ABSTRACTS FROM THE

# SIXTEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

**1**

**Methods for dense SNP marker analysis for genetic linkage analysis with application to rheumatoid arthritis**

C.I. Amos(1), W.V. Chen(1), B. Peng(1), X. Liu(2), D. Zhu(1), S. Shete(1), K. Siminovitch(2)
(1) U.T. M.D. Anderson Cancer Center, Houston, TX.
(2) Mount Sinai Hospital, Toronto

Analysis of densely located SNPs increases information for linkage studies, but can increases the false positive linkage evidence when SNPs are falsely assumed to be in linkage equilibrium. We compared several different approaches for dense SNP marker analysis when parental genotypes are not available. We simulated data using the SNPs that are present on the Affymetrix 100K platform. We performed extensive simulation studies using a model in which a single genetic locus influencing susceptibility to disease was present. The first approach for removing effects of linkage disequilibrium was an 'interleaving' method (Bacanu, Genetic Epidemiology 29:195–203). For this method, we divided the data into 10 disjoint sets of markers and then performed analysis using SNPLINK to remove any markers showing excess LD (R2>0.05). Results from these 10 runs were then averaged and the empirical p-value obtained at each point. Alternatively, we used Merlin (Am J Hum Genet. 77:754–67), an approach which accounts for LD by 'clustering' tightly linked markers. Results of both these analyses led to diminished Z-scores compared with those obtained without allowing for LD and eliminated false positive results. On average, the runs using Merlin were slightly less powerful and required extensive computational resources but interleaving showed an excess of false positive findings in some situations. Application in linkage analysis of sib pairs with rheumatoid arthritis supported linkage to 6p.

**2**

**Turbo Genomic Control**

W.J. Astle(1), C.C. Holmes(2) and D.J. Balding(1)
(1) Section of Biostatistics, Department of Epidemiology and Public Health, Imperial College London.
(2) Department of Statistics University of Oxford.

In the analysis of population association studies, Genomic Control[1] (GC) adjusts the Armitage test statistic to correct the type I error for the effects of population substructure, but its power is often sub-optimal. Turbo Genomic Control (TGC) generalises GC to incorporate co-variation of relatedness and phenotype, retaining control over type I error while improving power. TGC is similar to the method of Yu et al.[2], but we extend it to binary (case-control) in addition to quantitative phenotypes, we implement improved estimation of relatedness coefficients, and we derive an explicit statistic that generalizes the Armitage test statistic and is fast to compute. TGC also has similarities to EIGENSTRAT[5] which is a new method based on principle components analysis.

The problems of population structure[3] and cryptic relatedness[4] are essentially the same: if patterns of shared ancestry differ between cases and controls, whether distant (coancestry) or recent (cryptic relatedness), false positives can arise and power can be diminished. With large numbers of widely-spaced genetic markers, coancestry can now be measured accurately for each pair of individuals via patterns of allele-sharing. Instead of modelling subpopulations, we work instead with a coancestry coefficient for each pair of individuals in the study.

We explain the relationships between TGC, GC and EIGENSTRAT. We present simulation studies and real data analyses to illustrate the power advantage of TGC in a range of scenarios incorporating both substructure and cryptic relatedness.

[1] Devlin B. and Roeder K., Genomic control for association studies. Biometics 55(4) December 1999.
[2] Yu J. *et al.*, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38(2) February 2006.
[3] Clayton D.G. *et al.* Population structure, differential bias and genomic control in a large-scale case-control association study. Nature Genetics, 37(11) November 2005.
[4] Voight B.J. and Pritchard J.K., Confounding from cryptic relatedness in case-control association studies. Public Library of Science Genetics, 1(3) September 2005.
[5] Price A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8) (August 2006).

**3**

**Distinguishing true from false positive results in genome-wide association studies: quality control and statistical significance**

JH Barrett (1), MM Iles (1), DT Bishop (1), R Parisi (1), MD Tobin (2), JR Thompson (2), NJ Samani (2), AS Hall (1)
(1) University of Leeds, UK, (2) University of Leicester, UK

Genome-wide association (GWA) studies give rise to almost unprecedented levels of multiple hypothesis testing, and there is a need to interpret statistical significance with care. Error in genotype calling is an additional major source of false positive results, which cannot be addressed by increasing the sample size or adjusting significance levels. Here we present our experience of developing quality control (QC) criteria to eliminate the likelihood of results being due to calling error.

Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20262

Criteria are developed and illustrated using data from the Wellcome Trust Case Control Consortium cardiovascular study, based on the 500k Affymetrix SNP chips. Apparently significant associations are much more likely to arise in SNPs with higher, but still low, levels of missing (uncertain) genotype calls. To illustrate, when less than 14% of SNPs overall have>2% of calls missing, this frequency rises to 30% and 57% in SNPs showing evidence of association at P < 10-3 and 10-4 respectively. Even after applying very strict QC criteria (based on the frequency of missing genotype calls, departure from Hardy-Weinberg equilibrium and other measures) and levels of statistical significance, robust disease associations remain. Further preliminary QC findings will be presented from a GWA study of melanoma currently in progress which is based on the Illumina 317k chip.

**4**

**MASEL: marker selection for linkage analysis with high density SNP maps in large pedigrees**

C. Bellenguez (1,2), C. Ober (3) and C. Bourgain (2,1)
(1) Univ. Paris Sud, IFR69, UMR_S535, Villejuif F-94817, France
(2) INSERM, U535, Villejuif F-94817, France
(3) Department of Human Genetics, The University of Chicago, USA

SNP maps are becoming the gold standard today, even for linkage analyses. However, because they are much closer, SNPs present important linkage disequilibrium (LD), which biases classical nonparametric multipoint analyses. This problem is even stronger in population isolates where LD extends over larger regions with a more stochastic pattern. We investigate the issue of linkage analysis with a 500K SNP map in a large and inbred 1840-member Hutterite pedigree, phenotyped for asthma. Using an efficient pedigree breaking strategy, we first identified linked regions with a 5cM microsatellite map, on which we focused to evaluate the SNP map. The only method that models LD in the NPL analysis is limited in both the pedigree size and the number of markers (Abecasis and Wigginton, Am.J. Hum.Genet. 77:754–767, 2005) and therefore could not be used. Instead, we developed an algorithm, MASEL, that iteratively selects SNPs with LD lower than a threshold defined by the user. Each SNP is chosen according to two criteria: maximization of mean heterozygosity and minimization of inter-SNP distance variance. Weights can be applied to each criteria. Null simulations are performed to control that Zlr calculated with the SNP sets are not falsely inflated. The method efficiently identifies a limited number of SNPs, leading to similar results as compared with the dense microsatellite map, in terms of both information content and linkage detection.

**5**

**Rapid and Accurate Haplotype Phasing and Missing Data Inference for Whole Genome Association Studies using Localized Haplotype Clustering**

B.L. Browning(1, 2), S.R. Browning(1)
(1) Department of Statistics and (2) Discipline of Nutrition, The University of Auckland, New Zealand

Whole genome association studies present many new statistical and computational challenges due to the large quantity of data obtained. One of these challenges is haplotype inference: methods designed for small data sets from candidate gene studies do not scale well to the large number of individuals genotyped in whole genome association studies. We present a new method and software for inference of haplotype phase and missing data that can accurately phase data from whole genome association studies. Our method is based on fitting a localized haplotype cluster model to initial estimates of haplotype phase. The localized haplotype cluster model is extended to give a hidden Markov model, from which revised estimates of haplotype phase can be sampled. This process is iterated, with most likely haplotype phase inferred at the last iteration. Our method is compared with existing haplotype inference methods, including fastPHASE and HaploRec, on real and simulated data sets with thousands of genotyped individuals. We find that our method outperforms existing methods in both speed and accuracy for large data sets with thousands of individuals and densely spaced genetic markers, and we use our method to phase a real data set of 3002 individuals genotyped for 490,032 markers in 3.1 days computing time, with 99% of masked alleles imputed correctly. Our method is implemented in the Beagle software package which is available at http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html.

