion syndrome. Observed phenotypes are highly variable with 74% showing cardiac defects and 69% showing palate abnormalities; immune dysfunction, hypocalcaemia and learning difficulties are also present. It is estimated that 10% of patients inherit the deletion from an affected parent. Within families the phenotype is also highly variable. Due to the great inter- and intra-family variability of the phenotype, it is thought that modifiers of the phenotype must exist, either genetic, environmental or both. Here we report the results of a candidate gene analysis for modifiers of the cardiac phenotype in familial 22qDS cases. A total of 75 members of 29 22qDS families were evaluated for cardiac abnormalities and genotyped on the Affymetrix 6.0 SNP chip. Candidate gene lists were

assembled from literature reviews and the results of expression array analysis. We used the software program LAMP because of its ability to handle large and more complex pedigree structures. Comparison of 22qDS patients with a normal cardiac phenotype to those with any cardiac malformation yielded statistically significant association with markers in *ARHGAP24*, *TNFRSF11B* and *RXRa* after correction for the number of independent SNPs tested ($p$-values $<2.3 \times 10^{-5}$). These results suggest that variants of these genes may increase the risk of cardiac defects in 22qDS patients.

**180**

**Genome-wide Linkage and Association Reveals that the Tissue Factor (F3) Locus on Chromosome 1p22-p21 Likely Influences Quantitative Variation in D-dimer Levels**

Vincent P. Diego (1), Eugene Drigalenko (1), Melanie A. Carless (1), Joanne E. Curran (1), Thomas D. Dyer (1), Laura Almasy (1), David L. Rainwater (1), Michael C. Mahaney (1), Anthony G. Comuzzie (1), Shelley A. Cole (1), Russell P. Tracy (2), Jean W. MacCluer (1), Eric K. Moses (1), Harald H.H. Goring (1), John Blangero (1)
(1) Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio TX 78227
(2) Department of Pathology, University of Vermont, Colchester, VT 05446

D-dimer results from fibrin formation via coagulation and subsequent fibrinolysis, and is associated with increased risk of a cardiovascular disease (CVD). Studies on the genetics of D-dimer have found this trait to be significantly heritable. To further these results, we performed a genome-wide (joint) linkage and association study (GWLAS) of plasma D-dimer levels on 811 individuals from 35 extended families of Mexican Americans of San Antonio, Texas participating in the San Antonio Family Heart Study. Plasma D-dimer levels were transformed to normality by way of the inverse Gaussian transformation applied to residuals after accounting for age, sex, and their interactions as covariates. Using a maximum likelihood variance components model, we found that these normalized plasma D-dimer levels had a significant heritability of 0.17 ($p = 1.6 \times 10^{?4}$). We then performed a GWLAS on 542,994 single nucleotide polymorphisms (SNPs) using a joint linkage and measured genotype association approach. We found evidence of significant association with a single SNP, namely rs2022309 ($p = 1.3 \times 10^{?10}$) near the tissue factor gene (*TF, F3*) located at chromosome 1p22-p21. If found to be functional, since TF is a major control point for coagulation, this result would emphasize the role of coagulation in D-dimer formation and contribute to our understanding of the regulation of this important risk factor for CVD.

**181**

**Generalization of GWAS-based Genetic Effects for Diastolic Blood Pressure in American Indians: The Strong Heart Family Study**

Nora Franceschini (1), Kari E. North (2), V.S. Voruganti (3), Sandra Laston (3), Karin Haack (3), Elisa T. Lee (4), Lyle G. Best (5), Jean W. MacCluer (3), Jason G. Umans (6), Thomas D. Dyer (3), Shelley A. Cole (3)
(1) Epidemiology, University of North Carolina
(2) Epidemiology, University of North Carolina and Carolina Center for Genome Sciences
(3) Genetics, Southwest Foundation for Biomedical Research
(4) University of Oklahoma Health Sciences Center
(5) Missouri Breaks Industries Research Inc
(6) Medstar Research Institute

Genome wide association (GWA) studies have been successful at mapping loci that influence blood pressure (BP), primarily in European and European American populations. Translation of these findings requires an exploration of these associations in different ancestral groups. The purpose of the current study was to identify genetic loci influencing BP traits in an understudied minority population, American Indians. We selected 41 GWA single nucleotide polymorphisms (SNPs) in 23 loci in 3807 Strong Heart Family Study American Indian participants from three recruiting centers: Arizona, Dakotas and Oklahoma. Center-specific diastolic BP residuals were obtained from linear regression models adjusted for age, sex, age$^2$ and age-by-sex interaction. Residuals were regressed onto SNP dosage using variance component models to account for family relatedness and population history. Summary estimates across centers were combined using a weighted average of point estimates meta-analyses (fixed effects). Seven SNPs were associated with BP (alpha = 0.05): rs12046278 nearby *CASZ1* ($p = 0.03$), rs6749447 in *STK39* ($p = 0.05$), rs2509458 nearby *SPACA1* ($p = 0.001$), rs11014166 in *CACNB2* ($p = 0.03$), rs3184504 in *SH2B3* ($p = 0.04$), rs653178 in *ATXN2* ($p = 0.04$) and rs1378942 in CSK ($p = 0.03$). Overall findings suggest some similarities in the genetic architecture underlying BP across European and American Indian populations.

**182**

**The Consistent Association of Helicobacter Pylori seropositivity with Le Polymorphism and the Inconsistent Association with Se Polymorphism through Our Studies**

Yasuyuki Goto (1,2), Nobuyuki Hamajima (1)
(1) Department of Preventive Medicine, Nagoya University Graduate School of Medicine, Japan
(2) Depertment of Internal Medicine and Gastroenterology, Inuyama Chuo Hospital

*Helicobacter pylori* (*H. pylori*) attach to gastric mucosa with the blood group antigen-binding adhesion, which binds to Lewis b or H type 1 blood group carbohydrate structures. In our first study, *Se* gene polymorphism reduced *H. pylori* infection risk and *Le* polymorphism tended to increase the risk. However, in our second study, we reported that *Se* gene polymorphism was related to the risk in the opposite direction and *Le* showed the consistent positive relation to the risk. In this study, we scrutinized whether *Se* and *Le* gene polymorphisms were related to *H. pylori* infection risk. The study subjects were consisted of 505 healthy adults, who attended a health check-up examine at Yakumo town in Hokkaido. *le* allele and *se* allele are considered to have low/no enzyme activity. Those with *le* allele showed the consistent tendency to increase the risk of *H. pylori* infection; OR for *Lele* = 1.37, 95% CI:0.93–2.02 and OR for *lele* = 1.42, 95%CI:0.68–2.96. However, those with *se* allele in this study were associated with *H. pylori* infection in the opposed direction to in our first report without significance; OR for *Sese* = 1.03, 95%CI:0.68–1.57 and OR for *sese* = 1.56, 95%CI:0.91–2.68. We could not deny the possi-

bility that *Se* polymorphism was less likely to be associated with the risk of *H. pylori* infection. *Le* polymorphism could be more important in the establishment of *H. pylori* infection.

**183**

**ACTG DACS 250: Mitochondrial Genetic Variation Is Associated with CD4 T Cell Recovery in HIV-infected Persons Initiating Antiretroviral Therapy**

Benjamin J. Grady (1), David C. Samuels (1), Gregory K. Robbins (2), Doug Selph (1), Jeffrey A. Canter (1), Richard B. Pollard (3), David W. Haas (1), Robert Schafer (4), Spyros Kalams (1), Debbie Murdock (1), Marylyn D. Ritchie (1), Todd Hulgan (1)
(1) Vanderbilt University
(2) Harvard University
(3) University of California Davis
(4) Stanford University

Mitochondrial genetic variation has been associated with aspects of HIV infection, ranging from time to progression to AIDS to adverse effects from antiretroviral therapy (ART). In this study, we obtained full mitochondrial DNA sequence data from U.S.-based adult participants in the AIDS Clinical Trials Group (ACTG) study 384 to examine links between mitochondrial DNA variants and the CD4 T cell recovery with ART. Variation in the mitochondrial genome could influence apoptotic activity in cells with certain variants which might reduce CD4 T cell recovery during ART. In ACTG 384, data on CD4 T cell count was available at baseline, and 48 and 96 weeks after ART initiation on 423 participants, 30% of which self-identified as non-Hispanic black (black). The primary outcome was CD4 T cell count as an absolute change from baseline at 48 weeks dichotomized at a 100 cell increase. A race-stratified analysis of variants with a frequency greater than 1% showed several variants associated ($P < 0.05$ before Bonferroni correction) with increased and decreased magnitude of CD4 cell recovery. The most significantly associated mitochondrial DNA variant was found in the stratum of (self-identified) black and indicated a primary association between the African L2 subhaplogroup and decreased CD4 cell recovery ($p = 0.002$, Odds Ratio = 0.022). In addition to variants tagging the L2 subhaplogroup, other polymorphisms not associated with haplogroup status warrant follow up in a replication set.

**184**

**A Mixture Model with Random Effects Components for Classifying Families with Application to Healthy Aging.**

Jeanine Houwing-Duistermaat (1), Marian Beekman (1), Rudi Westendorp (1), Eline Slagboom (1), Francesca Martella (2)
(1) LUMC
(2) Sapienza University Rome

Model based clustering is a tool for joint analysis of multiple and diverse parameters. It identifies clusters of subjects with similar profiles for these parameters. To be able to use these tools for family data, we extended current methods by adding random family effects in the cluster specific regression model. The models are fitted by an EM algorithm with the random effects and cluster membership as unobserved data. The new methods were used to classify nonagenarian sibling pairs of the Leiden Longevity Study (LLS) ($n = 420$) according to their health status. The LLS is a multi generational family study showing excess survival. Healthy was defined as having beneficial values for established health parameters such as Mini Mental State Examination. Two out of seven considered health parameters appeared to discriminate pairs. The obtained percentages of concordant healthy, unhealthy, and discordant pairs were 63%, 28%, and 9% respectively. To verify whether the results were realistic, we analyzed parental ages at death and found that higher probabilities for healthy cluster membership were associated with parents who reached older ages ($p = 0.06$). Using simulations under a similar setting, we obtained high values of agreement (0.98). To conclude we successfully derived model based clustering methods for family data and used them to classify sibling pairs according to health status. The obtained cluster membership probabilities are used in gene mapping studies for healthy aging

**185**

**Evaluation of Muscle Signatures for Aging and HIV Infection**

Rebecca Kusko (1), Paola Sebastiani (2), Monty Montano (1)
(1) Boston University School of Medicine
(2) Boston University School of Public Health

The identification of genetic signatures that underlie age-associated muscle wasting conditions would be valuable in the development of therapeutics and in understanding muscle ageing. In our preliminary analysis of two distinct gene expression profile datasets from healthy young and healthy older subjects, we were able to identify a genetic signature that distinguished young from old. BADGE (Bayesian Analysis of Differential Gene Expression) was used to find the most significant set of expressed genes in *vastus lateralis* samples of younger (19–29 year old) and older (65–85 year old) males. From this analysis, we identified a ten-gene signature. To evaluate whether this signature was informative for accelerated aging in HIV infection associated wasting, *vastus lateralis* samples from mid-age male patients with HIV infection (30–55 years old, average = 43 years old), were evaluated and observed to cluster within the older subject group rather than the null hypothesis of randomly segregating between young and old, supporting the view that HIV infection resembles an accelerated biological aging phenotype. To assess the specificity of this signature we did a comparison of this muscle aging signature with multiple muscle disease datasets. The ten gene musculoskeletal ageing signature discovered in this study may be useful as a tool to help elucidate the relationship between HIV and musculoskeletal aging phenomena.

**186**

**Genetic Analysis of Neuroblastoma in African American Patients**

Valeria Latorre (1), Sharon Diskin (2), Maura Diamond (2), Haitao Zhang (3), Hakon Hakonarson (3), John Maris (2), Marcella Devoto (1)

(1) Division of Human Genetics, The Children's Hospital of Philadelphia
(2) Division of Oncology, The Children's Hospital of Philadelphia
(3) Center for Applied Genomics, The Children's Hospital of Philadelphia

Genome wide association studies in neuroblastoma have revealed several associated loci, including *FLJ22536/FLJ44180* in 6p22, *BARD1* in 2q35, *LMO1* in 11p15, and a CNV in 1q21. These studies have been carried out in patients of European descent, which comprise most neuroblastoma cases. Extension to other ethnic groups is important to confirm their validity beyond the Caucasian population. The different LD pattern may also help to better define the limits of the association signals and identify causal variants. We have collected a group of 326 self-reported African-American neuroblastoma patients and genotyped them using the Illumina HumanHap 550 K and 610Quad. A group of 2500 African-American unaffected children was used as controls. Standard QC procedures were applied to remove samples with low call rates, and SNPs with low call rates, deviation from HWE in controls, and MAF less than 1%. Analysis was performed by Cochran-Mantel-Haenszel test after multi-dimensional scale clustering to account for population substructure. Among the known loci, the one most consistently replicated was *BARD1*, with six SNPs with $P < 0.05$ (smallest $p = 0.0004$ for rs7587476). Two SNPs in the *FLJ22536/FLJ44180* region had nominally significant *p*-value (smallest $p = 0.02$ for rs1928174), and one in *LMO1* ($p = 0.02$ for rs4237769). These results indicate that the same genetic factors affect risk of neuroblastoma in both populations.

**187**
**Association of Y402H CFH Polymorphism Status and Mortality**

Kristine E. Lee (1), Ronald Klein (1), Barbara E.K. Klein (1), Sudha K. Iyengar (2), Theru A. Sivakumaran (2)
(1) University of Wisconsin
(2) Case Western Reserve University

The Complement Factor H (CFH) gene is associated with age-related macular degeneration (AMD) and may be associated with cardiovascular disease. We will explore the association of the Y402H polymorphism of the CFH gene to mortality in the Beaver Dam Eye Study. The study was designed to look at risk factors for age-related eye diseases including AMD in a population of men and women between the ages of 43 and 86 years and has cause of death information for up to 15 years following the baseline examination in 1988–1990. The Y402H polymorphism is being genotyped in the entire population, but is currently available for 1930 participants. Cause of death is obtained from death certificate records. All models are for age at death and are adjusted for gender, blood pressure, diabetes, smoking status, body mass index and total serum cholesterol. There is no association of the Y402H polymorphism with all cause mortality (Hazard ratio and 95% CI for YY compared to HH: 1.16 (0.92, 1.46)) or death due to cancer (1.14 (0.72,1.79)) but is associated with death from heart disease (1.52 (1.07,2.17)). Further adjustment for history of cardiovascular disease at baseline attenuates this relationship. Similar results hold when we consider the Y402H polymorphism risk as the YY vs HH or HY. Our results suggest the Y402H polymorphism in the CFH gene may increase risk of cardiovascular disease.

**188**
**WITHDRAWN**

**189**
**HLA Class I A And B Determinants of Type 1 Diabetes in the Belgian Population and Interaction with Autoantibodies**

Eric Mbunwe (1), Eva Van der Stockt (1), Ines Hellemans (1), Chris Van Schravendijk (1), Bart Van der Auwera (1)
(1) Vrije Universiteit Brussel

**Background:** Genome wide association studies have confirmed HLA-DR, -DQ independent effects of HLA class I to type 1 diabetes (T1D) genetic risk.
**Aims:** We explored the association of HLA class I alleles with T1D, age at onset and T1D antibodies.
**Methods:** DNA samples of 1793 T1D patients and 877 control subjects drawn from amongst Caucasians of Belgian residence were typed for presence or absence of *HLA-A*, *-B* alleles by PCR-ASO using non-radioactive probes. Data were available for HLA-DQ genotype and antibody status.
**Results:** Presence of *HLA-A∗24*, *-B∗18/39* were over-represented in patients carrying protective HLA-DQ genotypes vs. genetically matched control subjects ($p_c = 0.001$ OR = 2.0; $p_c = 0.03$ OR = 1.9; $p_c < 10^{-6}$ OR = 5.8 respectively). Also, patients carrying susceptible or neutral HLA-DQ genotypes were frequently positive for *HLA-A∗24* ($p_c = 0.03$ OR = 2.0) or *HLA-B∗18* ($p_c = 0.01$ OR = 2.6) respectively. We observed a strong negative association ($P < 10^{-4}$) of *A∗24, B∗18/39* positivity with presence of high risk DQ8 (OR, 0.6, 0.5, 0.3 respectively). *HLA-A∗24* showed a weak association with age at onset ($p = 0.03$ OR = 1.3). Finally, *HLA-A∗24* positivity was preferentially associated with absence of IA-2A, GADA or ICA ($p_c = 0.01$, 0.01 or 0.04) respectively amongst antibody positive patients.
**Conclusion:** The data suggest an additional value for HLA class I typing in T1D immunogenetic risk assessment in Belgium and highlights the relevance of conventional autoantigens in *HLA-A∗24* related T1D autoimmunity.

**190**
**Prevalence of Genetic Disorders in the Northwest of Iran**

Tuba Mizani (1), Saeed Dastgiri (2), Morteza J. Bonyadi (3)
(1) Department of Community Medicine, Tabriz University of Medical Sciences
(2) Department of Community Medicine, Tabriz University of Medical Sciences
(3) Department of Medical Genetics, Tabriz University of Medical Sciences

**Background and Aim:** genetic disorders are responsible for a major proportion of mortality, morbidity and handicap across the world varying by racial, ethnicity and cultural differences. The aim of this study was to estimate the prevalence of genetic disorders in the northwest of Iran.
**Methods:** in this study, 2968 cases were identified between 2005 and 2009. Prevalence rate, descriptive statistics and 95 percent confidence intervals (CI95%) were used for data analysis.

**Results:** the study subjects included patients, carriers, suspected cases and unknown cases. The most prevalent disorders (per 100000 population) were identified as Familial Mediterranean Fever (20.6, CI95%: 19.1–22.07), Inherited Deafness (11.4, CI95%: 10.6–12.2), Spinal-Muscular Atrophy (11.1, CI95%: 10.1–12.1), Cystic Fibrosis (7.9, CI95%:7.1–8.2), Duchenne Muscular Dystrophy (7.8, CI95%:6.2–8.7) and Down Syndrome (5.1, CI95%:4.5–5.4).
**Conclusion:** estimating the true prevalence of the genetic disorders may help in the planning of health care and screening programs. The prevalence of these diseases in the region indicates the necessity to establish a population-based center for genetic disorders in the area. More population-based investigations are however needed to develop effective preventive strategies to control genetic disorders in the region.

**191**

**Common Variation in Genes that Regulate Circadian Rhythms, Interaction with Rotating Shift Work, and Breast Cancer Susceptibility**

Genevieve M. Monsees (1), Peter Kraft (2), Susan E. Hankinson (3), David J. Hunter (3), Eva S. Schernhammer (4)
(1) Fred Hutchinson Cancer Research Center, University of Washington, Harvard School of Public Health, Brigham and Women's Hospital
(2) Harvard School of Public Health
(3) Harvard School of Public Health, Brigham and Women's Hospital
(4) Brigham and Women's Hospital

Rotating shift work is associated with breast cancer (BC) risk. Variation in genes that regulate circadian functioning (clock genes) may influence BC risk. This is the first study to our knowledge to investigate whether the effect of common variation in these genes is modifiable by exposure to rotating shift work. A mixed candidate- and tag- single nucleotide polymorphism approach was used to select 178 markers across 15 clock genes. Logistic regression was used to test for association between variants and BC risk among 609 prevalent and incident BC cases and 1216 matched controls from the Nurses' Health Study II. Potential interaction between genotype and rotating shift work on incident BC risk was investigated in a subset of 438 incident cases and 880 matched controls. When accounting for potential effect modification, an NPAS2 coding variant (rs2305160:G > A) was most strongly associated with BC risk (nominal test for interaction $p$-value = 0.0005). The observed multiplicative increase in odds of BC per minor allele (A) was 0.65 (95% CI: 0.51–0.82) among individuals with <24 months of cumulative rotating shift work, and 1.19 (95% CI: 0.93–1.54) among individuals with ?24 months of cumulative rotating shift work. Our findings suggest that common variation alone in clock genes plays at most a small role in determining BC risk among women of European ancestry. The impact of NPAS2 rs2305160 in the presence of shift work requires further investigation.

**192**

**Genetic Variation at Adenylate Cyclase 5 (ADCY5) is Associated with Glycemic Control in type 1 Diabetes**

Andrew D. Paterson (1), Mohsen Hosseini (1), Daryl Waggott (2), Andrew P. Boright (3), Enqing Shen (2), Marie-Pierre Sylvestre (2), Patricia A. Cleary (4), John M. Lachin (4), Jennifer E. Below (5), Dan Nicolae (5), Nancy J. Cox (5), Niina Sandholm (6), Carol Forsblom (6), Per-Henrik Groop (6), Angelo J. Canty (7), Lei Sun (8), Shelley B. Bull (2), DCCT/EDIC Research Group (9)
(1) Hospital for Sick Children
(2) Lunenfeld
(3) UHN
(4) George Washington University
(5) U Chicago
(6) University of Helsinki
(7) McMaster University
(8) University of Toronto
(9) NIDDK

**Objective**: Evidence from twins and genetic association studies point to common variants for glycemia in both type 1 diabetes (T1D) and non-diabetics. 19 loci associated with glucose in individuals without diabetes were evaluated for association with glycemic traits in T1D cohorts.
**Research Design and Methods**: 22 SNPs were obtained from GWAS data in 3 T1D studies: DCCT (637 intensively (INT) and 667 conventional treated subjects); GoKinD (nephropathy cases ($n = 531$) and controls ($n = 851$)); FinnDiane ($n = 3340$).Linear regression analysis was performed.
**Results**: Of 19 loci, rs11708067 at *ADCY5* was associated with mean A1C ($p = 0.00024$) and mean 7-point glucose ($p = 0.0045$) in DCCT/INT, in the same direction as in non-diabetics. This SNP also showed borderline association with A1C in the same direction ($p = 0.057$) in GoKinD cases. Consistent with the direction of effect on glycemia, associations of this SNP with time-to-diabetes complications in DCCT/INT ($p = 0.0039$, 0.034 for clinically significant macular edema and severe non-proliferative retinopathy, respectively) were attenuated by adjusting for glycemia. No significant association of these SNPs with A1C was observed in the other T1D cohorts.
**Conclusions**: *ADCY5* is associated with glycemia in both T1D and non-diabetics. Detection of association in DCCT/INT may be explained by the A1C level most similar to non-diabetics, smaller within individual variation of A1C compared to the other cohorts, and better precision via repeated measures.

**193**

**Association between SLC2A9 Transporter Gene Variants and the Serum Uric Acid Phenotype Reveals Within-gene Differences between Persons of European and African Ancestry**

Andrew D. Rule (1), Mariza de Andrade (1), Martha E. Matsumoto (1), Tom H. Mosley (2), Sharon L.R. Kardia (3), Stephen T. Turner (1)
(1) Mayo Clinic
(2) Univ of Mississippi Medical Center
(3) Univ of Michigan

It is well known that increased uric acid levels lead to gout, kidney stones, hypertension, and cardiovascular disease. Uric acid is filtered, reabsorbed, and secreted in the kidney under regulation by the SLC2A9 transporter in renal

tubules and its genetic variation has an established association with serum uric acid in European ancestry populations. Using subjects from the Genetic Epidemiology Network of Arteriopathy family-based cohorts that had Affymetrix SNP Array 6.0 genotype, we identified SLC2A9 gene variants showing association with serum uric acid in 1155 subjects of African ancestry (53 SNPs) and 1132 subjects of European ancestry (63 SNPs) using a linear mixed effects model for correlated data. Analyses were adjusted for age, sex, diuretic use, body mass index, homocysteine, and triglycerides in a parsimonious model. The most statistically significant SNP was in the latter half of the gene and explained 2.77% and 2.71% of the variation in serum uric acid in subjects of European and African ancestry, respectively. After adjustment for the most statistically significant SNP, 0.86% of the variation in serum uric acid in subjects of African ancestry was explained by a SNP in the first half of the gene. These findings show that SLC2A9 transporter variants associate with serum uric acid levels in subjects of African and European ancestry, but there are race-differences in the specific gene regions that most affect serum uric acid levels.

**194**

**Association Between Fuchs Corneal Dystrophy and Age-related Macular Degeneration**

Euijung Ryu (1), Nirubol Tosakulwong (1), Albert O. Edwards (2)
(1) Mayo Clinic
(2) University of Oregon

**Background:** Both Fuchs corneal dystrophy (FCD) and Age-related macular degeneration (AMD) are progressive eye disease that could lead to vision loss in older individuals. Our previous study found genetic variations that were associated with FCD, which included transcription factor 4 (TCF4) gene in chromosome 18.
**Aim:** We attempt to see an association between AMD and FCD severity. We also investigate an effect of FCD associated genetic variants on AMD disease status.
**Methods:** AMD status and FCD severity information were collected from a total of 176 subjects: samples from a collaborative genome-wide association study for AMD were graded for FCD severity and samples from Mayo Clinic Fuchs project were evaluated for AMD status. Ordinal effect size measure was used to see if there is any relationship between AMD status and FCD severity grading. Conditional logistic regression was used for a relationship between FCD associated genetic variants and AMD status.
**Results and Conclusion:** A measure of superiority for FCD severity in terms of AMD status was 0.53 with 95% CI equal to (0.45, 0.60), which implied no relationship between two variables. Single nucleotide polymorphisms (SNPs) associated FCD status did not show any association with AMD status.