**6**

**Bias Reduction in Genome-wide Association Studies with Time-to-Event Phenotypes**

SB Bull (1,2), X Xie (1), L Faye (1,2), L Sun (2,3), AD Paterson (2,3)
(1) Samuel Lunenfeld Research Institute, Toronto (2) Dept of Public Health Sciences, U of Toronto (3) Hospital for Sick Children, Toronto, Canada

To date, genetic analysis has been based largely on trait data collected cross-sectionally, and thinking around GWA studies has focused on the efficiency of case-control designs due to advantages of cost and duration. There are however recent examples of studies conducted by genotyping existing cohorts with longitudinal measurements, which can track phenotype changes through time and detect the points of development of multiple phenotypes. While statistical criteria are necessary to control the occurrence of false positive findings, genome-wide selection of SNP markers with large test statistics introduces upward bias into effect estimates for the genetic associations detected. Motivated by a study of the genetics of complications of type I diabetes in a rigorously followed cohort of participants in the Diabetes Control and Complications Trial (DCCT/EDIC), we develop analytic expressions for expected bias in the hazard ratio (HR) of a proportional hazards survival model which coincide remarkably with the patterns observed in bootstrap samples of the DCCT/EDIC study individuals. Building on work in bias-reduced effect estimation in genome-wide linkage, we extend statistical resampling techniques to single-SNP and multiple-SNP association settings and evaluate them via simulations based on observed LD in DCCT/EDIC SNP data. Replication study sample size requirements determined by bias-reduced estimates are substantially larger than those based on the original estimates.

**7**

**Multiple Myeloma, Chronic Lymphocytic Leukemia, Non-Hodgkin Lymphoma: Evidence for Overlapping Genetic Etiologies**

NJ Camp(1), TL Werner(2), LA Cannon-Albright(1)
(1)Depts of Biomedical Informatics&(2)Oncology, University of Utah, USA

The relationship of MM to other hematological and solid cancers is unclear. Identifying genetic overlap will aid design of genetic studies and power of subsequent genetic analyses. Here we perform familiality analyses using the Utah Population Database (UPDB) to identify potential genetic etiological overlap. This powerful resource includes a genealogy linked to all cancer records diagnosed or treated in Utah since 1966. We used 2 million individuals in the UPDB with genealogical data in our analyses. We performed familial relative risks (FRR) for 1st, 2nd and 3rd degree relatives. The benefit of analyzing beyond 1st degree relatives is important because shared environment decreases for distant relatives and familiality can more readily be interpreted as evidence for a genetic component. Analysis of 3rd degree relatives has not previously been done. We investigated MM, Hodgkin Lymphoma (HL), Non-Hodgkin (NHL), NHL B cell (NHLB), Leukemia, Acute Myeloid (AML) and Chronic Lymphocytic (CLL). MM, NHL, NHLB and CLL exhibited significantly increased FRR in 1st, 2nd and 3rd degree relatives (NHLB driving NHL). MM and NHLB were significantly increased in 1st and 2nd degree relatives of CLL cases, and vice versa, suggesting etiological overlap. Also, prostate cancer was significantly increased in 1st, 2nd and 3rd degree relatives of MM, NHLB and CLL, and in the cases themselves. These results, in addition to the similarity of characteristics such as onset age, gender bias and survival, suggest genetic etiological overlap for MM, CLL and NHLB.

**8**

**A Genome-Wide Association Study of Skin Pigmentation in a South Asian Population**

Tony Dadd (1), Renee P. Stokowski (2), P.V. Krishna Pant (2), Amelia Fereday (1), David A. Hinds (2), Carl Jarman (1), Wendy Filsell (1), Rebecca S. Ginger (1), Martin R. Green (1), David R. Cox (2), Frans J. van der Ouderaa (1)
(1) Unilever Corporate Research, Colworth Park, Bedford, UK
(2) Perlegen Sciences, Mountain View, California, U.S.A.

We conducted a multi-stage, genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with natural skin colour variation in a population of South Asian descent. Skin reflectance measurements, known to correlate with melanin content in skin, were taken from sun-protected sites on volunteers resident in the UK, and the extremes of the resulting distribution used to create a threshold-defined case-control phenotype. A pooled genotyping experiment on 1.6 million SNPs, using 571 samples matched for inferred ancestry using the structure program, was followed by two stages of individual genotyping with 737 and 231 samples respectively. Genetic association with phenotype was tested within a logistic regression framework using likelihood ratio tests; covariates were included to account for gender and population structure as measured by principal components analysis. Non-synonymous SNPs in SLC24A5, SLC45A2 and TYR demonstrated highly significant and replicated associations, which collectively account for a large fraction of the variation in natural skin colour in this South Asian population.

**9**

**The "bunny ears" syndrome: covariates with confounding genetic effects**

M. de Andrade (1), B.L. Fridley (1), S.T. Turner (2)
(1) Biostatistics, and (2) Nephrology and Hypertension, Mayo Clinic, USA

In quantitative trait linkage analyses, investigators may sometimes observe two linkage peaks in the same genomic region. It is usually thought to be the consequence of two causative loci in the same region. However, this may due to a hidden effect of a covariate. We call this the "bunny ears" syndrome because the linkage plot looks like "bunny ears". This syndrome was observed by us in two distinct data sets, one using real data from GENOA Phase II sibships and the other using GAW 13 simulated data. In the simulated data, the trait of interest was systolic blood pressure (SBP) and the covariate was height. Height was simulated to have an impact in the level of SBP, and also had a major gene effect in the same region as the major gene for SBP. When the linkage analysis was performed for SBP without adjusting for height, two linkage peaks appeared causing the "bunny ears" syndrome. After adjusting for height, the bunny ears disappeared and only the peak due the SBP major gene remained. In the real data, this syndrome was also observed when performing a linkage analysis with brain atrophy (BA) as the trait of interest and total intracranial volume (TIV) as the covariate. BA was calculated as the difference between brain volume and TIV. We observed the "bunny ears" syndrome in the chromosome 17p region when the linkage analysis was performed without adjusting for TIV. After adjusting for TIV the bunny ears disappeared and only one peak remained for BA. These results emphasize the importance of adjusting for covariates that may otherwise exert confounding genetic effects.

**10**

**Using a Propensity Score Versus a Mendelian Risk Prediction Score to Increase Genetic Homogeneity for Association Analyses**

BQ Doan(1,2) G Parmigiani(2) A Chakravarti(1) JE Bailey-Wilson(3)
(1)IGM/JHMI (2)Onc Biostat/JHMI (3)IDRB/NHGRI/NIH

When a significant linkage peak is identified, further follow-up is warranted to scrutinize the region for a putative disease gene. However, genetic and disease heterogeneity can reduce the power to identify genes of modest effects for complex traits. We previously showed that the use of a propensity score(PS) in a nonparametric linkage analysis of lung cancer can identify a more genetic subset by discriminating among

those who may be affected due to environmental exposures (higher scores), as its use increased the LOD score from 1.0 to 3.9. We have since developed a Mendelian risk prediction model for lung cancer. Because the risk prediction score (RPS), the probability of having a deleterious mutation, is highly dependent upon family history, its use as a covariate does not discriminate as well among siblings as the PS. However, the use of such scores can possibly identify a more homogeneous subset for follow-up association analyses. Consequently, we have developed an association simulation study based upon a mixed approach using affected individuals from the lung cancer dataset and unrelated controls. We generate genotype data considering a range of λ's, Disease allele frequencies, LD, and phenocopy rates. We then compare the power and type I error to detect an association using the entire dataset, and subsets identified by the PS or the RPS as being more genetic. The subsets consist of those individuals below a given threshold for the propensity score, and those above a given threshold for the Mendelian RPS score. Various thresholds are also considered.

**11**

**A Powerful Multilocus Association Test for Quantitative Traits**

M.P. Epstein(1), L.C. Kwee(2)
(1)Dept. of Human Genetics, Emory University, USA, (2) Dept. of Biostatistics, Emory University, USA

There is considerable debate regarding the most efficient approach for association mapping of a quantitative trait using tagSNP genotype data. A common strategy uses statistics based on individual tagSNPs, but such an approach may have low power due to incomplete LD between tagSNPs and the causal variant. An alternative strategy bases inference on statistics that simultaneously consider the joint effects of all tagSNPs in a region (using either genotypes or haplotypes), but the resulting test can have many degrees of freedom that also compromises power. Here, we propose a novel semiparametric regression model for association mapping that incorporates all tagSNP information simultaneously in analysis but results in test statistics with small degrees of freedom. We fit this model using least-square kernel machines, which we show is analogous to model fitting using a linear-mixed model. Using simulated tagSNP data from the International HapMap Project, we demonstrate our approach has superior performance relative to existing approaches for association mapping of quantitative traits. Our approach is also flexible, as it allows easy modeling of covariates and, if interest exists, high-dimensional interactions among genetic and environmental predictors.