**195**

**Genotype-Phenotype Correlation for Phelan-McDermid Syndrome (22q13 Deletion Syndrome)**

Sara M. Sarasua (1), Julianne S. Collins (1), Alka Chaubey (2), R. Curtis Rogers (1), M.C. Phelan (3), Barbara R. DuPont (1)
(1) Greenwood Genetic Center, Greenwood, South Carolina; Department of Genetics and Biochemistry, Clemson University, Clemson, SC
(2) Greenwood Genetic Center, Greenwood, South Carolina
(3) Molecular Pathology Laboratory Network, Maryville, Tennessee

Phelan-McDermid Syndrome, also known as 22q13 Deletion Syndrome, is characterized by global developmental delay, hypotonia, mildly dysmorphic features, and occasional autistic-like characteristics. The syndrome is caused by deletions ranging from 0.1 to 9.0 million bases on the distal long arm of chromosome 22. *SHANK3*, located in the subtelomeric region, is generally assumed to be the primary cause of neurologic phenotypes. The severity and expression of the syndrome are highly variable suggesting other genes in the deletion region may be important. With oligo array comparative genomic hybridization (CGH) techniques becoming clinically available, more affected individuals are being discovered and deletion break points can be more accurately determined. The Greenwood Genetic Center used a custom designed high density oligo array CGH to interrogate 22q12-qter to pinpoint deletion breakpoints down to a resolution of 100 bp in a cohort of more than 100 individuals with this rare syndrome. These array CGH data have been combined with physical exam and medical history data to better understand the developmental and behavioral characteristics of the syndrome. Using these data, the investigators performed a genotype-phenotype correlation study to determine if additional genes in the deletion region are responsible for the many associated features and their severity.

**196**

**New Loci Associated with Lung Function and Chronic Obstructive Pulmonary Disease**

Maria Soler Artigas (1), Louise V. Wain (1), Ma'en Obeidat (2), Emmanouela Repapi (1), Ian Sayers (2), Ian P. Hall (2), Martin D. Tobin (1), SpiroMeta Consortium SpiroMeta Consortium (3)
(1) Departments of Health Sciences and Genetics, Adrian Building, University of Leicester, University Road, Leicester, UK
(2) Division of Therapeutics and Molecular Medicine, Nottingham Respiratory Biomedical Research Unit, University Hospital of Nottin
(3) See Repapi E et al. (2010) Nat Genet. 42(1):36–44 for authors and affiliations

Lung function measures are heritable traits that predict morbidity and mortality. They define chronic obstructive pulmonary disease (COPD), the fourth leading cause of death in the US. We tested genome-wide association with forced expiratory volume in 1s ($FEV_1$) and the ratio of $FEV_1$ to forced vital capacity (FVC) in the SpiroMeta consortium ($n = 20{,}288$, with follow-up in 54,276 individuals). We confirmed the reported association at 4q31 (near HHIP) and identified 5 new associations ($P < 5 \times 10^{-8}$): 2q35 in *TNS1*, 4q24 in *GSTCD*, 5q33 in *HTR4*, 6p21 in *AGER* and 15q23 in *THSD4*. To follow up the loci we tested association of each locus with COPD. The reported locus and 3 novel loci (TNS1, GSTCD, HTR4) showed association

($P < 2 \times 10^{-3}$). Then, we assessed the combined effect of risk alleles at all 6 loci. Compared with a baseline of 7 risk alleles, carrying 10 to 12 risk alleles was associated with an effect equivalent to 4.2 years decline in $FEV_1$ in the non-smoking population (? = −72.2 ml, 95%CI −111.2 ml to −32.8 ml, $p = 3.8 \times 10^{-4}$), and also with $FEV_1$/FVC (? = −1.6%, 95%CI −2.2% to −1.0%, $p = 5.7 \times 10^{-6}$) and with increased risk of COPD (OR = 1.6, 95%CI 1.3 to 2.0, $p = 1.5 \times 10^{-5}$). We also carried out a comprehensive evaluation of previously published lung function associations in the genome-wide data of the SpiroMeta consortium. Overall we found no more associations with lung function than would be expected by chance, suggesting that many reported candidate gene associations may have been false positives.

**197**

**Segregation Analysis of Familial Barrett's Esophagus**

Xiangqing Sun (1), Amitabh Chak (2), Robert Elston (1)
(1) Case Western Reserve University
(2) University Hospitals of Cleveland

Barrett's Esophagus (BE), esophageal adenocarcinomas (EAC), and esophagogastric junction adenocarcinomas (EGJAC) aggregate within families, and are termed Familial Barrett's Esophagus (FBE). A segregation analysis of FBE reported evidence of rare autosomally inherited dominant susceptibility alleles with incomplete penetrance and a polygenic component (Sun et al., 2010. Cancer Epidemiol Biomarkers Prev 19:666–674). Although it is known that the incidence of BE has increased over the last few decades and the prevalence of BE increases with age, in that analysis age and date of birth (DOB) were not explicitly considered because age was missing on 40% and DOB was missing on 83% of the family members. In this study, in order to improve the estimation of the inheritance parameters, we impute the missing age and DOB according to the known age and/or DOB information of all members in each nuclear family. This is made possible by the fact the sum of age and DOB is approximately a constant for all members of a nuclear family, as well as by using the estimated mean age differences between these relatives (Schnell et al. 2000. Am J Med Genet 92:212–219). The imputing error is evaluated, and a new segregation analysis of FBE using the imputed ages and DOBs is shown to result in better estimates of the inheritance model.

**198**

**Genomewide Association with Alpha1-Antitrypsin Blood Levels in a Subset of the SAPALDIA Cohort Study**

Gian Andri Thun (1), Ashish Kumar (1), Medea Imboden (1), Ivan Curjuric (1), Thierry Rochat (2), Erich W. Russi (3), Maurizio Luisetti (4), Nicole M. Probst-Hensch (1)
(1) Swiss Tropical and Public Health Institute, Basel, Switzerland
(2) Division of Pulmonary Medicine, University Hospital of Geneva, Switzerland
(3) Pulmonary Division, University Hospital of Zurich, Switzerland
(4) Institute for Respiratory Disease, IRCCS San Matteo Hospital Foundation, University of Pavia, Italy

Reduced concentration of alpha1-antitrypsin (AAT) in the blood is an established risk factor for pulmonary emphysema. The two alleles of the *SERPINA1* coding region determine the AAT phenotype and AAT blood levels. However, serum concentrations are also modified by external factors and potentially other gene regions within or outside the *SERPINA1* locus. We performed a genome wide association analysis for determinants of circulating AAT in the population-based Swiss Cohort on Air Pollution and Lung Disease in Adults (SAPALDIA). A SAPALDIA subgroup ($n = 1612$) was genotyped using the Human Illumina 610 quad array. The association of 497K quality controlled single nucleotide polymorphisms (SNPs) with serum AAT was assessed using fixed effects linear regression in an additive genetic model adjusted for gender, age, study center, and population stratification. SNPs that reached genome wide significance were located in the SERPINA gene cluster at 14q32.1. The two top ranking SNPs were found in close proximity to the *SERPINA6* gene ($p$-value: $2.90 \times 10^{-12}$ and $1.28 \times 10^{-10}$; $D' = 0.4$, $r^2 = 0.05$). Along the line of our previous observation of gender differences in circulating AAT levels, we observed sex-specific differences in the association of SNPs within the SERPINA gene cluster with serum AAT. Our findings confirm that *SERPINA1* is a major genetic determinant of AAT blood levels. However, our preliminary analysis points to a more complex cis-regulated control of the *SERPINA1* gene expression.

**199**

**Mitochondrial DNA Variation in Human Metabolic Rate, Energy Expenditure and Mortality: The Health, Aging and Body Composition Study**

Gregory J. Tranah (1), Todd M. Manini (2), Kurt K. Lohman (3), Michael A. Nalls (4), Stephen Kritchevsky (5), Anne B. Newman (6), Tamara B. Harris (7), Alessandro Biffi (8), Steven R. Cummings (1), Yongmei Liu (5)
(1) California Pacific Medical Center
(2) University of Florida, Department of Aging and Geriatric Research
(3) Department of Biostatistical Sciences, Wake Forest University School of Medicine
(4) Laboratory of Neurogenetics, Intramural Research Program, National Institute on Aging
(5) Sticht Center on Aging, Wake Forest University School of Medicine
(6) Department of Epidemiology, University of Pittsburgh
(7) Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging
(8) Center for Human Genetic Research, Massachusetts General Hospital

The majority of the body's energy needs are met by OXPHOS and the genes encoded by the mtDNA are crucial for assembling the OXPHOS machinery. There is a long-standing debate over the role of climate in driving adaptive selection of mtDNA as Homo sapiens migrated out of Africa into temperate and arctic Eurasia. We evaluated the role of mtDNA variation in resting metabolic rate (RMR) and total energy expenditure (TEE) among 294 older, community-dwelling African and European American adults from the Health, Aging and Body Composition Study. A total of 137 mtDNA variants were genotyped and common African and European haplogroups were defined. Following multivariate adjustment, participants from African haplogroup L had lower RMR (1221?10 kcal/d vs.1327?9 kcal/d; $P < 0.0001$) and

TEE (2073?27 kcal/d vs.2241?25 kcal/d; $P<0.0001$) when compared to European haplogroup N participants. Common African haplogroups L0, L2 and L3 had significantly lower RMRs than European haplogroups H, JT and UK. African haplogroup L1 had an RMR intermediate to the European and remaining African haplogroups. Ten year survival was greatest for European haplogroup J and the highest mortality was observed among African haplogroups L3 (HR = 1.70, 95% CI = 1.14–2.55, $p = 0.009$) and L2 (HR = 1.68, 95% CI = 1.12–2.53, $p = 0.01$) after adjusting for baseline age, sex, and education. The role of mitochondrial-nuclear gene epistasis in human aging will also be explored.

**200**

**Familial Aggregation of Age at Diagnosis of Prostate Cancer**

Wei Wang (1), Lisa A. Cannon Albright (2), Peter V. Tishler (3)
(1) Brighan & Women's Hospital, Division of Sleep Medicine
(2) Medicine, University of Utah School of Medicine
(3) Brigham & Women's Hospital, Channing Laboratory

The intrafamilial age at onset of prostate cancer (PrCa) may be correlated. To assess this, we studied the familial aggregation of age at diagnosis (Dx), a surrogate for age at clinical onset. Methods: The Utah Population Database provided 2,071 families with an index case & >1 1? relative with PrCa. Using a proband-predictive model, we correlated age at Dx of affected relatives with index case age at Dx. The statistic is a regression coefficient (ss) & 95% CI (CI). Results: Index case age at Dx was significantly correlated with brothers age at Dx (ss = 0.14; 95% CI = 0.09, 0.19; $P<0.0001$). Significant correlations remained for index cases with Dx ≤ age 60 & >70, for sibships in which the brothers' Dx preceded or succeeded the index case Dx, & for sibships with Dx periods that were similar (both ≤1988 or >1988) or dissimilar (index case Dx ≤1988 or >1988, brother's Dx >1988 and ≤1988). The correlations were significant in sibships with a different index case, but were not significant in ''pseudosibships'' of index cases & unrelated brothers. Conclusions: The age at Dx/clinical onset of PrCa is highly correlated among brothers throughout their age range. The correlation is not influenced by the order or the calendar year (the method) of Dx. This reflects the lifelong operation of biologic, especially genetic, factors.

**201**

**WITHDRAWN**

**202**

**Excessive Contiguous Homozygous Runs in the Human Major Histocompatibility Complex Region in Rheumatoid Arthritis Patients**

Hsin-Chou Yang (1), Lun-Ching Chang (1), Yu-Jen Liang (1), Chien-Hsing Lin (2)
(1) Institute of Statistical Science, Academia Sinica
(2) Division of Molecular and Genomic Medicine, National Health Research Institutes

Rheumatoid arthritis (RA) is a chronic inflammatory disorder with multifactorial etiology and a polygenic mode of inheritance. This study aims to examine excessive contiguous homozygous runs (ECHR) in RA patients, which provides a new hypothesis for the genetic mechanisms of RA. A genome-wide non-parametric ECHR association test is proposed, and using this powerful method, the first genome-wide ECHR association study is conducted by analyzing 2,000 RA patients and 3,000 normal controls from the Wellcome Trust Case Control Consortium. The genome-wide ECHR association scans consistently pinpoint a region from ~32.5 Mb to 32.7 Mb ($-\log10(p) > 8$) within the human major histocompatibility complex (MHC) region at 6p21.3. RA susceptibility genes such as *HLA-DRB1* are contained in this region. More than 20% of RA patients can be characterized by the pattern of ECHR in this region. The results illustrate that ECHR in the human MHC region plays a role in the complex etiology of RA. Genome-wide ECHR mapping provides a new alternative to allelic association mapping for the identification of RA susceptibility genes.

**203**

**A Repeated Measures Genome Wide Association Study of Blood Pressure in Type 1 Diabetes**

Chang Ye (1), Angelo J. Canty (1), Daryl Waggott (2), Marie-Pierre Sylvestre (2), Enqing Shen (2), Mohsen Hosseini (3), Andrew P. Boright (4), Lei Sun (5), Shelley B. Bull (2), Andrew D. Paterson (3), DCCT/EDIC Research Group (6)
(1) McMaster University
(2) Samuel Lunenfeld Research Institute
(3) Hospital for Sick Children
(4) University Health Network
(5) University of Toronto
(6) NIDDK

Most GWAS of quantitative traits use single measures although many traits vary within an individual over time. Participants with type 1 diabetes (T1D) in the Diabetes Control and Complications Trial (DCCT) had blood pressure (bp) measured quarterly for an average of 6.5 years. We take advantage of the repeated measures using a mixed effects model to identify loci associated with blood pressure in T1D.

**Methods**: We used data from 1302 white DCCT participants and tested 814 K SNPs with MAF?1%. Systolic and diastolic bp were analyzed separately. A univariate GWAS was conducted using a random time effect for the repeated measures. A second GWAS included design covariates, gender, diabetes duration and age. A forward selection procedure identified BMI, triglycerides and HbA1c as also associated with blood pressure. These were measured annually so we included annual measures of the trait and covariates in an extended GWAS.

**Results**: RS7842868 (chr 8) was associated with diastolic bp in the univariate GWAS ($p = 4.5 \times 10^{-8}$). Other SNPs in the same locus also showed suggestive association. This SNP also showed suggestive evidence of association with systolic bp ($p = 6 \times 10^{-6}$). Inclusion of covariates had little effect on the signal. We also examine interactions between rs7842868 and some of the covariates.

**204**

**Incorporating Age-Dependent Effects for Alzheimer's Disease in Genome-Wide Association Study: Framingham Heart Study**

Jacqueline L. Buros (1), Gyungah Jun (1)
(1) Boston University School of Medicine

Recent analysis of Framingham data for genome-wide association of SNP data with Alzheimer's disease (AD) yields no novel findings and replicates no result outside of APOE. Multifactorial diseases such as AD often have residual, unexplained heritability concurrent with non-informative GWAS results for association with disease status. We hypothesize that SNP data may instead demonstrate age-dependent association with AD status, that they may interact with AD-correlated traits in association with AD, or that SNPs may show interaction with the rate of change of these traits with respect to AD status. We analyzed Framingham subjects genotyped on the Affy 550 k chip per protocol with non-missing data for AD, MRI traits, stroke, smoking and cognitive testing. SNPs and persons are cleaned for quality control including minor allele frequency >2%, call rate >98%, & HWE p>10e−6, leaving 346,228 SNPs and 2400 subjects. Covariates values, rates of change, and interaction terms are all time-varying. All analyses use survival models with age of diagnosis or last ascertainment as the time to event, and all are evaluated as a generalized linear latent and mixed model (GLLAMM) using the kinship correlation coefficient matrix. Results are compared with those from testing for association with AD status by GEE using age as a covariate. Our analysis using rate of changes and survival models produced new findings that were previously undetected using AD affection status as a trait.

**205**

**Identification of Genetic Variants Related to Hepatitis B Virus Infection in Hepatocellular Carcinoma Families and Chronic Liver Diseases: A Genome-Wide Association Study**

Su-Wei Chang (1), Dar-In Tai (2), Chia-Lin Hsu (1), Cathy S.J. Fann (1)
(1) Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
(2) Liver Research Unit, Chang Gung Memorial Hospital, Chung Gung University College of Medicine, Taipei, Taiwan

Hepatitis B is a major health problem worldwide and a potentially life-endangering liver infection. Previous studies have shown that male gender was a significant host factor related to hepatitis severity. However, a large part about the biomedical mechanism of persistence and protection for HBV infection is still unknown. In this study, we conducted a genome-wide association analysis on 625 male Taiwanese. The sample comprised 321 HBsAg positive cases and 304 HBsAg negative controls with genotype data from Illumina HumanHap550 K and 610 K beadchips. A total of 456,262 SNPs were used in the GWAS analysis after quality control filters with an average call rate of 99.95%. We performed single-locus association tests and identified 61 SNPs that were potentially associated with persistent HBV infection. Among those, 38 out of the 61 markers were clustering in the HLA Class II region on chromosome 6. In addition, we found association at six loci rs10484569, rs2281388, rs3117222, rs3128917, rs9277535, and rs9380343 ($10^{-9}<P<10^{-5}$) and confirmed previous findings of Kamatani et al.'s (Nature Genetics 2009;41:591–595). Moreover, we found that BTNL2, HLA-DQA2, HLA-DQB2, and CDC2L5 showed significant associations with the HBV persistence ($P<10^{-5}$). The underlying function remains uncertain and requires further investigation.

**206**

**Set Level Association Testing in a GWA Study on Preterm Birth**

Frank Geller (1), Bjarke Feenstra (1), Hao Zhang (1), Heather A. Boyd (1), Kelli K. Ryckman (2), John R. Shaffer (3), John Dagle (2), Daniel E. Weeks (3), Mary Marazita (3), Eleanor Feingold (3), Jeffrey C. Murray (2), Mads Melbye (1)
(1) Statens Serum Institut, Denmark
(2) University of Iowa
(3) University of Pittsburgh

**Background:** Initial analyses of genome-wide association studies (GWAS) mainly analyze the data at the single SNP level. In subsequent analyses specific genomic regions, e.g. candidate genes, should be analyzed, allowing for the possibility of multiple independently associated SNPs. Here, an approach that takes the linkage disequilibrium between SNPs into account is needed.
**Methods:** We analyzed our GWAS on preterm birth applying Set Level Association Testing, a method recently introduced by de la Cruz et al. A total of 20,756 genes from the Panther Database were covered by 1 to 1,653 (mean 28) SNPs.
**Results:** None of the genes reached the Bonferroni-adjusted significance level for the sets (0.00000247), but some provided *p*-values below 0.0001 that were smaller than the minimum *p*-value observed for any SNP in the region, warranting further investigation. Several sets including SNPs that were among the ten top hits on the SNP level were not among the top 100 sets.
**Conclusions:** Given the fact that some GWAS have identified multiple basically independent signals in narrow genomic regions, a set based method has the potential to identify such regions swiftly. It remains to be seen, whether regions with significant set *p*-values turn out to harbor associated SNPs.

**207**

**Imputed Genome Wide and Candidate Gene Studies Reveals Novel Genetic Variant Associated with Venous Thromboembolism (VTE)**

John A. Heit (1), Martha E. Matsumoto (1), Yan Asman (1), Elysia Jeavons (1), Sebastian M. Armasu (1), Tanya M. Petterson (1), Julie M. Cunningham (1), Mariza de Andrade (1)
(1) Mayo Clinic

Factor V Leiden, prothrombin G20210A (*F2*) and ABO blood type non-O are known genetic risk factors for VTE. To identify novel VTE-susceptibility genes, we performed candidate gene (CG) and genome-wide association (GWA) studies. The CG study included 1486 VTE cases and 1439 controls frequency-matched on case age, sex, and state of residence with 12,497 functional or haplotype tag SNPs within 754 genes relevant to the anticoagulant, procoagulant, fibrinolytic and innate immunity pathways. The GWA study included a GC patient subset (1270 VTE cases; 1302 controls) with 557,391 SNPs (Illumina 660W Quad). Four discovery phase analysis strategies were used: 1. CG only, 2. GWA only, 3. Imputed GWA using BEAGLE and HapMap Phase 3 (96 CEU and 88 TSI) as reference sample, 4. Imputed merged data (CG+GWA SNPs) using MACH and HapMap Phase 2 (60 CEU). Using CG only, we identified *FV*, *ABO*, *F2*, and three novel genes (Bonferroni corrected p?10E-5). Using GWA only, the same four genes (3 on 1q and 1 on 9q) reached genome-wide significance (p?10E-8). Using imputed GWA,

we identified the same 4 genes. Using the imputed merged data, we identified the same 4 genes plus a novel gene on chromosome 1. In a detailed *ABO* haplotype analysis using 1,000 Genomes Project, a novel *ABO* tag SNP was is in high LD with a conserved *ABO* region that was independent of the ABO blood type coding regions, providing new insights regarding the association between *ABO* and VTE.

**208**

**Extreme-phenotype Small Sample Genome-wide Association Study of Early-onset Acute Coronary Syndrome Cases and Age-discordant Controls**

Benjamin D. Horne (1), John F. Carlquist (2), Joseph B. Muhlestein (2), Muredach P. Reilly (3), Mingyao Li (4), Mary-Susan Burnett (5), Joseph M. Devaney (5), Stephanie L. DerOhannessian (3), Stephen E. Epstein (5), Daniel J. Rader (3), Jeffrey L. Anderson (2)
(1) Cardiovascular Dept., Intermountain Medical Center; Genetic Epidemiology Div., University of Utah
(2) Cardiovascular Dept., Intermountain Medical Center; Cardiology Div., University of Utah
(3) Institute for Translational Medicine & Therapeutics and the Cardiovascular Institute, University of Pennsylvania
(4) Biostatistics and Epidemiology, University of Pennsylvania
(5) Cardiovascular Research Institute, MedStar Research Institute, Washington Hospital Center

Acute coronary syndromes (ACS) are the most severe coronary heart disease (CHD) phenotypes. A genome-wide association study (GWAS) of single nucleotide polymorphism (SNP) associations with extreme CHD case and control phenotypes was performed in a small sample size. Age-discordant GWAS studied 80 patients with early-onset ACS and ?2-vessel coronary disease (age of females: 48.6?3.0 years [range: 43–53]; males: 38.5?2.9 y [30–42]). Controls were 80 angiography patients free from CHD, stroke, and all other cardiovascular diseases (age of females: 80.6?3.4 y [76–89]; males: 76.4?3.9 y [71–85]). All were unrelated, never-smokers, and non-diabetic. The HumanHap 550 K chip was used and analysis tested allelic associations. Replication was performed by meta-analysis of *in silico* and wet lab data for standard phenotypes in 3 populations (Univ Pennsylvania, Washington Hosp Cntr, Intermountain Med Cntr; case $n = 5846$, control $n = 4154$). GWAS statistical power was 8–13% to find $P < 0.05$ for the CHD 9p21.3 locus based on published minor allele frequencies for rs1333049 (2 case/control: 0.55/0.47, 0.54/0.48 [Samani NEJM 2007;357:443–53]), but the SNP had $p = 0.004$. Five loci had $P < 1 \times 10^{-5}$ and one replicated (rs2868882, *EYA2*, meta-$p = 0.0054$). A small sample of 80 cases and 80 controls theoretically imparted low power to an extreme-phenotype GWAS, but actual power was sufficient to replicate published CHD SNPs and to discover a novel, replicable CHD locus at the transcription factor *EYA2*.

**209**

**Association of Asbestos Exposure and Genetic Modification on Lung Cancer Risk**

Chen-yu Liu (1)
(1) HSPH

To assess whether there was an association between lung cancer linked to asbestos and genetic variability, we conducted a genome-wide association analysis in 1,000 Caucasian cases and 1,000 Caucasian controls using Illumina Human 610-Quad BeadChips. Cumulative lifetime asbestos exposure score (AES) was calculated from self-reported duration and intensity of occupational and nonoccupational exposures. A total of 12.1% of cases and 8.5% of controls had ''high'' AES, which was found to be associated with pleural or parenchymal abnormalities from the previous study.

**210**

**Analysis of Heterogeneity by Autopsy-Confirmation Status in Genome-wide Association of Late-Onset Alzheimer Disease Identifies Limited Heterogeneity in the Strongest Associations**

Adam C. Naj (1), Gary W. Beecham (1), Eden R. Martin (1), Paul J. Gallins (1), Ruchita Rajbhandary (1), Kara Hamilton (1), Ioanna Konidari (1), Patrice L. Whitehead (1), Guiqing Cai (2), Vahram Haroutunian (2), John R. Gilbert (1), Jonathan L. Haines (3), Joseph D. Buxbaum (2), Margaret A. Pericak-Vance (1)
(1) John P. Hussman Institute for Human Genomics
(2) Mount Sinai School of Medicine
(3) Vanderbilt Center for Human Genetics Research

Alzheimer Disease (AD) is highly genetic (estimated $H^2 \sim 70\%$), but until recently, genome wide association studies (GWAS) have only observed consistent genetic associations with late-onset AD (LOAD) in *APOE*, possibly due to heterogeneity such as etiologic heterogeneity and ascertainment bias. To determine whether associations are being masked potentially by heterogeneity in a previous GWAS of LOAD, we used multinomial logistic regression analyses in SAS (v9.2) to re-evaluate associations of 483,399 SNPs among 931 cases and 1104 controls by, stratifying cases by autopsy confirmation of AD clinical diagnosis (370 autopsy-confirmed/561 clinically-confirmed). We confirmed associations with SNPs in/near *APOE* on chromosome 19, and a SNP association meeting genome-wide statistical significance, rs11754661 (OR (95% CI): 2.03 (1.58, 2.62); $p = 4.70 \times 10^{-8}$) still demonstrated strong associations for autopsy-confirmed (OR (95% CI): 1.84 (1.26, 2.68); $p = 0.00146$) and clinically-confirmed (OR (95% CI): 2.04 (1.55, 2.70); $p = 4.20 \times 10^{-7}$) comparisons. Our stratified analyses found that effect sizes of associations within stratum were similar to effect sizes observed in non-stratified GWAS analyses of LOAD, though with reduced statistical significance likely due smaller sample sizes of the strata. Weak heterogeneity by autopsy-confirmation status may reflect improved accuracy of clinical AD diagnosis, however further investigation is needed.