**12**

**A method for assessment of gene-wide significance of association with disease**

N.W. Galwey, Hao Li, M.C. Irrizarry, R. Upmanyu, S. Wetten, R.A. Gibson
GlaxoSmithKline

When association with a phenotype is sought in a whole-genome scan (WGS), the underlying aim is usually to assess the evidence for association of each gene studied, rather than each individual marker locus. Any method for doing so must take account of the varying number of loci studied per gene, as lower p-values are expected to be obtained by chance from genes in which a large number of loci are tested. However, due to linkage disequilibrium (LD) among loci, the effective number of significance tests conducted within each gene is less than the nominal number, and this also must be taken into account in the adjustment applied.
A new method of adjustment is presented, based on the eigenvalues of the correlation matrix among the marker loci, and is compared with other methods. Each method of adjustment is applied to two approaches to the joint interpretation of multiple p-values, viz:

- the minimum p-value within each gene.
- a meta-p-value for each gene, obtained from the individual p-values by Fisher's method or the inverse-Normal method.

All these methods are computationally economical, and hence applicable to a WGS. The gene-wise p-values obtained are validated by comparison with those obtained from a permutation-based, computationally-intensive method. In a sample of 34 genes from a WGS for Alzheimer's Disease, correlation coefficients above 0.9 were obtained between each computationally economical method and the corresponding permutation-based method, indicating that the proposed methods will be effective for ranking genes for further investigation.

**13**

**Shifting Paradigm of Association Studies: Value of Rare Single Nucleotide Polymorphisms**

Ivan P. Gorlov(1), Olga Y. Gorlova(1), Shamil R. Sunyaev(2), Margaret R. Spitz(1), Christopher I. Amos(1)
(1) Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas, United States of America, (2) Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America

We explored the usefulness of rare single nucleotide polymorphisms (SNPs) having minor allele frequencies (MAFs) less than 5% for detecting causal genetic variants for common human diseases. By combining the International HapMap data with bioinformatics tools we found SNPs that are more likely to disturb protein function/structure have lower MAF compared to benign SNPs. For a given MAF we computed the proportion of protein damaging SNPs among all nonsynonymous SNPs. We then estimated the joint probability that a SNP is functional and is detected as significant in a case-control study. We found that for a given sample size there was a MAF for which the joint probability is maximal – the most powerful MAF (mpMAF). We further found that mpMAF was negatively correlated with the sample size, suggesting that targeting rare SNPs is advantageous when large sample size is used. We discuss practical implications of the results of the analysis for design of case-control association studies.

**14**

**Catching Local Replications: a Local Score-based approach to replicated association studies**

M. Guedj(1,2), J. Wojcik(2) and G. Nuel(1)
(1)Statistics and genome (UMR Evry Univ, CNRS 8071, INRA 1152), Evry, France, (2)Serono, Geneva, Switzerland

In gene-mapping, replication of initial findings has been put forwards as the approach of choice for filtering false-positives from true signals for underlying loci. In practice, such replications are however too poorly observed. Besides the statistical and technical-related factors (lack of power, multiple-testing, stratification, quality control...) inconsistent conclusions obtained from independent populations might result from real biological differences. In particular, the high degree of variation in the strength of LD among populations of different origins is a major challenge to the discovery of genes.
Seeking for Local Replications (defined as the presence of a signal of association in a same genomic region among populations) instead of strict replications (same locus, same risk allele) may lead to more reliable results. A simple extension of the Local Score approach (Guedj et al 2006) adapted to replicated association studies seems to be promising and constitutes, to our knowledge, a first framework dedicated to the detection of such local replications. On simulations and genome-wide case-control association data, it appears more powerful than single-marker-based replications and robust against genetic heterogeneity among populations.

**15**

**Simultaneous analysis of genome-wide SNP data**

C.J. Hoggart(1), M. De Iorio(1), J.C. Whittaker(2), D.J. Balding(1)
(1) Department of Epidemiology and Public Health, Imperial College London, Norfolk Place, London W2 1PG, (2) Non-communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

The ideal analysis of a genome-wide association study for a complex disease would involve analyzing all the SNP genotypes simultaneously to find a set of SNPs most associated with disease risk. The computational challenge of handling up to one million SNPs simultaneously is daunting, but it could greatly improve performance over single-SNP analyses, since a weak effect may be stronger and a false signal weakened when other causal effects are accounted for. Our algorithm estimates regression coefficients for each SNP by maximizing the likelihood subject to a penalty that strongly favors zero values, corresponding to no association. For each causal variant our algorithm typically reports one SNP that best captures the association and not other SNPs in strong LD with it. We consider two forms for the penalty corresponding to the Laplace and normal-exponential-gamma prior distributions. The Laplace prior improves SNP selection in comparison with single-SNP tests, and the normal-exponential-gamma prior improves selection further. We demonstrate the performance of the algorithm using simulated and real datasets of up to 500K SNPs. These analyses require only a few hours on a desktop workstation, exploiting an approximate calibration of the type-I error that avoids the need for permutation analyses.

**16**

**GWAs – Sifting through the evidence with machine learning tools**

I.R. König(1), D.F. Schwarz(1), J. Erdmann(2), S. Szymczak(1), N. Samani(3), H. Schunkert(2), A. Ziegler(1)
(1)Institut für Medizinische Biometrie und Statistik and (2)Medizinische Klinik II, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany, (3)Department of Cardiovascular Sciences, Glenfield Hospital, University of Leicester, UK

Standard statistical methodology for genome wide association studies (GWAs) is based on separate analyses of single SNPs with classical methods like trend tests on genotypes. Significant SNPs are then subjected to, e.g., haplotype analysis of adjacent SNPs or regression modeling of a few SNPs simultaneously. Because of the instability of models, a disadvantage with this repertoire is that many or even all SNPs cannot be analyzed jointly. Furthermore, if the significance levels are adjusted to a genome-wide level, many studies will lack sufficient statistical power to reliably detect interesting effects. As an alternative, we investigate the use of machine learning algorithms to simultaneously analyze all SNPs from a chip, thus taking possible interdependencies into account. Specifically, based on random forests, interesting SNPs are selected and then sent to regression analysis to develop a model including the chosen SNPs and possible interactions. We illustrate this general approach with the analysis of a simulated and a real data set. For the latter, data from a GWA on myocardial infarction (MI) is utilized in which 875 cases with MI and 1644 population-based controls were genotyped on the 500K Affymetrix chip. We provide evidence that our approach is computationally feasible yielding valid results in the simulated and plausible findings in the real data.

**17**

**Statistical Methods to Identify Loss of Heterozygosity in Matched Samples of Normal and Lung Tumor Tissues**

AT Kraja (1), Q Zhang (1), K Chen (2), L Ding (2), ER Mardis (2), RK Wilson (2), IB Borecki (1), MA Province (1)
(1)Division of Statistical Genomics, Washington University School of Medicine, (2)Genome Sequencing Center, Washington University School of Medicine, Saint Louis, MO, USA

Chromosomal loss of heterozygosity (LOH) can identify regions harboring tumor suppressor and oncogenes. High-density genotyping platforms expand our ability to identify LOH regions, but defining the LOH borders remains a challenge. We developed and assessed statistics to identify LOH regions in 357 matched normal and lung adenocarcinoma tissues from the same patients (Tumor Sequencing Project), with the Affymetrix 250K SNP chip. We compared the informative LOH statistic (the ratio of LOH SNPs to

informative ones); to the LOH corrected for genotyping error (estimated from apparent "gain" of heterozygosity); to sliding windows of informative LOH averaged over a region (R(x)). P-values were calculated from bootstrap resampling and LOH block boundaries were estimated using a Hidden Markov Model. LOH regions were found on chromosomes 5q, 6q, 9p, 17p, 18q, and 19p. To compare methods, we simulated SNPs for 400 matched pairs, comparable to the TSP data, which showed that using large R(x) sliding windows lowers its sensitivity, especially for short significant LOH regions. In combination, these methods appear to improve LOH boundary estimation, which can help identify carcinogenesis genes.

**18**

**Two-stage design for genomewide association studies revisited: power, sample size, and cost trade-offs**

J.P. Lewinger, D.C. Thomas
Division of Biostatistics, Department of Preventive Medicine, University of Southern California, USA

Despite ever decreasing genotyping costs, genomewide association studies (GWAS) remain very costly. To reduce the overall genotyping cost, two-stage designs have been proposed as an alternative to one-stage association studies. Wang et al [2006] extended the two-stage introduced by Satagopan et al [2003] to GWAS and showed that a minimum-cost two stage study can achieved the same power than a single-stage study at a fraction of the cost. However, to achieve equal power, a two-stage design requires a larger total sample size than a one stage study.
We revisit the two-stage design for GWAS to solve the three related problems of finding the minimum cost design subject to power and sample size constraints, the minimum sample size design subject to power and cost constraints, and the maximum power design subject to cost and sample size constraints. The latter has not been addressed before and it is the most relevant in practice since researchers usually face fixed budgets and limited number of samples. Our results have important implications for the planning of GWAS. We show that the maximum-power two-stage design is more powerful than a one stage design (usually dramatically so) for any combination of cost and sample size constraints such that the cost constraint is binding. Equivalently, with only a slight increase in total sample size, a two-stage design can achieve a large cost reduction over an equally powered one-stage study.