**211**

**Genome-Wide Association Study (GWAS) and Fine Mapping of NT-proBNP Level: Novel Loci and Novel Variants in the MTHFR-CLCN6-NPPA-NPPB Gene Cluster**

Cristian Pattaro (1), Fabiola Del Greco M. (1), Andreas Luchner (2), Irene Pichler (1), Thomas Winkler (3), Andrew A. Hicks (1), Christian Fuchsberger (1), Andre Franke (4), Scott A. Melville (5), Annette Peters (6), H-Erich Wichmann (6), Stefan Schreiber (7), Iris M. Heid (3), Michael Krawczak

(8), Cosetta Minelli (1), Christian J. Wiedermann (9), Peter P. Pramstaller (1)
(1) Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy
(2) Klinik und Poliklinik fur Innere Medizin II, Klinikum der Universitat Regensburg, Regensburg, Germany
(3) Department of Epidemiology and Preventive Medicine, Regensburg University Medical Center, Regensburg, Germany
(4) Institute for Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany
(5) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA
(6) Helmholtz Zentrum Munchen, Institute of Epidemiology, D-85764 Neuherberg, Germany
(7) Institute for Clinical Molecular Biology and popgen biobank, Christian-Albrechts-University Kiel, Kiel, Germany
(8) Institute of Medical Informatics and Statistics and popgen biobank, Christian-Albrechts-University, Kiel, Germany
(9) Department of Internal Medicine, Central Hospital of Bolzano, Bolzano, Italy

High blood concentration of the N-terminal cleavage product of the B-type natriuretic peptide (NT-proBNP) is strongly associated with cardiac dysfunction and is increasingly used for heart failure diagnosis.
To identify genetic variants associated with NT-proBNP level we performed a GWAS in 1,325 individuals from South Tyrol, Italy and followed-up the most significant results in 1,746 individuals from two German population-based studies. One GWA signal was replicated after correction for multiple testing, the *MTHFR-CLCN6-NPPA-NPPB* gene cluster (replication 1-sided *p*-value = 8.4E-10), and two other signals were replicated at statistical nominal level, the *MEIS1* (*p*-value = 0.026) and *NR2F1* (*p*-value = 0.014) genes. A conditional regression analysis of 128 single nucleotide polymorphisms in the *MTHFR-CLCN6-NPPA-NPPB* locus identified novel variants in the *CLCN6* gene as independently associated with NT-proBNP. In this locus, four haplotypes were associated with increased NT-proBNP levels (haplotype-specific combined *p*-values from 8.3E−03 to 9.3E−11). The increase in NT-proBNP level was proportional to the number of haplotype copies (dosage effect), with the increase associated with two copies varying between 20 and 100 pg/ml across populations.
The identification of two new candidate genes for NT-proBNP, known to be involved in cardiac muscle development, and novel variants in the *MTHFR-CLCN6-NPPA-NPPB* cluster provide new insights in the biological mechanisms of cardiac dysfunction.

**212**

**Identification of Novel SNPs Significantly Associated with Melanoma and Prostate Cancer in a Genome-wide Association Analysis.**

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

It has been shown that genetic risk variants at a single locus may contribute to multiple cancer types, and that such variants may exist for melanoma and prostate cancer. We conducted a genomewide association study (GWAS) for a compound cancer phenotype using 75 familial prostate cancer and 87 familial melanoma cases genotyped on the Illumina 550 K platform. For controls, we used a set of 2400 genetically matched Caucasian iControls from Illumina's iControlsDB, similarly genotyped. We conducted a cursory scan for association ignoring known relationship information using Plink software. We then re-analyzed the 29 SNPs with *p*-values exceeding $5 \times 10^{-6}$ resulting from the cursory scan with Genie software, which does account for known relationships in determining empirical *p*-values for tests of association. The Genie analysis identified significant SNPs on chromosome 2q12.1, the most extreme having an empirical *p*-value of $2.2 \times 10^{-8}$ (OR = 2.8). The genomic inflation factor between cases and controls was 1.09, indicative of genetic admixture among the controls. After correcting for this using the genomic control strategy, the result was still genomewide significant (empirical $p = 8.54 \times 10^{-8}$). The large magnitude of the OR for this finding is likely due to the familial nature of the cases, where the number of sporadic cases is lower than might be encountered in a cohort of unrelated cases. We are now confirming this result in a set of independent cases and local controls.

**213**

**Genome-wide Association Study of Endometriosis Shows Differential Etiology by Stage and Identifies a Locus at 7p15.2 Associated with the Development of Moderate-severe Disease**

Krina T. Zondervan (1), Jodie N. Painter (2), Carl A. Anderson (1), Dale R. Nyholt (2), Stuart Macgregor (2), S. Hong Lee (2), Peter M. Visscher (2), Peter Kraft (3), Nicholas G. Martin (2), Andrew P. Morris (1), Susan A. Treloar (2), Stephen H. Kennedy (4), Stacey A. Missmer (5), Grant W. Montgomery (2)
(1) Wellcome Trust Centre for Human Genetics, Oxford
(2) Queensland Institute of Medical Research, Brisbane
(3) Harvard School of Public Health, Boston
(4) Nuffield Dept of Obstetrics and Gynaecology, Oxford
(5) Brigham and Women's Hospital and Harvard Medical School, Boston

Endometriosis is a common gynaecological disease associated with severe pelvic pain and sub-fertility. There is considerable debate whether different endometriosis stages represent disease progression, or whether moderate-severe (rAFS III/IV) disease is pathological and minimal-mild (rAFS I/II) an epiphenomenon. We conducted a GWA study using 540,082 SNPs in 3,194 surgically confirmed endometriosis cases and 7,060 controls from Australia and the UK. We used novel statistical methods to estimate the proportion of common variation explained by all markers and performed polygenic predictive modelling for disease stage, both showing significantly increased genetic loading among the 42% of cases with moderate-severe endometriosis. We subsequently genotyped 72 SNPs in an independent US dataset comprising 2,392 endometriosis cases and 1,646 controls. An association with rs7798431 on 7p15.2 for moderate-severe endometriosis ($p = 6.0 \times 10^{-8}$, OR = 1.34 (1.21–1.49)) was replicated, reaching combined genome-wide significance ($p = 1.7 \times 10^{-9}$; OR = 1.26 (1.17–1.35)). The implicated inter-genic region involves a 48 kb segment of high LD upstream of plausible candidate genes *NFE2L3* and *HOXA10*. This locus is the first to be robustly implicated in the etiology of endometriosis, with evidence of association limited to moderate-severe disease. The results show how

well-defined subphenotype analyses can help elucidate differential etiologies of heterogenous complex diseases.

**214**

**Genome-wide Bootstrap Bias Reduction for Point and Interval Estimation that Accounts for Ranking- and Threshold-selection Bias in Discovery GWAS, with Implications for Replication Study Sample Size**

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 Institute of Mount Sinai Hospital, Canada
(2) Dalla Lana School of Public Health, University of Toronto; Department of Statistics, University of Toronto, Canada

Genetic effect estimates obtained from discovery studies are often biased upward due to a form of selection known as the winner's curse. In GWAS the effect is exacerbated by the stringent selection threshold and ranking over a half million or more SNPs. Both bootstrap-based and likelihood-based bias-reduced estimators have been proposed, however, the joint effect of threshold-based selection, ranking of SNPs and complex correlation structure among linked SNPs can be difficult to model. We extend the multi-locus linkage bootstrap method to the GWAS setting, and develop confidence interval construction for the bootstrap effect estimate. In this setting, we introduce two corrections crucial to GWAS that improve the accuracy of the original bootstrap estimates. In extensive simulations we demonstrate that the genome-wide bootstrap method effectively reduces estimation bias, and show that replication study sample sizes computed from the uncorrected estimates are almost never adequate for the desired power, while those computed from the bootstrap bias-reduced estimate are adequate more than 80% of the time. We illustrate the flexibility of the method in various applications, including SNP selection by rank and SNP detection by multiple genetic model tests, and shed light on some apparent failures of replication studies. The method has been implemented in an efficient and flexible program suitable for large-scale GWAS.

**215**

**A Flexible, Efficient and User-friendly Tool for Genome-wide Association Analyses**

Jian'an Luan (1), Jing-Hua Zhao (1), Daniel Barnes (1), Ruth J.F. Loos (1)
(1) MRC Epidemiology Unit, Institute of Metabolic Science, Cambridge, UK

Genome-wide association studies (GWAS) have been conducted extensively to study genetic variants in relation to human diseases and other quantitative traits. It often uses imputed SNPs based on genotype data obtained from the HapMap Project and/or the 1,000 Genomes Project, leading to a total of 2–8 million SNPs. Our analysis for GWAS has revealed that a substantial amount of work is related to data management and preparation for analytical software such as SNPTEST. It is often necessary to conduct trait transformations and specific analyses appropriate for binary or continuous traits with and without adjustment for particular covariates. Not only do these tasks demand statistical knowledge and computing skills, they can also be prone to programming errors or errors from other sources. We have developed Stata packages to automate data preparation and analyses. Our implementation can deal with trait transformation, specification of covariates, sample stratification, collection and reformatting of results, and more. The software is currently available for Linux and has the following advantages: 1) one can run a GWAS analysis using a very basic knowledge of Stata and Linux; 2) the computation speed can be increased using Linux clusters. For example, a case-control score test on a sample of 1400 individuals each with 8.6 million SNPs based on the 1000 Genomes Project can be run on 24 nodes within an hour; 3) other analytical software such as MACH2DAT/MACH2QTL can be employed.

**216**

**A New Region-based Association Test for GWAS Using Partial Least Squares**

Angel Martinez-Perez (1), Alfonso Buil (1), Alexandre Perera (2), Leonor Rib (1), Anna Bricks (1), Jose M. Soria (1)
(1) Unit of Genomics of Complex Diseases. Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau
(2) Centre de Recerca en Enginyeria Biomedica. Universidad Politecnica de Barcelona

**Introduction:** We present a new multi-SNP association test for Genome Wide Association Studies (GWAS), that uses the genetic variability in regions. These regions can be genes or areas with low genetic recombination. The aim is to increase the ability to detect modest genetic effects by using biological information and alleviate the multiple testing problem.
**Methods:** This method divides the genome into regions and, for each region, it performs a single test of association with the phenotype under study. If the genetic regions have a biological function, the results will be interpretable more easily. For each region, we estimate a genetic distance matrix between individuals using a multilocus identical by state metric. This distance summarizes the information in the region into a single measure for each pair of individuals. With this distance matrix and the phenotype, we conducted a distance-based Partial Least Squares analysis (PLS) and retain the first latent variable. Finally, we test the association between the latent variable and the phenotype.
**Results:** Our results showed situations in which the new method is more powerful than the usual SNP by SNP method. Conclusions: It appears that both methods use the genetic information in different ways. Therefore, both tests are complementary.

**217**

**A Two-stage Gene-based Approach for Genome-wide Association Studies**

Jingyuan Zhao (1), Anbupalam Thalamuthu (1)
(1) Genome Institute of Singapore

Genome-Wide Association Studies (GWAS) focusing on single SNP analysis successfully identified a number of important SNPs for complex diseases, but they are only a small proportion of disease phenotype-associated variants partly due to small marginal effects. The gene-based analysis provides an alternative way to understand GWAS data on complex diseases. We propose a two stage gene-based approach to select associated genes in GWAS. We first screen all the genes based on single SNP *p*-values. Only

the genes that survive in the screening step will enter the next step for further association testing. To handle the large number of genes in GWAS, we carry out the selection using the Least Absolute Shrinkage and Selection Operator (LASSO) penalized logistic regression combined with the gene-based scores. Here, we implement the supervised principal component score for each gene as the gene-based score. The *p*-values of top genes are calculated using permutations. In simulated data sets, it has been shown that our proposed method enjoys a higher power to detect the associated genes than some other available methods. We illustrate our proposed method through its application to one of the published GWAS of psoriasis. Our gene-based analysis has identified several relevant genes for psoriasis, which demonstrates that the method is a promising approach to perform the gene-based analysis in GWAS.