Satagopan et al (2003) Genet Epidemiol 25:149–57.
Wang et al (2006) Genet Epidemiol 30:356–368

**19**

**Detecting Associations in the Presence of Extreme Allelic Heterogeneity: Application to the Rare Variant Common Disease Hypothesis**

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

Association studies are frequently utilized to map variants which are susceptibility loci for common diseases. Critical assumptions of this approach are that the disease is due to a common functional variant which is in strong linkage disequilibrium with genotyped SNP(s) and there is minimal allelic heterogeneity. If the rare variant common disease hypothesis holds, current association based methods will be underpowered due to allelic heterogeneity, low allele frequencies and poor correlation (r2) with tagSNPs. For common diseases where the underlying etiology is believed to involve extreme allelic heterogeneity, large scale candidate gene sequencing is currently underway to discover multiple causal rare variants. However, which methods are optimal for analyzing this type of data to detect associations is unknown. In this study, we analytically demonstrate that collapsing genotypes and rare variants across multiple loci is more powerful than multi-marker test (Hotelling's T2) and single marker test (Fisher exact test). Collapsing methods are also robust against misclassifications unless the non-causal variants are common. In that case, collapsing only rare variants gained significant robustness with little loss of power. Empirical findings from simulation studies were consistent with analytical results and, additionally, it was shown empirically that for collapsing methods type I error is well controlled.

**20**

**Reduce the Disturbance of Bias in Genome-wide Association Studies: Extreme-value Genome-wide Association Studies (EGWAS)**

Dalin Li and David V. Conti
Department of Preventive Medicine, University of Southern California, Los Angeles, CA

Although many genetic epidemiological studies have been carried out in complex diseases, with many positive results, few have been replicated. This lack of replication is often attributed to the lack of power of the studies, although studies with larger sample sizes also fail to replicate. Here, we argue that potential bias is another important reason for this lack of replication, and bias would make the research result unpredictable when the true effect is weak. We propose the extreme-value analysis for genome-wide association studies (EGWAS) in which only individuals with extreme phenotypes are recruited in a genome-wide association study to reduce the disturbance of bias. Our simulation shows that the EGWAS design yields strong associations between the underlying causal SNPs and the outcome, while the accompanying bias effects do not increase correspondingly – thus, reducing the disturbance of bias significantly. We compare our approach to traditional genome-wide association studies (GWAS) where true effects can be easily obscured from the biased results of the non-causal SNPs, even if the bias is extremely rare. Finally, we discuss the cost efficiency of this design when there is absolutely no bias. We show that when the number of subjects needed to be screened to generate the sample(s) is fixed, by carefully choosing the cutoff to the underlying continuous trait, EGWAS can be more efficient in terms of power than the traditional GWAS design, while the number of subjects needed to be genotyped is greatly reduced.

**21**

**Modelling extended haplotypes in genetic association studies**

P.M. McKeigue(1), D. O'Donnell(1), M. Blizinski(1), M. Dunlop(2), A. Tenesa(2), H. Campbell(3), M.D. Shriver(4)
(1)Conway Institute, University College Dublin, Ireland, (2)MRC Human Genetics Unit, Western General Hospital, Edinburgh, UK, (3)Department of Public Health Sciences, University of Edinburgh, UK, (4)Department of Anthropology, Penn State University, USA

We describe methods for testing for association at all untyped HapMap loci in a genetic association study, using phase-known haplotypes for the corresponding continental group in the HapMap database. Each gamete is modelled as a mosaic of K hidden states representing modal haplotypes (as in the fastPHASE program), and score tests for association are constructed by averaging over the posterior distribution of missing genotypes. The ability to predict missing genotypes was examined using HapMap data with some genotypes masked, comparing predictive distributions with the true genotypes. In a model with K=8, HAPMIXMAP outperformed fastPHASE in the prediction of missing genotypes in Africans and in Europeans. To evaluate the ability of a tag SNP array to test for association at untyped loci, we analysed 200 cases and controls of European ancestry from a case-control study of colon cancer, typed with the Illumina Hap550 tag SNP array. These methods make it feasible to analyse any association study that uses tag SNPs as if all loci in the HapMap had been typed, and to combine in a meta-analysis studies in which different tag SNP arrays or candidate gene polymorphisms have been typed.

**22**

**Impact of Linkage Disequilibrium and Effect Size on the Ability of Machine Learning Methods to Detect Epistasis in Case-Control Studies**

K.K. Nicodemus(1), Y.Y. Shugart(1)
(1) Epidemiology, Johns Hopkins SPH, USA

A novel approach to detect epistasis in case-control studies is using methods designed for high-dimensional data, e.g., machine learning methods. Using a simulation study, we sought to evaluate the ability of 3 algorithms (random forests (RF), Monte Carlo logic regression (MCLR), generalized boosted regression (GBM)), using classification trees as the base learner, to detect epistasis, using each method's importance measure. In each condition, we simulated data (1000 replicates) for 5 genes; in 2 genes SNPs interacted to increase risk for case status. Additional genes were unassociated. We varied the LD between SNPs within genes, effect size and type of interaction. Detection of epistasis was defined as the # of replicates where both causal loci ranked in the top 5% of important SNPs/total # of replicates. LD was the strongest determinant in detecting epistasis. When both causal loci were in low LD with other SNPs and using an odds ratio (OR) of 2.5, RF detected both loci in 100% of replicates; MCLR and GBM detected both loci in ~80% of replicates. When both causal SNPs were in strong LD with other SNPs and using OR=2.5, RF was unable to detect (<10%) correct loci, whereas MCLR and GBM were able to detect correct SNPs in ~80% of replicates. We show that reducing the correlation between trees and increasing the number of trees improves the ability of RF to detect the correct SNPs in high LD conditions. Machine learning methods are a viable way of detecting epistasis in case-control studies; however, LD should be considered during algorithm selection.

**23**

**Confounding between recombination and selection, and a novel genome-wide method for detecting selection**

P.F. O'Reilly (1), E. Birney (2), and D.J. Balding (1)
(1)Dept. Epi & Public Health, Imperial College London, UK
(2)European Bioinformatics Institute, UK

In recent years there have been major developments of population genetics methods to estimate both rates of recombination and levels of natural selection. However, genomic variants subject to positive selection are likely to have arisen recently, and consequently had less opportunity to be affected by recombination. Thus, the two processes have an intimately-related impact on genetic variation, and inference of either may be vulnerable to confounding by the other. We illustrate here that even modest levels of positive selection can substantially reduce population-based recombination rate estimates in humans. We also show that genome-wide scans to detect loci under recent selection in humans have tended to highlight loci in regions of low recombination, suggesting that confounding with recombination rate may have reduced the power of these studies. Motivated by these findings we introduce a new genome-wide approach for detecting selection, based on the ratio of pedigree-based to population-based estimates of recombination rate. Simulations suggest that this "Ped/Pop" approach has good power to discriminate between neutral and adaptive evolution. This power is maintained many generations after fixation of the selective sweep. Unusually for a multi-marker method our approach also shows good power in regions of high recombination. We apply the method to human HapMap and Perlegen data sets, finding confirmation of reported candidates as well as identifying new loci that may have undergone recent intense selection.

**24**

**Gathering the gold dust: Identification small-effect complex trait genes**

M.A. Province, I.B. Borecki
Division of Statistical Genomics, Washington University School of Medicine

Genomewide association scan (GWAS) data mining has found moderate-effect "gold nugget" complex trait genes. But for many traits, much of the explanatory variance may be truly polygenic, more like gold dust, whose small marginal effects are undetectable by traditional methods. Yet, their

collective effects may be quite important in advancing personalized medicine. We consider a novel approach to sift out the genetic gold dust influencing quantitative (or qualitative) traits. Out of a GWAS, we randomly grab handfuls of SNPs, modeling their effects in a multiple linear (or logistic) regression. The model's significance is used to obtain an iteratively updated pseudo-Bayesian posterior probability associated with each SNP, which is repeated over many random draws until the distribution becomes stable. A stepwise procedure culls the list of SNPs to define the final set. Results from a benchmark simulation of 5 trait genes among 1,000, in 1,000 subjects, are contrasted with marginal tests using nominal significance, Bonferroni-corrected significance, false discovery rates, as well as with serial selection methods. Random handfuls produced the best combination of sensitivity (0.94) specificity (0.99) and true positive rate (0.89) of all methods tested and better replicability in an independent subject set. From more extensive simulations, we determine which combinations of signal to noise ratios, SNP typing densities, and sample sizes are tractable with which methods to gather the gold dust. The methods are also evaluated in real data from the NHLBI Family Heart Study.