**218**

**Schoenfeld Residuals and the 0.632 Estimator for Assessment of Prediction Capability of Survival Traits with a Genetic Variant.**

Yesilda Balavarca (1), Heike Bickeboeller (1)
(1) Dept. of Genetic Epidemiology, University of Medicine, Goettingen

Different versions of coefficient of determination ($R^2$) have been proposed to judge the explained variation of survival models. Our interest is the evaluation of prediction capability of a genetic survival model without the availability of new independent data. On that matter, the $R^2$ version of the 0.632 bootstrap estimator for survival data (Gerds et al. 2007, Biometrics 63:1283–1287) based on Brier's score is helpful. Also, the $R^2$ derived from the Schoenfeld residuals has shown to be best for assessing goodness of fit in genetic survival models (Muller et al. 2008, Genet Epidemiol 32:574–585). We propose using the 0.632 bootstrap estimator to evaluate prediction errors based on the Schoenfeld residuals. In short, each bootstrap sample is split into training and testing data. The expected values of a genetic variant obtained from modeling the training data plays as the prediction rule for the true values of the variant for individuals failing in the testing data. The method was tested through simulations of real settings of genetic survival data with a biallelic variant. The $R^2$ versions based on Brier's score and Schoenfeld residuals showed good agreement with their corresponding valid $R^2$ computed with a much larger data set. Thus, both estimators proved good performance for evaluation of prediction capability, although the estimator based on Schoenfeld residuals handed higher $R^2$ values, which is better for discrimination between genetic survival models.

**219**

**Exploring the Uptake and Effectiveness of KRAS Testing for Metastatic Colorectal Cancer (mCRC): Can We Improve Health Outcomes and Influence Decision-making in Personalized Medicine?**

Katrina A.B. Goddard (1)
(1) Kaiser Permanente Center for Health Research

Abstract presented on behalf of the CERGEN Study Team KRAS is a pharmacogenomic test used to help clinicians make treatment decisions for patients with metastatic colorectal cancer (mCRC). KRAS testing allows oncologists to use anti-EGFR therapy on patients who will benefit the most, thereby increasing treatment effectiveness and minimizing adverse events. Since its introduction in 2008, KRAS testing has been rapidly adopted in numerous clinical settings. EMR data can help evaluate the use of genomic applications in clinical practice and associated patient outcomes. This study spans 7 managed-care organizations with a combined population of 5 million covered lives; the study's size will provide critical real-world information on the impact of KRAS testing in clinical care. This study includes 800 mCRC cases diagnosed between 2004 and 2009. We will conduct KRAS genotyping, using archival specimens, for all subjects who did not receiving testing as part of their clinical care. We will present our findings on characteristics of patients, providers, and health systems that are related to diffusion of KRAS testing. We will also compare overall survival for patients who do or do not receive KRAS testing as part of their clinical care among patients treated in these community practices. This study will serve as a model for future evaluations of the ability of genomic applications to improve health outcomes through personalized medicine.

**220**

**Modeling the Genetic Burden in Multiple Sclerosis Predicts a Higher Genetic Load of Risk Variants in Multicase Compared to Single Case Families: A Tool for Future Study Experiment in MS.**

Pierre-Antoine Gourraud (1), Joseph M.c. Elroy (1), Cailler Stacy (1), Britt Johnson (1), Adam Santaniello (1), Stephen L. Hauser (1), Jorge R. Oksneberg (1)
(1) UCSF

**Background**: Multiple sclerosis (MS) is a multifactorial neurologic disease characterized by modest but tractable heritability. Genome Wide Association Studies (GWAS) have identified and validated multiple polymorphisms in approximately 16 genes associated with susceptibility.
**Methods**: The corresponding genetic burden and its transmission was analyzed in 1213 independent MS families encompassing both sporadic (single-case) and multi-case families. In a weighted log-additive integrative approach termed MS Genetic Burden (MSGB), we accounted for the well-established genetic variants from previous association studies and meta-analysis.
**Results**: MSGB analysis demonstrated a higher aggregation of susceptibility variants in multi-case, compared to sporadic, MS families. The aggregation of non-MHC SNPs depended neither on gender nor on the presence of HLA-DRB1∗15:01 alleles. While a greater MSGB in siblings sibs of MS patients was associated with an increased risk of MS ($OR = 2.1$, $p = 0.001$), ROC curves of MSGB differences between probands and sibs (AUROC 0.57) show that case-control status prediction of MS cannot be achieved with the currently available data.
**Discussion**: The primary interest in the MSGB concept resides in its capacity to integrate cumulative genetic contributions to MS risk. Multi-affected families remain an invaluable resource for advancing the understanding of the genetic architecture of complex autoimmune traits.

**221**

**A Bayesian Model For Computation of Genetic Risk of Complex Traits**

Stephen W. Hartley (1), Martin H, Steinberg (2), Paola Sebastiani (1)
(1) Department of Biostatistics, Boston University School of Public Health
(2) Department of Medicine, Boston University School of Medicine

While simple genetic tests exist for monogenic traits such as Huntington's Disease, useful prognostic tests have not yet been developed for more complex polygenic traits. Our objective is to create a software utility that will generate a prognostic classifier for a phenotype of interest via naive Bayesian networks, based on a given case/control dataset. The software utility is a fast and efficient command-line utility designed to rapidly generate optimal Naive Bayesian Classifiers from genome wide data, and test those classifiers' performance. To find the classifiers, the software calculates the Bayes Factors given a selection of inheritance models. The SNPs are then sorted by the maximum Bayes Factor of each SNP over all such models. This ordering is used to produce a series of hierarchical naive Bayesian classifiers. Training and test sets—mutually exclusive subsets of the original discovery set—are repeatedly generated at random to test the specificity, sensitivity and robustness of these classifiers. In minutes, the software utility can create a hierarchical series of classifiers based on a million SNPs and run thousands of iterations of accuracy tests on randomly-generated test sets. We will demonstrate the usefulness of our methods by generating a genetic risk model that can predict the risk for complication of leg ulcer in patients with sickle cell anemia using genome wide data of 1400 sickle cell anemia patients from the Cooperative Study of Sickle Cell Disease.

**222**

**WITHDRAWN**

**223**

**Gene-gene and Gene-environment Interactions Involving GWAS-identified Loci Unlikely to Drastically Improve Breast Cancer Risk Prediction**

Peter Kraft (1), Hugues Aschard (1), Jinbo Chen (2)
(1) Harvard School of Public Health(2) University of Pennsylvania School of Medicine

The discriminatory ability of published genetic risk profiles for breast cancer incorporating GWAS-identified markers is modest. However, these profiles have not incorporated gene-gene or gene-environment interactions (departures from additivity on a log relative risk scale). We investigated the increase in discriminatory ability from including gene-gene and gene-environment interactions. We considered a wide range of potential risk models incorporating interactions with 17 known breast cancer risk markers; all models were constrained to be consistent with the published marginal effects of these markers. Except for a few extreme models, incorporating interactions led to at most a modest gain in discrimination (change in $C<5\%$). We also examined differences in discriminatory ability for genetic risk scores across strata defined by non-genetic factors in 1,145 postmenopausal breast cancer cases and 1,142 controls from the Nurses Health Study. The discriminatory ability of genetic risk profiles decreased with increasing age ($C = 64\%$ for women 44–61 and 58% for women 71 to 83; $p = 0.11$) and with increasing Gail score ($C = 62\%$ for women below median risk and 57% for women above median risk; $p = 0.04$). Our results suggest that incorporating interactions involving known risk markers is unlikely to improve the average discriminatory ability of genetic risk profiles. However, genetic profiles may have slightly greater discrimination in some strata defined by non-genetic factors.

**224**

**Additional Value of a 31 SNP Risk Score in Predicting Incident type 2 Diabetes in non-diabetic Women**

Nina P. Paynter (1), Daniel I. Chasman (1), Aruna D. Pradhan (1), Paul M. Ridker (1)
(1) Brigham and Women's Hospital

**Background**: Genetic risk scores with up to 20 variants have not improved prediction of type 2 diabetes mellitus (T2DM) over clinical risk factors.
**Methods**: 21,710 initially healthy Caucasian women without T2DM and with hemoglobin A1c $<6.5\%$ were followed for a median of 13.3 years for incident T2DM as part of the Women's Genome Health Study. In addition to clinical risk factors ascertained at baseline (age, blood pressure, lipids, hemoglobin A1c, body mass index, exercise, smoking, family history, and hormone use), we also collected information on 31 non-correlated single nucleotide polymorphisms (SNPs) with a previously published association with either T2DM, fasting glucose or hemoglobin A1c. A genetic risk score was calculated by summing the risk alleles from the 31 SNPs. Predictions of incident T2DM were generated from Cox proportional-hazards models.
**Results**: After adjustment for clinical risk factors, the genetic risk score increased the hazard of incident T2DM 4% per risk allele (hazard ratio 1.04, 95% confidence interval 1.03–1.06). However, the addition of the genetic risk score did not improve prediction of T2DM over clinical risk factors as measured by discrimination (c-statistic 0.877 vs. 0.879, $p = 0.121$) or reclassification (net reclassification improvement 0.2, $p = 0.76$).
**Conclusions**: An expanded genetic score was independently associated with incident T2DM but did not improve prediction over clinical risk factors.

**225**

**Representing Epistasis Effects of Susceptibility SNPs in a Risk Prediction Model for Lung Cancer**

Olaide Y. Raji (1), Stephen W. Duffy (2), George Xinarianos (1), Triantafillos Liloglou (1), Robert P. Young (3), John K. Field (1)
(1) Liverpool UK-CR Centre, University of Liverpool
(2) Wolfson Institute of Preventive Medicine, Queen Mary University of London
(3) Department of Medicine, Auckland Hospital

The proliferation of genomic research has led to identification of many genetic variants predisposing individuals to cancer risk, most importantly Single Nucleotide Polymorphisms (SNPs) from the Genome-Wide Association

Studies (GWAS) or candidate gene approach. SNPs are often incorporated into risk models independently; thereby, omitting the effect of correlations among different SNPs on predicted risks. As each SNP causal variant is known to contribute only a small effect to disease risk, there is a need to simultaneously incorporate more than one SNP into risk models if future risk predictions are to benefit from addition of genetic/molecular makers of disease.
Data from two independent lung cancer studies were used to investigate the contributions of factors representing the interactive (epistasis) effects of the SNPs in a risk prediction model for lung cancer. These factors were derived through methods such as Multi-factorial Dimensionality Reduction (MDR) and partial least squares regression. Additionally, SNPs on a similar biological pathway were identified and combined to form factors representing SNPs that are likely to function as a unit and share common biological role in the risk model.
Using the recently developed relative utility curve, the best epistasis factor was identified among the best SNP combinations by comparing the clinical utility of the various versions of the risk models incorporating different epistasis factors.

**226**
**Improving Prediction in Genetic Models Based on Additive Effects from Dependent Variants**
Kerby Shedden (1), Ming-Chi Hsu (1), Ji Zhu (1), Yan V. Sun (2)
(1) University of Michigan, Department of Statistics
(2) University of Michigan, Department of Epidemiology

For several common human diseases, genome-wide association studies (GWAS) have resulted in strong evidence that genetic risk factors for the disease exist in multiple distinct regions of the genome. Within each such region, one or more genetic variants may confer risk, but efficiently summarizing the information contained in multiple dependent variants is challenging. To date, only a fraction of known genetic heritability for common diseases has been explained by multigenic models. Here we focus on the potential to improve prediction by incorporating information from multiple informative SNPs in a region. Due to the strong dependencies within a region, traditional regression analysis applied to all variants in all regions performs poorly, but several regularized regression analysis alleviate this breakdown due to variance. As a baseline for comparison, we considered a widely-used approach that selects a single variant from each region. We find that this method is competitive with the other methods only when there truly is only a single informative variant per region, while the methods incorporating information from multiple dependent SNPs improve the prediction substantially when multiple informative SNPs exist. Putting this into a larger context, we discuss possible mechanisms by which closely-linked variants may have independent predictive value, or including the possibility of distinct mechanistic roles that they capture distinct information about untyped variants.