**25**

**Simple Methods for High-Density Copy Number Variation Data**

G.A. Satten(1), J.G. Mulle(2), A.S. Allen(3), M.P. Epstein(2) S.T. Warren(2)
(1)Centers for Disease Control and Prevention, (2)Dept. of Human Genetics, Emory University, (3)Dept. of Biostatistics and Bioinformatics, Duke University

The extent of copy number (CN) variation in the human genome is an exciting new discovery with important implications for complex disease studies. The amount of data generated by current methods has outpaced the ability of available algorithms to predict CN variants efficiently or accurately. In array-based systems, sample DNA (labeled with Cy5) is compared to a reference DNA (labeled with Cy3) by measuring the log intensity ratio (LIR) at each of many probes. When the CN of sample and reference differ, the LIR deviates from 0. We describe simple methods for finding probes corresponding to changes in CN. We first compute the left-smoothed and the right-smoothed LIR at each probe j, denoted L(j) and R(j), where L(j) and R(j) smooth LIR data only from probes to the left or right of locus j, respectively. L(j) and R(j) can be calculated recursively, which also gives improved behavior near the telomeres. We then calculate candidate jump points by applying a novel peak-finding algorithm to D(j)=R(j)-L(j-1). Finally, we fit the LIR data using ordinary least squares and step functions that change at the candidate jump points. Backward elimination is used to remove jumps smaller than a user-defined threshold. We evaluate our approach using data from NimbleGen whole-genome oligonucleotide arrays, containing 2.1 million probes tiled at 1.1 kb intervals throughout the genome, measured on phenotypically normal individuals from an Ashkenazi Jewish population.

**26**

**Using Haplotype Clustering Techniques to Perform Genome-Wide Disease Association Studies**

S.-Y. Su, D.J. Balding, L. Coin
Department of Epidemiology & Public Health, Imperial College London, UK

Association studies using population genetic data are expected to have greater power to detect small and moderate variants in complex disease versus linkage studies. Haplotype-based association studies are a promising approach for identifying such genetic variants due to the captured information about the interdependence among alleles and ancestral structure in the haplotype distribution. However, there are some potential problems with haplotype-based analysis including: unobserved haplotypes, abundant rare haplotypes and the definition of block boundary along the sequence.
Aiming to solve these problems, we propose a likelihood-based approach with clustered haplotype techniques to assess haplotype association with disease in population-based studies. First, we employ a Hidden Markov Model (HMM), an algorithm similar to HINT, to cluster haplotypes into few groups of assumed founder states from genotype data. Under this model, recombination patterns along the sequence are captured by modeling the hidden states with estimated jump events between markers; and haplotypes are clustered according to the relation between hidden states and observed genotype data modeled by emission probabilities in the HMM. Then, we conduct haplotype grouping association test based on a likelihood of the counts of case and control with the number of haplotype in a cluster. The results of analyzing a mice dataset show that our method can have higher rates of correct identification of causative loci in the additive, dominant and recessive disease models compared to the Armitage trend test.

**27**

**A Fast Implementation of Scan Statistic for Identifying Chromosomal Patterns of Genome Wide Association Studies**

Y.V. Sun(1), D.M. Jacobsen(1), S.T. Turner(2), K.R. Bailey(3), E. Boerwinkle(4), S.L.R. Kardia(1)
(1)Dept. of Epid, Univ. of Michigan, (2)Div. of Nephrology and Hypertension, Mayo Clinic, (3)Div. of Biostat, Mayo Clinic, (4)Hum Genet Ctr, Univ. of Texas Health Sciences Ctr, USA

We have developed a single nucleotide polymorphism (SNP) association scan statistic that accounts for the complex distribution of genomic variation when identifying chromosomal regions with significant SNP effects. To address the computational needs for analyzing genome wide association study (GWAS) data of hundreds of thousands SNPs, we implemented a fast JAVA application combining single-locus SNP tests and the scan statistic for identifying chromosomal patterns. It can quickly identify chromosomal regions associated with phenotypes based on GWAS genotype data. To illustrate this application, we analyzed SNP associations in

a pharmacogenomic study using the Affymetrix Human Mapping 100K Set. We selected 55,335 tagSNPs (pair-wise linkage disequilibrium $R2 < 0.5$) to reduce the correlation between SNPs. A typical workstation can complete the whole genome scan of 10,000 permutation tests within hours. We found the most significant regions located on chromosome 7 that contained eight clusters of SNPs, six of which contain at least one known gene that may be involved in the biological mechanism of drug response. The average size of these regions was 200 kb with a range of 49 to 418 kb. This scan statistic application can be used to make regional inference about disease association, and is expected to provide a better statistical foundation for identification of these regions and comparisons across GWAS.

**28**

**Genetic Association Mapping via Evolutionary-Based Clustering of Haplotypes**

I. Tachmazidou(1), C.J. Verzilli(2), M. De Iorio(1)
(1)Department of Epidemiology and Public Health, Imperial College London, UK, (2)Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, UK

Multilocus analysis of single nucleotide polymorphism (SNP) haplotypes is a promising approach to dissect the genetic basis of complex diseases. We propose a coalescent-based model for association mapping which potentially increases the power to detect disease-susceptibility variants in genetic association studies. The approach uses Bayesian partition modelling (BPM) to cluster haplotypes with similar disease risks by exploiting evolutionary information. We focus on candidate gene regions with densely spaced markers and model chromosomal segments in high linkage disequilibrium (LD) therein assuming a perfect phylogeny. To make this assumption more realistic, we split the chromosomal region of interest into sub-regions or windows of high LD. The haplotype space is then partitioned into disjoint clusters within which the phenotype-haplotype association is assumed to be the same. For example, in case-control studies we expect chromosomal segments bearing the causal variant on a common ancestral background to be more frequent among cases than controls, giving rise to two separate haplotype clusters. The novelty of our approach consists in the fact that the distance used for clustering haplotypes has an evolutionary interpretation, as haplotypes are clustered according to the time to their most recent common ancestor. Our approach is fully Bayesian and we develop a Markov Chain Monte Carlo (MCMC) algorithm to sample efficiently over the space of possible partitions. We compare the proposed approach to both single-marker analyses and recently proposed multi-marker methods and show that the BPM performs similarly in localizing the causal allele while yielding lower false positive rates. Also, the method is computationally quicker than other multi-marker approaches. We present an application to real genotype data from the CYP2D6 gene region, which has a confirmed role in drug metabolism, where we succeed in mapping the location of the susceptibility variant within a small error.

**29**

**Combined analysis of SNP and gene expression data in TLR genes with respect to infection disease in Ghana**

H.-W. Uh(1), F.C. Hartgers(2), M. Yazdanbakhsh(2), J.J. Houwing-Duistermaat(1)
(1) Dept. of Medical Statistics and Bioinformatics, (2) Dept. of Parasitology, Leiden University Medical Center, Leiden, The Netherlands

The study population consists of 565 school children from two urban and two rural schools in Ghana. Phenotypes of interest were helminth, malaria, and atopy. While 29 polymorphisms in the Toll-like receptor (TLR) genes were typed, gene expression data were available for 21 genes.
Differences in disease patterns between rural and urban areas exist. Allergies are relatively rare in rural areas compared with urban centers where allergic reactivites are highly prevalent. Pathogen exposure may be different in rural versus urban communities and it is known that microbial exposure, such as helminths, can be protective against allergic diseases. TLRs might play an important role in the link between infections and allergies, since products from pathogens have been shown to bind to TLRs and thereby activate cells of the immune system.
Our first interest is the identification of TLR genes that influence the risk of disease. Secondly, we explore highly correlated gene expression data. To overcome overfitting, we conduct Principal Component Analysis (PCA) to determine the combinations of gene expressions that best capture most of the underlying variance. These components are then used as variables in the logistic regression model.
TLR3 genes were appeared to be associated with malaria. The first principal component that explains most of the variance was associated with both helminth and malaria infection.