**227**
**Incorporating Genetic Ancestry into Interaction and Risk Prediction Models**
Nadia Solovieff (1), Clinton T. Baldwin (2), Martin H. Steinberg (3), Thomas T. Perls (3), Paola Sebastiani (1)
(1) Department of Biostatistics, Boston University, Boston, MA
(2) Center for Human Genetics, Boston University, Boston, MA
(3) Department of Medicine, Boston University, Boston, MA

The era of genome wide association studies (GWAS) has produced a wealth of genetic data and has identified numerous disease associated genes. The effect of population stratification on genetic association is well established and numerous methods exist for detecting and adjusting for it. Generally, investigators focus on removing the effect of population stratification when performing association tests or building prediction models. However, SNPs that have a different association in different ethnic groups are not considered or detected by methods typically used in GWAS. For example the allele frequencies of rs405509, a SNP near APOE known to be associated with exceptional longevity and age related diseases, vary across ethnic groups in Europe. The association between this SNP and exceptional longevity also varies among ethnic groups with odds ratios ranging from 0.38 in Italians, 1 in Ashkenazi Jewish to 1.09 in the Irish. The association is only detected when considering an interaction model between ethnic ancestry and the SNP. We present various models of interaction between SNPs and ancestry using both logistic regression and directed graphical models. We then explore methods of incorporating ancestry into genetic risk prediction models to improve the sensitivity and specificity of case control classification. We evaluate these models using both simulated data and real data from the New England Centenarian Study.

**228**
**Genome Wide Profiling: Joint Polygenic Profiles for LDL and Cognitive Function**
Cornelia M. van Duijn (1), Ayse Demirkan (1), Najaf Amin (1), Ben A. Oostra (1), Yurii S. Aulchenko (1), A Cecile J.W. Janssens (1)
(1) ErasmusMC

While association studies show a key role of lipid genes in Alzheimer disease, there is no evidence for a role of this gene family in cognition. We tested the hypothesis that a large number of lipid genes, each with a small contribution, explain part of the heritability of cognitive function according to a polygenic model. We used a genome wide profiling approach in which profiles for circulating triglyceride, HDL-, LDL-, total cholesterol levels were defined based on a GWAS of lipids in the ENGAGE consortium ($n = 17{,}798$–$22{,}256$). Different profiles were created with increasing cut-off scores for $p$-values for the selection of SNPs. All profiles were added to the baseline model including gender, age and education and associated to cognitive function in the target study, the Erasmus Rucphen Family study ($N = 2300$). Only the genetic risk scores based on LDL-cholesterol significantly explained variance in cognition according to a polygenic model: the more (non-significant) SNPs were included in the score derived from the discovery set, the higher the percentage of variance was explained. The risk score including SNPs

with a $p$-value $<0.90$ in the ENGAGE lipid meta-analysis explained 1.6% of the variance in intelligence ($p = 1.33*10^{-5}$) and 1.2% of the variance in executive function ($p = 4.12*10^{-4}$). Although the variance explained is small, our study shows a joint polygenic origin of lipid metabolism and cognitive function.

**229**

**A Clustered Optimal ROC Curve Method for Family-based Genetic disease Prediction**

Chengyin Ye (1), Jun Zhu (2), Qing Lu (3)
(1) Department of Epidemiology, Michigan State University; Institute of Bioinformatics, Zhejiang University
(2) Institute of Bioinformatics, Zhejiang University
(3) Department of Epidemiology, Michigan State University

Risk prediction that capitalizes on the emerging genetic findings holds great promise to improve public health and clinical care. However, statistical methods for genetic risk prediction research, in particular for correlated data, are still lacking. We propose a clustered optimal ROC curve (CORC) method here, to build predictive genetic tests using data from family-based genetic research. For the proposed method, we have extended the conventional optimal ROC curve method to handle multiple genetic markers, taking sample correlation into consideration, and implemented a forward selection algorithm to allow for high-dimensional data. We have evaluated the CORC method in both simulations and a real-data application, showing that the method performed better than the existing methods under various pedigree structures and underlying disease models. In the real-data application, we have applied the method on the large scale International Multi-Center ADHD Genetics Project dataset and formed a predictive genetic test for conduct disorder. The test reached a low to medium classification accuracy with an AUC value of 0.6908.

**230**

**FTT—Adaptive, Fast and Robust Association Testing**

Mathew J. Barber (1), Matthew Stephens (1)
(1) University of Chicago

We introduce a general three-stage analytical strategy for the statistical problem of testing for a difference in means of a continuous response between two or more groups. The strategy can conveniently be summarised as ''(i) Fit, (ii) Transform, (iii) Test'', or FTT. The intuitive motivation for this strategy is that step (i) aims to produce a more powerful test by adapting the test based on the observed response data to learn about the true and unknown residual distribution, while the transformation of the quantitative trait in step (ii) aims to ensure that the test is robust, in step (iii), to any misspecification of the parametric model in (i).
We have implemented a specific version of the FTT strategy that transforms to a mixture of two normal distributions. Primarily, we show the considerable power gain that can be achieved over ordinary least squares regression (for example, 65% versus 40% power), especially when sample size is high (say, 500 or larger) and effect size is small (say, 3% variation or smaller). We also show that the type I error for the testing stage can reliably calculated by a chi-squared approximation for reasonable sample sizes (say, 100 or larger).
This strategy enables the automation of the decision as to which test to perform, as well as performing a statistical test. Over and above the gains in power, this aspect is especially useful when association is to be tested between a large set of univariate responses and a large set of univariate predictors.

**231**

**Use of Survival Methods to Identify Loci Underlying Diseases with Variable Age of Onset**

Emmanuelle Bouzigon (1), Hugues Aschard (1), Florent Monier (1), Florence Demenais (1)
(1) INSERM, U946, Paris, France

Asthma is a heterogeneous disease and age of onset is one the simplest features that can be used to differentiate asthma phenotypes. Various approaches can be used in genetic association analyses to take into account a variable age of onset of disease. One way is to stratify affected subjects according to age of onset, as previously done by ordered-subset regression analysis which showed an effect of 17q21 locus in early-onset asthma (Bouzigon et al. N Engl J Med 2008;359:1985–94). Our goal was here to identify genetic factors underlying time to onset of asthma by using an approach based on martingale residuals (MR) from Cox survival model. We conducted a genome-wide association study (GWAS) of time to asthma onset in 1,511 subjects (750 asthmatics and 1085 non-asthmatics) from the French EGEA study genotyped by Illumina 610K chip. We compared the results obtained by applying the MR method to 1) affecteds only or 2) both affected and unaffected subjects. Analysis of asthmatics only led to detect the 17q21 locus as before plus 7 loci at $P?10^{-5}$. When both asthmatics and non asthmatics were examined, we found 4 loci associated with time to asthma onset at $P?5 \times 10^{-7}$: one locus within IL33 gene recently identified by a meta-analysis of asthma GWAS (Gabriel Consortium) plus three new loci on chromosomes 1, 2 and 15. Thus considering time to disease onset can be a powerful approach to identify new loci underlying complex diseases such as asthma. Funded by European Commission (GABRIEL)

**232**

**Bayesian Centroid Inference for Genome-Wide Association Studies**

Luis E. Carvalho (1)
(1) Boston University

Genome-Wide Association Studies (GWAS) attempt to identify a (possibly small) subset of single nucleotide polymorphisms (SNPs) from a large number of measured candidates that are associated with a specific observable trait. Identification of associated genetic variants is particularly hard due to very large genotype sizes in comparison to case-control group sizes. We formally frame this problem as Bayesian variable selection in a logistic regression model with spike-and-slab priors in the coefficients. To help overcome the curse of dimensionality, we further explore a co-dependency structure between SNPs by setting an Ising hyper-prior on the space of possible SNP associations to trait. We introduce and derive centroid and graph centroid estimators and contrast them to the classical maximum a posteriori (MAP) estimators in

this setup. We illustrate this approach with a toy example and a larger simulated dataset based on HapMap. Finally, we offer a few concluding remarks and directions for future work.

**233**
**Multifactorial Diseases: The Polygenic Model Paradigm**
Francoise Clerget-Darpoux (1)
(1) INSERM UMR 669- University Paris-Sud

As an alternative to the monogenic model paradigm, geneticists have switched to a polygenic model to explain susceptibility to multifactorial diseases, assuming no interaction between factors. However, interaction is very likely, although statistical demonstration of its existence is complicated. In particular, the SNPs identified through GWAS do not provide enough power to demonstrate interaction because, most often, these SNPs poorly represent the gene effects in terms of differential risks. This is illustrated in Multiple Sclerosis where the joint information from two SNPs of IL2RA corresponds to a 4-fold differential risk between the least and most at-risk genotypes, whereas it is less than 1.6 for the SNP reported in the literature. For Rheumatoid Arthritis, the differential risk is 2.7 for the *PTPN22* SNP most strongly associated, but reached 4.7 for a combination of 3 SNPs in this gene. For geneticists, leaving the polygenic model paradigm behind is as difficult as it was to abandon the monogenic model. When assuming factors with no interaction and no causal hierarchy, missing information and risk can be quantified. However, in contrast to monogenic diseases, the primary cause(s) of a multifactorial disease may be non-genetic. Most often, the environmental factors are unknown. Nevertheless, exposure to them may be essential to disease initiation. Where this is the case, risk estimations for an individual based on his genomic information alone do not make sense.

**234**
**An Evaluation of Residual-Outcome Regression Analysis and Covariate Effect Adjustment in Genetic Association Studies**
Serkalem Demissie (1), L. Adrienne Cupples (1)
(1) Boston University School of Public Health, Department of Biostatistics

Association studies of genetic or environmental risk factors and complex diseases require a careful assessment of potential confounding factors. Multiple regression analysis, whereby the primary risk factor (exposure), covariates and the outcome are modeled jointly, is an effective method of accounting for a confounding effect. Another approach that has been gaining acceptance in genetic and some epidemiologic studies is a two-stage (residual-outcome) approach, in which residuals are first calculated after regressing the outcome variable on confounding factors and then the residuals are modeled as a function of the exposure variable in a simple regression analysis. Using theoretical and simulation analyses of single nucleotide polymorphisms (SNPs) we examined the performance of the two-stage approach as compared with traditional multiple linear regression. When a SNP and a covariate were correlated, the two-stage analysis resulted in under-estimation of the genotype effect, loss of power and conservative Type I error rate. Bias was always toward the null and was directly proportional to the squared-correlation (r2) between the exposure variable and the covariate. For example, for $r^2 = 0.0$, 0.1 and 0.5, the two-stage approach resulted in, respectively, 0%, 10% and 50% attenuation in the exposure effect. In summary, in studies involving correlated independent variables, the residual-outcome and the multiple regression analyses yield different results.

**235**
**WITHDRAWN**

**236**
**X Chromosome Association Testing in Genome Wide Association Studies**
Peter F. Hickey (1), Richard Huggins (2), Melanie Bahlo (1)
(1) The Walter and Eliza Hall Institute of Medical Research
(2) The University of Melbourne

The problem of testing for genotype-phenotype association with loci on the X chromosome in genome wide association studies (GWAS) has received surprisingly little attention. There exist several methods in the literature, however, to date there has been no study comparing these methods to one another and it is unclear which is the optimal method. To address this issue we have performed a simulation study under a wide variety of study designs, allele frequencies and disease models to evaluate the performance of eight popular statistical tests from the literature. Our results show that two tests proposed by David Clayton [Biostatistics 2008;9(4):593–600] have high power across a range of study designs and genetic models, however, the optimal test depends on the study design and disease model. We show that two tests implemented in the popular GWAS analysis software PLINK [Am J Hum Genet 2007;81(3):559–575] are vastly underpowered when compared to other existing methods for X chromosome analysis. Our results also highlight that study design features, such as the proportion of females in the study, greatly affect the power to detect associations on the X chromosome. This study will benefit researchers analysing GWAS data, particularly of diseases where the X chromosome is believed to have a role.

**237**
**Bayesian Multivariate Regression with Singular Value Decomposition (BMRSVD) A Way to Identify Association with Multiple Genes and Traits in Modest Sample Sizes**
Soonil Kwon (1), Leslie J. Raffel (1), Kent D. Taylor (1), Mark O. Goodarzi (1), Y.-D. Ida Chen (1), Thomas A. Buchanan (2), Jerome I. Rotter (1), Xiuqing Guo (1)
(1) Medical Genetics Institute, Cedars-Sinai Medical Center
(2) University of Southern California

Genetic association studies often involve many SNPs ($m$), multiple traits, and modest sample sizes ($n$). To evaluate multiple SNPs when $m \gg n$, we previously developed Bayesian Regression with Singular Value Decomposition (BRSVD), which reduces the dimension from m to n by

applying SVD to the design matrix. We now describe *Bayesian multivariate regression with singular value decomposition* (*BMRSVD*) method that incorporates simultaneous analysis of multiple quantitative traits. Utilizing Markov chain Monte Carlo with Gibbs sampler, we constructed a model from posterior densities driven with conjugate priors. Permutation testing was incorporated to generate empirical *p*-values. We applied BMRSVD to a hypertension family cohort, in which blood pressure and insulin resistance cosegregate. Prior analysis had identified 4 blood pressure genes, ADRB2, NPPA, SCNN1A, and NOS3, to be associated with insulin resistance, derived by glucose clamp (M). To explore genes contributing to both SBP and M, we applied BMRSVD to 80 independent offspring with complete phenotype and genotype data. Association was tested for 227 SNPs from 109 candidate genes simultaneously. SNPs in 23 genes (IL4R, ADRB2, ICAM1, CCR5, IL5RA, IL6, SCYA11, GNB3, NOS3, NPPA, F2, ACE, GPRK2L, MMP3, PDE4D, REN, CCL2, IL18, CRP, CAV1, CD36, GCG, and SORCS1) were significantly associated with both traits. This demonstrates that BMRSVD is effective in identifying genes for multiple phenotypes.