**30**

**A shrinkage regression approach to tackle the HLA region**

C. Vignal(1,2), A. Bansal(2), D. Balding(1)
(1)Imperial College London, UK, (2)GlaxoSmithKline, UK

Many autoimmune diseases have been associated with the HLA region, but the presence of linkage disequilibrium (LD) has meant that finding causal elements has been difficult. Multivariate association analyses can perform better than univariate methods, however, there can be problems when the number of variables exceeds the number of observations or in the presence of correlated predictors.
We adopt a Bayesian-inspired shrinkage regression approach for multilocus analysis of correlated data. Each regression coefficient is assigned a Laplace (double-exponential) prior distribution with mode zero. Parameter inference is based on the posterior mode and terms with non-zero posterior modes indicate marker-disease associations.
We applied this approach to a case-control association study on rheumatoid arthritis (RA) using SNPs spanning the HLA region, together with genotypes from the multiallelic HLA-DRB1 locus. The latter is a known RA risk factor that

was included in all our models without shrinkage. After controlling for type-I error, we found fewer positive SNP associations than in single-point tests, suggesting that LD might be better-handled. These results were supported by a simulation study. We selected a set of SNPs in various degrees of LD with HLA-DRB1. For each marker, case-control labels were randomised within the HLA-DRB1 allelic classes to simulate causal SNPs, while maintaining LD with HLA-DRB1. Our results showed that the shrinkage approach provides a substantial benefit, both in terms of maintaining statistical power to detect multiple causal variants and in the reduction of false positive associations.

**31**

**A novel multi-locus method for modeling disease association**

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

Several multi-locus methods have been proposed to improve the localization of a disease associated SNP by smoothing the output of an association study. We have developed a novel method which applies the Malecot model to a multinomial haplotype likelihood using individual subject case-control data. The Malecot constraint models the decline in association between a SNP and disease as a function of map distance and LD structure. Applying this constraint to the observed haplotype data allows us to estimate the underlying strength of LD. In contrast to previous approaches, ours uses individual subject data rather than summary haplotype frequencies. We have derived proper estimates of standard errors for the model parameters using a robust estimator, giving a confidence interval on the location of the disease. The method works with haplotype data and, assuming Hardy-Weinberg equilibrium, unphased genotype data. Model checking has been incorporated using an empirical goodness of fit method. Results from both simulated data and real genetic data with a simulated disease indicate that the model works well, with the correct position of the disease causing SNP being estimated accurately and with narrow confidence intervals. Our method shows potential for improving the fine mapping of disease and quantitative trait loci with scope for including environmental covariates.

**32**

**Detection of Homozygous Segments: Search for Genomic Deletions**

C.C. Wu, S. Shete, C.I. Amos
Department of Epidemiology, UT M. D. Anderson Cancer Center, Houston, Texas, U.S.A.

There is increasing evidence that many human genetic disorders are associated with deletions of DNA sequences in chromosomes. Identified microdeletions range in size from < 1kb to 4Mb. Large deletions are usually involved with loss of genetic material encompassing contiguous genes. Among the most common contiguous-gene deletion syndromes in humans are DiGeorge syndrome and Velo-Cardio-Facial syndrome on chromosome 22q11. Smaller chromosomal deletions have been implicated as causally related to complex diseases such as autism, schizophrenia, and childhood mental retardation. Because the current molecular techniques are not efficient for direct identification of genetic deletions, we have developed marker-based statistical methods for detecting regions of microdeletion using case-control designs. We propose to test for an excess of homozygosity of contiguous genetic markers in patients presenting with a genetic disorder. Our approaches evaluate the frequency of contiguous homozygous segments in a subject and determine the heterozygosity level at each region among subjects. Identification of overlapping microdeletions in multiple patients provides strong evidence for the role of genes within the deleted region predisposing to risk for the disease. The sensitivity and power of our proposed methods depend on marker density, allele frequency, and size of microdeletion. We will perform analyses for simulation studies and lung cancer data.

**33**

**What type of genotyping error most inflates type I error for differential misclassification in case-control studies?**

K. Ahn(1), D. Gordon(2), S.J. Finch(3)
(1) Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033,
(2) Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854,
(3) Applied Math and Statistics, Stony Brook University, Stony Brook, NY 11790.

Genotyping error adversely affects the statistical power or type I error and parameters for case-control association studies. Most research has assumed non-differential genotype misclassification; that is, the same error mechanism and probabilities apply to affected and unaffected individuals. Research dealing with differential genotype misclassification in genetic studies has started to appear, but there is little research on what type of genotyping error will most inflate the type I error. Therefore, we have examined the non-centrality parameter of chi-squared test and LTT when there is no difference between case and control genotype frequencies but there is differential misclassification with SNP data. The parameters examined are minor allele frequency (MAF), disease allele frequency and sample size. We found that: 1) when $MAF << 0.5$, genotyping errors lead to a greater inflation of the type I error; 2) the problem increases as the sample size increases; and 3) the most problematic errors to inflate type I error are recording homozygote as another homozygote and the more common homozygote as the heterozygote.

**34**

**Repair phenotype, polymorphisms in repair genes and genotoxicity in radiation exposed workers**

M. Kirsch-Volders (1), R. Mateuca (1) and P. Aka (1,2)
(1) Laboratory of Cell Genetics, Vrije Universiteit Brussels, Belgium
(2) Department of Genetics, Medical Research Council Laboratories, The Gambia

We studied two exposed groups of nuclear plant workers, one chronically exposed (n=28) and the other made up of seasonal cleaners acutely exposed (n=32). The third and fourth groups were controls to the first (n=19) and second (n=31) groups respectively made of office staff. Using the Comet assay, we assessed DNA damage and the DNA strand break repair phenotype. The frequencies of micronuclei were assessed by means of the in vitro micronucleus assay. Genotyping for DNA repair genes OGG1, XRCC1 and XRCC3 was performed on blood samples of the acutely exposed workers using restricted fragment length polymorphis (RFLP).
Our data show that exposed workers repaired damage to their DNA more proficiently than their controls. Also, the exposed smokers had higher levels of DNA damage and micronuclei frequencies than non smokers. In the acutely exposed workers, a significant contribution of the OGG1 genotypes to the in vitro DNA strand break repair capacity was found. A multivariate analysis revealed that genetic polymorphisms in XRCC1 resulted in higher residual DNA values and the Met/Met variant of XRCC3 gave an increased frequency of micronuclei. We conclude that a combined analysis of the three genotypes, OGG1, XRCC1 and XRCC3 polymorphisms is advised in order to assess individual susceptibility to ionising radiation. As an alternative or complement, the in vitro DNA strand break repair phenotype which integrates several repair pathways is recommended.

**35**

**Estimating genetic disease risk from family data: choosing the optimal method to correct for ascertainment**

F. Alarcon(1), C. Bourgain(1), V. Planté-Bordeneuve(2), D. Stoppa-Lyonnet(3), C. Bonaïti-Pellié(1)
(1) INSERM U535, Villejuif
(2) Dept Neurology, CHU Bicêtre, Le Kremlin Bicêtre
(3) Genetic Oncology, Institut Curie, Paris, France

Providing valid risk estimates of a genetic disease using family data requires an adjustment for ascertainment bias, since families are selected through affected individuals. When families have been ascertained through one affected individual (proband), the Prospective likelihood (Kraft and Thomas, A.J.H.G. 66:1119-1131, 2000) conditions on the existence of at least one ascertained individual among affected family members but requires knowledge of the probability of ascertainment. Another widely used method simply excludes the proband's phenotype and analyses the rest of the family without any correction (which we refer to as the "Proband's phenotype Excluded Likelihood" or PEL). The PEL implicitly assumes that the ascertainment probability is low.
We compared these two methods in terms of bias by simulating family samples under various disease risk models and various selection patterns with at least one affected member. We show that the optimal method depends on both the genetic model and the ascertainment scheme.
As an illustration, we estimated the disease risks for transthyretin amyloid neuropathy from a French and a Portuguese sample and for breast and ovarian cancer from a sample ascertained on early-onset breast cancer cases. In each case, the optimal estimation method used was different.

**36**

**Multiple SOD1 SNPs are associated with the development and progression of Diabetic Nephropathy**

H. Al-Kateb(1), AP. Boright(2), X. Xie(3), L. Mirea(3,4), R. Sutradhar(3), L. Sun(1,4), SB. Bull(3,4), AD. Paterson(1,4) and the DCCT/EDIC Research Group
(1) Genet & Genome Biol, SickKids (2) Med, UHN (3) Prosserman, SLRI (4) Dept Pub Hlth Sci, UofT, Toronto, Canada

Despite familial clustering of nephropathy (DN) in type 1 diabetes (T1D), few SNPs have been consistently detected in association studies. We performed an individual-based association study of time to persistent microalbuminuria (PM) and severe nephropathy (SN) in 1,362 white T1D probands from the Diabetes Control & Complications Trial/Epidemiology of Diabetes Interventions and Complications study with 1,414 SNPs in 212 candidate genes. SNPs were analyzed for association with the time from DCCT baseline to event using multivariate Cox models. To adjust for multiple hypothesis testing, q values from the false discovery framework were calculated for each outcome. We observed association of a SNP 3–TM of SOD1 (rs17880135) with both SN (HR=2.62, 95%CI 1.64–4.18, p=5.6 x 10−5, q=0.06) and PM (HR=1.82, 95%CI=1.29−2.57, p=6.4 x 10-4, q=0.46). Sequencing and fine-mapping identified 4 additional SOD1 variants associated with PM, two of which were also significantly associated with SN. Attempts to replicate the findings in cross-sectional case-control studies produced non-significant results, which may be explained by the Beavis effect and insufficient power. We observed no significant differences between risk genotypes in serum SOD activity, in serum SOD1 mass, nor in SOD1 mRNA expression in lymphoblasts. Multiple variations in SOD1 are significantly associated with the time to PM and SN, although the effect size is small.

**37**

**Higher order interaction: how do you minimize multiple comparisons and integrate biological pathways? One approach**

C. Aragaki (1), K. Klos (1), K. Volcik (1), E. Boerwinkle (1)
(1)Human Genetics Center, University of Texas Health Science Center at Houston, USA

In order to elucidate the pathways involved in complex disease, new analytic strategies need to be developed to integrate the human genome and environmental factors. In particular, the sheer number of potential combinations and integration of the multiple biologic pathways are two problems which need to be solved by new methods. One analytic strategy is to combine current differing analytic techniques to minimize each problem. In the following approach, we use Moore's multifactor dimensionality reduction (MDR) within a biological pathway on an intermediate endpoint to determine combinations that impact disease and then integrate multiple pathways in a Hierarchical Bayes approach. METHODS: Using candidate nonsynonymous SNPs in the lipid metabolism and renin-angiotensin pathways and cardiovascular risk factors measured in

the Atherosclerosis Risk in Communities cohort study, we determined SNP-environmental factor pathway combinations that impact risk for hypertension and hyperlipidemia. We then regressed these results on cardiovascular events in a hierarchical Bayes Cox regression. RESULTS: We found that genes and environmental factors explained risk variation better than either alone.

**38**

**A multiple-marker strategy based on the Local Score is a useful tool for positional cloning**

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

Genotyping large numbers of SNPs across the genome or in linkage regions raise analytical issues, among which the multiple testing problem occupies a central role. We have recently shown, in the context of a genome-wide association study, that a multiple-marker strategy as compared to a single marker approach reduces the false discovery rate. Our aim was to compare the performance of these two strategies in a linkage region selected from a genome screen for positional cloning. A total of 1,505 individuals belonging to 368 families from the French EGEA asthma study were genotyped for 1047 SNPs, spanning a 18 Mb region linked to a measure of lung function. The single marker approach used the FBAT method applied to each SNP and p-values were corrected by Bonferroni correction (BC). For the multiple-marker strategy, we used the Local Score method to select sets of markers with a positive score; each set was then analyzed by FBAT and BC was applied to p-values within sets and across sets. The single marker approach led to detect 4 association signals with $p < 0.0005$ spanning 3 Mb, none of them being significant after BC. When using the multiple-marker strategy, the Local score method selected 11 sets of markers with positive scores and 4 of these sets remained significant after BC (p-corrected ranging from 0.001 to 0.008), the two first sets spanning 900 kb. The multiple-marker strategy helps in controlling the multiple testing problem and in selecting small regions to be further explored towards gene identification.

**39**

**Investigation of the IL4-IL13/IL4R pathway in French Multiple Sclerosis patients**

M.C. Babron (1), E. Genin (1), I. Cournu-Rebeix (2), H. Perdry (1), B. Fontaine (2), F. Clerget-Darpoux (1)
(1) INSERM UMR535 and Univ. Paris-Sud, Villejuif, France
(2) INSERM UMR546 and Univ Paris 6, Paris, France

Multiple Sclerosis (MS) is multifactorial disease, in which both genetic and environmental factors intervene. Apart from HLA, little is known however on the genetic factors involved. Two genome scans [Broadley et al, 2001; Babron et al, 2004] have suggested the presence of a risk factor in the chromosomal region 5q31. The location of the IL4 and IL13 genes, coding for cytokines involved in the immune response, in this region makes them good candidates for susceptibility to MS. Previous studies on type I diabetes, another autoimmune disease, have suggested interactions between IL4/IL13 and their receptor IL4R [Bugawan et al, 2003].
We investigated the role of this pathway in a sample of 124 French trios (an affected child and his two parents). Sequencing of all the individuals allowed identification of a total of 249 SNPS. Among them, 14, 23 and 86 for IL13, IL4 and IL4R, respectively, had a minor allele frequency $< 1\%$. None of these, taken individually, showed significant association with MS.
However, in multifactorial diseases such as MS, combinations of two or more variants in the same pathway may be involved. Thus, we investigated combinations of two variants, one in IL4 or IL13 and one in IL4R. 7 combinations out of 3219 were significant at the corrected 1% level, and were more significant than each SNP analysed separately. These combinations will need to be tested in independent samples for replication.
We thank REFGENSEP for data sharing and ARSEP for funding.

**40**

**Gene-Environment Interaction in Mesothelioma**

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

About 50% of deaths in three villages in Cappadocia, Turkey, are attributed to malignant mesothelioma (MM) (BARIS and GRANDJEAN 2006; BARIS et al. 1978). Although MM is typically associated with asbestos exposure in the industrialized world, the epidemic in Cappadocia has been linked to exposure to erionite, a fibrous zeolite mineral suggested to be a potent carcinogen for MM. Preliminary pedigree analysis demonstrates that in these villages, MM is common in some families and not in others, and when high-risk MM family members marry into a low-risk family, MM appears in the descendents. Analysis of a 6-generation extended pedigree of 526 individuals revealed that MM has a transmission pattern consistent with an autosomal dominant genetic model, with reduced penetrance due perhaps to gene-environment interaction. Additionally, MM does not seem to develop in members of high-risk families born and raised outside these villages (CARBONE et al. 2002). In the complex etiology of MM, environmental carcinogens and genetic factors may cause malignancy alone, in concert, or synergistically. To identify genes that confer susceptibility to mesothelioma, families in the Cappadocian villages of Karain, Old Sarihidir, and Tuzkoy that are at very high risk for MM have been identified, and pedigrees of these families have been established. DNA is being collected from both non-malignant (blood) cells and tumor biopsies of living and deceased members of these families. Analyses of these data will focus on both traditional methods of linkage and association

analysis, and novel approaches designed to take into account the genomic information on association, DNA copy number variation and the possibility that affected individuals will share genetic risk factors identical-by-descent from a recent common ancestor. We present preliminary results of the studies primarily focused on the novel methods of analysis.

BARIS, Y. I., and P. GRANDJEAN, 2006 Prospective study of mesothelioma mortality in Turkish villages with exposure to fibrous zeolite. J Natl Cancer Inst 98:414–417.
BARIS, Y. I., A. A. SAHIN, M. OZESMI, I. KERSE, E. OZEN et al., 1978 An outbreak of pleural mesothelioma and chronic fibrosing pleurisy in the village of Karain/Urgup in Anatolia. Thorax 33:181–192.
CARBONE, M., R. A. KRATZKE and J. R. TESTA, 2002 The pathogenesis of mesothelioma. Semin Oncol 29:2–17.

**41**

**Model Selection in Case-Parent Triad Studies**

Tracy L. Bergemann
Division of Biostatistics, School of Public Health, University of Minnesota

Studies that genotype individuals within nuclear families are now widespread. Generally, samples are drawn from an affected offspring, manifesting a disease or phenotype of interest, as well as from the parents [Ahsan H et al, 2002]. Case-parent triad designs avoid the potential for spurious association results due to admixture that can occur in case-control studies. And, if parents of the offspring are alive and consent to genotyping, the use of nuclear families can be a powerful method in certain disease settings.
It is of interest to test for association, not only of single SNPs, but also any possible gene-gene interactions and gene-environment interactions using methods outlined in Umbach DM and Weinberg CR, 2000. The number of potential log-linear models to fit the data is therefore quite large. We suggest a strategy to find optimal models that incorporate both the biological information, e.g., the linkage disequilibrium patterns of the SNP data, as well as traditional methods for model selection such as the Bayesian Information Criterion. Further, we modify existing methods to impute missing data so that we get more accurate estimates of the variance of our model parameters.