**238**

**Case-Control Association Tests that Assume Hardy-Weinberg Equilibrium**

MyoungKeun Lee (1), Eleanor Feingold (1)
(1) University of Pittsburgh

To test association between a SNP and a binary trait in a case-control or cohort study, the standard analysis is a $X^2$ test or logistic regression. However, alternative tests have been proposed that attempt to improve power by assuming that either controls or the population are in Hardy-Weinberg Equilibrium (HWE). We hypothesized, however, that such tests might behave poorly in the context of a GWAS, because they might preferentially pick out SNPs that are out of HWE by random chance or due to genotyping error. We tested this hypothesis using both single-SNP and genome-wide simulations based on real data.

For the single-SNP simulations, we considered four different HWE scenarios: (1) when HWE holds in both cases and controls, (2) when HWE does not hold in cases but does in controls, (3) when HWE does not hold in controls but does in cases, and (4) when HWE does not hold in either cases or controls. For genome-wide simulations we used real GWAS data, but simulated the binary phenotype.

As expected, when HWE holds in both cases and controls, the standard test and the test that assumes HWE are equivalent, except for variation attributable to type I error. When HWE fails to hold in either cases or controls or both, the test that assumes HWE in controls detects this HWE departure and can therefore find a ''case-control difference'' even if there is not an allele frequency difference or a genotype frequency difference. We quantify the effect of this problem in the genome-scan context.

**239**

**Performance of a Latent Variable Framework for the Analysis of Genes and Biomarkers**

Won Lee (1), David Conti (1)
(1) University of Southern California, Department of Biostatistics

As part of the Pharmacogenetics of Nicotine Addiction Treatment Consortium we are studying how pharmacotherapies for smoking cessation vary across individuals. A pharmacokinetic mechanism is implied with CYP2A6, a critical gene in nicotine metabolism. The association of CHRNB2 with quit rates suggests a pharmacodynamic response. Along with genetic variation, there are associations with biomarkers, such as nicotine metabolite ratio (NMR), which reflects CYP2A6 genetic and environmental influences. In an effort to tease apart the combined effects of genes and biomarkers, we present a latent variable (LV) analysis treating the intermediate biomarker as a flawed measure of the underlying mechanism. The LV framework follows the model from Thomas (Lifetime Data Anal 2007;13(4):565–581): (1) a ''process'' model (PM-LV regressed on a formative exposure), (2) a ''measurement'' model (MM-a reflective indicator regressed on LV), (3) a ''disease'' model (DM-the outcome regressed on LV). Via simulations we evaluate the performance of the approach to identify and estimate effects. We also discuss the sensitivity and specificity of this framework in prediction, and compare performance to standard regressions of measured variables. We also explore cases in which numerous genes are included in PM and how including additional genetic effects in DM impacts performance. Finally, we apply this model to data from two clinical trials of smoking cessation with NMR and SNPs across 57 candidate genes.

**240**

**Association Tests for X-Chromosomal Markers—A Comparison of Different Test Statistics**

Christina Loley (1), Inke R. Konig (1), Andreas Ziegler (1)
(1) Institut fuer Medizinische Biometrie und Statistik, Universitaet zu Luebeck

Genetic association studies were successful to elucidate the genetic background of complex diseases. However, X chromosomal data have usually not been analyzed (Erdmann et al., 2009; Samani et al., 2007), so that information from the X chromosome is neglected. A reason for this is that there is so far no standard instrument for the statistical analysis. While females carry two copies of the X chromosome, males have only one. Thus, for loci not in the pseudo-autosomal region, special tests are required. Another peculiarity is inactivation of one of the female X chromosomes which may be a mechanism of dosage compensation, resulting in equal effects for one copy of the X chromosome in males and two copies in females. But since this X inactivation is not complete one might either neglect it entirely (Zheng et al., 2007). Thus, the influence of one risk allele in males is equated with one in females. Alternatively, inactivation may be considered (Clayton 2008, WTCCC 2007) by treating men with one risk allele like females homozygous for this allele. This contribution evaluates the test statistics regarding type one error rates and power. Hereto we performed extensive simulation studies covering a wide range of different settings.

References:

[1] Clayton D. 2008. Biostatistics 9:593–600.
[2] Erdmann J, et al. 2009. Nat Genet 41:280–282.
[3] Samani NJ, et al. 2007. N Engl J Med 357:443–453.
[4] WTCCC. 2007. Nature 447:661–678.
[5] Zheng G, et al. 2007. Genet Epidemiol 31:834–843.

**241**
**Comparison of Genetic Association Measures to Identify Causal Susceptibility Variants**
Justo Lorenzo Bermejo (1), Alfonso Garcia Perez (2), Kari Hemminki (3), Asta Foersti (3), Andreas Brandt (3), Abigail G. Matthews (4)
(1) Institute of Medical Biometry and Informatics, University Hospital Heidelberg, Germany
(2) Department of Statistics, Spanish Open University
(3) Division of Molecular Genetic Epidemiology, German Cancer Research Center
(4) Ott Laboratory, Rockefeller University NY, USA

Genome wide association (GWA) studies rely on the common-disease common-variant hypothesis: polymorphisms are genotyped and their association with disease is investigated. The identified, indirect associations are assumed to reflect a shared inheritance of genotyped markers and linked causal variants.
Probability values and Bayes factors are the commonest summary of results from GWA studies. Probability values and Bayes factors do not take into account the genetic inheritance of susceptibility variants, which results in familial aggregation of disease. Alternatively, genetic association can be represented by the attributable familial relative risk, a measure of the contribution of susceptibility loci to familial aggregation.
We have used simulation to investigate probability values, Bayes factors and attributable familial relative risks for markers in linkage disequilibrium with causal alleles, according to the recombination rate, allele frequency, penetrance and disease prevalence. Established high-risk variants and recently identified markers were used to illustrate theoretical results.
Our data demonstrate that the representation of genetic association by attributable familial risks instead of probability values and Bayes factors may facilitate the identification of causal variants. Results under several scenarios will be presented, including practical application to imputation-based association analysis.

**242**
**Comparison of Two Multilocus GWAS Methods to Detect Weaker Associations between Genetic Variants and Disease in Case-Control Data**
Radoslav Nickolov (1), Valentin Milanov (1)
(1) Fayetteville State University

One of the challenges of recent genome-wide association studies (GWAS) is the detection of weaker associations, possibly due to rare variants.
Here we present the results of a comparative study of two recently proposed efficient multilocus GWAS methods, V-Bay and BEAGLE, to detect weaker associations between genetic polymorphisms and disease. Both of these methods could have better power to identify rare disease-susceptibility variants than single-marker tests.
The V-Bay method [1] is a variational algorithm for Bayesian hierarchical regression that has been designed to detect weaker associations with low false positives rates. It consists of two modules: a hierarchical regression model with marker class partitioning and a variational algorithm for approximate Bayesian inference. The BEAGLE's method [2] is based on localized haplotype clustering achieved by fitting a variable-length Markov chain model. It uses phased haplotypes inferred by a method implemented in BEAGLE itself [3]. We compare the performance of the above two methods using wide range of simulated data under different realistic scenarios.

References:
[1] Logsdon BA, Hofman GE, Mezey GJ. 2010. BMC Bioinformatics 11:58.
[2] Browning BL, Browning SR. 2007. Genet Epidemiol 31:365–375.
[3] Browning SR, Browning BL. 2007. Am J Hum Genet 81:1084–1097.

**243-WITHDRAWN**
**Measurement Error of Non-genetic Factors—A Curse for Mendelian Randomization Studies?**
Brandon L. Pierce (1), Habibul Ahasn (1), Tyler Vanderweele (2)
(1) University of Chicago
(2) Harvard University

Mendelian randomization (MR) refers to the use of instrumental variable analysis to assess the causality of exposure-disease associations for exposures that have known genetic determinants. In MR studies, data on such a genetic determinant (i.e., an instrumental variable) is analyzed jointly with data on the exposure and the outcome. In theory, MR generates effect estimates that are not biased by unmeasured confounding or "reverse causation". In this work, we show that random error in the measurement of the exposure biases MR estimates away from the null, resulting in an inflated type 1 error rate. This measurement error can be interpreted as a violation of one of the key assumptions for valid MR: the genetic "instrument" cannot influence the outcome independently of the "measured" exposure. Substantial measurement error of non-genetic factors can occur for a variety of reasons, including temporal variation in the exposure, sample handling/storage, errors in measurement protocol, and measurement outside of the relevant "etiologic time window". In addition, we show that in random error in the outcome biases the MR estimate towards the null, while random genotyping error does not bias the MR estimate. We demonstrate our results analytically and using simulations. In conclusion, great caution should be taken when interpreting MR studies with error-prone exposure measures.

**244**
**TiledReg: Software Implementation of Tiled Regression**
Alexa J.M. Sorant (1), Juanliang Cai (1), Heejong Sung (1), Yoonhee Kim (1), Alexander F. Wilson (1)
(1) Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, NIH

The Tiled Regression method of association analysis combines simple linear or logistic regression models, stepwise selection of variables and a staged approach. Groups ("tiles") of potentially correlated SNPs (e.g. defined by hotspots) are first considered separately and discarded if they show no evidence of association with a trait. Stepwise regression is then used to select independent significant SNPs from remaining tilies, which are combined in a chromosome-wide model, where further selection may reduce the number for genome-wise consideration. Quanti-

tative and binary traits can be analyzed, and family structure among individuals can be addressed with generalized estimating equations, assuming residual correlation constant within family groupings or clusters. Fitted models assume additive effects within and among SNP variables, but first-order interactions may be considered in the final model. Tiled Regression methodology has been implemented in TiledReg, a software package written in the freely available R language. Functions are provided for assigning SNPs to hotspot-based tiles, data input, analysis and output of results. The package is structured modularly, so that it may be used as a single program or with user-written functions to allow for alternate tile definition or data format. Statistical parameters controlling details of the procedure can be determined by the user, but defaults are provided for most, combining flexibility and simplicity.

**245**

**Bayesian Variable Selection for Survival Regression in Genetics**

Ioanna Tachmazidou (1), Michael Johnson (2), Maria De Iorio (2)
(1) MRC Biostatistics Unit
(2) Imperial College

Variable selection in regression with very big numbers of variables is challenging both in terms of model specification and computation. We focus on genetic studies in the field of survival, and we present a Bayesian-inspired penalized maximum likelihood approach appropriate for high-dimensional problems. In particular, we employ a simple, efficient algorithm that seeks maximum a posteriori (MAP) estimates of regression coefficients. The latter are assigned a Laplace prior with a sharp mode at zero, and non-zero posterior mode estimates correspond to significant single nucleotide polymorphisms (SNPs). Using the Laplace prior reflects a prior belief that only a small proportion of the SNPs significantly influence the response. The method is fast and can handle datasets arising from imputation or resequencing. We demonstrate the localization performance, power and false positive rates of our method in large simulation studies of dense-SNP datasets and sequence data, and we compare the performance of our method to the univariate Cox regression and to a recently proposed stochastic search approach. In general, we find that our approach improves localization and power slightly, while the biggest advantage is in false positive counts and computing times. We also apply our method to a real prospective study, and we observe potential association between candidate ABC transporter genes and epilepsy treatment outcomes.

**246**

**Cure-rate Models for Time-to-event Phenotypes in Genetic Association Studies**

Yildiz E. Yilmaz (1), Jerald F. Lawless (2), Irene L. Andrulis (1), Shelley B. Bull (1)
(1) Samuel Lunenfeld Research Institute
(2) University of Waterloo

Time-to-event phenotypes typically arise in genetic association studies of complex traits in which individuals are followed through time and disease status is assessed longitudinally. When the population of interest consists of susceptible and non-susceptible individuals, and a significant proportion of individuals will never experience the disease event, the conventional Cox proportional hazards model is inappropriate because it assumes that all individuals continue to be at risk as followup continues. Furthermore, the genetic factors associated with susceptibility may differ from those that affect time to event in those at risk. As an alternative, we describe a cure-rate model in which time to event in the subgroup of susceptible individuals is modeled by a proportional hazards regression model with piecewise constant baseline hazard function and the probability of being disease-free in the long-term is modelled by a logistic model. Using maximum likelihood methods for estimation and inference, we apply the proposed method in a large breast cancer cohort to assess the association of molecular genetic prognostic factors with long-term disease-free survival and with time to disease recurrence. Our results illustrate the additional information that can be obtained via cure-rate models.