**42**

**Association between the ACCN1 Gene and Multiple Sclerosis**

L. Bernardinelli(1,2), S.B. Murgia(3), P.P. Bitti(4), L. Foco(1), R. Pastorino(1), D.R. Cox(5), C. Berzuini(2)
(1)Dip. Scienze Sanitarie Applicate, Univ. Pavia, I, (2)MRC Biostatistics Unit, Cambridge, UK, (3)Div. Neurologia, ASL N°3 Nuoro, I, (4)Centro Tip. Tiss., ASL N°3 Nuoro, I, (5)Dept.of Stat., Oxford, UK

We study the very special population of Nuoro, Sardinia, an isolated, old and genetically homogeneous population with high prevalence of Multiple Sclerosis (MS). We first typed microsatellites in the 17q11.2 region in MS nuclear families and unrelated MS cases and controls. An association signal was found at D17S798. Next, a bioinformatic screening of the region surrounding this marker, highlighted an interesting candidate MS susceptibility gene: the Amiloride-sensitive Cation Channel Neuronal 1 (ACCN1) gene. We resequenced the ACCN1 gene, and investigated the MS association of identified SNPs. We developed a method of analysis where complete, phase-solved, posterior-weighted haplotype assignments are imputed for each study individual from incomplete, multi-locus, genotyping data. The imputed assignments provide an input to a number of proposed procedures for testing associations at a microsatellite level or of a sequence of SNPs. These include a Mantel-Haenszel type test based on expected frequencies of pseudocase/pseudocontrol haplotypes, as well as permutation based tests, including a combination of permutation and weighted logistic regression analysis. We found a statistically significant association between MS and a SNP located in the 3′ UTR of ACCN1. This result is consistent with several recent experimental findings which suggest that ACCN1 may play a role in the in the pathogenesis of MS.

**43**

**A comparison of approaches for identifying gene-gene interactions in a study of genetic and clinical predictors of severe alcohol withdrawal**

J.M. Biernacka(1,2), B.L. Fridley(1), S.R. Stevens(1), C. Colby, V.M. Karpyak(2)
(1) Division of Biostatics, Dept. of Health Sciences Research, (2) Dept. of Psychiatry and Psychology, Mayo Clinic, Rochester MN, USA

It is believed that gene-gene interactions play an important role in the development of most phenotypes. Despite this widespread belief, few studies in psychiatric genetics attempt to study gene-gene interactions. A recent candidate gene study of clinical and genetic predictors of severe withdrawal symptoms in alcoholism identified a potentially important interaction involving genes in the serotonergic and dopaminergic pathways using logistic regression. Here we apply a number of model selection methods to the same data, with the aim of identifying gene-gene interaction effects. Methods considered include logistic regression, bootstrap logistic regression, random forests, Bayesian model averaging, and Bayesian variable selection, as well as multifactor dimensionality reduction (MDR). Results across methods are fairly consistent in terms of which interaction effects are detected during the modeling process. However, the final selected model differs across the statistical methods. Simulations are used to compare properties of the methods for several scenarios with gene-gene interactions contributing to a binary trait, including data generating models that give rise to similar main effect and interaction patterns as those observed in the alcohol withdrawal data. We discuss the differences between the various methods and suggest strategies for searching for gene-gene interactions.

**44**

**Estimating Association Parameters in Family-based Association Studies**

Stefan Boehringer (1), Ruth Pfeiffer (2)
(1) Institut fuer Humangenetik, Universitaetsklinikum Essen, Essen, Germany, (2) National Cancer Institute, Biostatistics Branch, DCEG, USA

Genetic association studies, using either independent cases or controls or based on families, are the primary means to elucidate genetic contributions to diseases including diabetes, Alzheimer's, autoimmune diseases and many others. The aim is to find variations in the genome that influence disease phenotypes and characterize the prevalence of disease alleles and effect sizes. For family based association studies, Mendelian inheritance can be used to make statistical inference about parameters which describe a disease locus based on observations of adjacent genetic loci in a given region of the genome. We model an unobserved true disease locus in linkage disequilibrium with observed markers in a genomic region using observations on family members. A random effect based on a normal distribution, characterizes the residual polygenic effects from other latent unlinked loci. The model also allows correction for ascertainment of the families. Using a maximum likelihood approach, we estimate the joint probability distribution of the observed markers and the latent true disease locus, a penetrance parameter measuring the impact of the disease allele on disease risk and the variance of the random effect. We show using simulations that parameter estimates are consistent and provide a powerful means of analyzing family-based data and apply the method to an ApoE/Alzheimer's data set.

**45**

**Meta-analysis of genome-wide linkage studies for asthma and atopy phenotypes, using an extended GSMA method**

E. Bouzigon(1,2) for the asthma meta-analysis group
(1)INSERM U794, France, (2)Med. & Mol. Genet., King's College London, UK

Nineteen linkage genome-wide scans have been conducted for asthma and asthma-related phenotypes in different populations and have led to the identification of many genomic regions that may harbour susceptibility genes. In the context of GA2LEN (Global Allergy and Asthma European Network), we carried out a meta-analysis of all available genome-wide linkage screens, conducted worldwide for asthma and two atopy-related traits (total IgE level and positive skin test response [SPT]), using the Genome Search Meta-Analysis method (GSMA). The GSMA is a rank-based analysis assessing the strongest evidence for linkage within bins of traditionally 30cM width across the genome. To explore how these results could be affected by bin definition, we analysed the data using different bin widths (20cM and 40cM) and using a shifted 30cM bin by moving bin boundaries by 15cM.
This study identified four regions of interest found for various grouping of populations of different geographical origins: 5q23-31 (p=0.002) for asthma, 6p21 (p=0.008) for IgE, 3q21-q24 (p=0.001) and 17q12-q24 (p=0.004) for SPT. The results remained similar when varying bin widths. Interestingly, the 17q region reached the genome-wide evidence of linkage for SPT when restricting the analysis to Caucasian families (p=0.0001).
These regions can provide targets for future linkage disequilibrium mapping studies or prioritise outcomes from genome-wide association studies.
Supported by EU Framework programme for research, contract n° FOOD-CT-2004-506378, the GA2LEN project.

**46**

**Gene-Phenotype Association Using Mutual Information-Based Cluster Analysis**

A. Buil(1), A. Perera(2), H. Brunel(2), J.C. Souto(1), J. Fontcuberta(1), M. Vallverdu(2), J.M. Soria(1), P. Caminal(2)
(1) U. Hemostasia i Trombosi. Hospital de la Santa Creu i Sant Pau, Barcelona, Spain, (2) C. Recerca Enginyeria Biomedica. UPC. Barcelona, Spain

The relation between a gene and a phenotype can be established testing for association between a set of SNPs on that gene and the phenotype. Testing SNPs one by one presents the problem of multiple testing, while haplotype analysis in unrelated individuals has to deal with phase uncertainty. We present an alternative method that avoids these problems.
This method is based on a cluster analysis of individuals by means of a dissimilarity matrix computed through a normalized version of the mutual information between patterns formed by the SNPs of the individuals. For each cluster configuration, one-way ANOVA seeks differences in the phenotypes between clusters. Finally, the SNPs that uniquely characterize the groups of individuals are determined using a floating feature selection procedure with a supervised learning algorithm. We have applied this method to test association between two phenotypes, FVII and FXII coagulation levels, and their respective structural genes, F7 (47 SNPs) and F12 (26 SNPs), in 100 unrelated individuals of the Genetic Analysis of Idiopathic Thrombophilia (GAIT) sample.
In both cases we found significant differences in the phenotype among clusters. The method correctly identified the different underlying genetic structure of the tested genes, corresponding to the existing haplotype configurations. The algorithm is computationally cost-effective and the clustering is controlled through a single parameter.

**47**

**Interleukin 10 promoter polymorphisms in HIV-HCV Co-infected African Americans**

L.M. Bull(1), R.C. Arduino(2), C.C. Aragaki(1),
L.Y. Hwang(1)
(1)University of Texas School of Public Health, Houston, Texas
(2)University of Texas Medical School; Houston, Texas

Methods:We determined the cytokine genotypes from 216 HIV positive patients and 216 HIV and HCV negative healthy controls, and specifically matched 51 HIV/HCV co-infected African American cases to 51 HIV mono-infected

and 102 HIV and HCV negative healthy African American controls to explore the hypothesis that specific SNPs facilitate the acquisition of HIV/HCV infection and may modulate the host ability to control the viruses. We measured nadir CD4 cell levels and zenith HIV RNA viral loads, HCV RNA viral loads or clearance of the virus as surrogate markers for host immune control of the virus.
Results:There was no association between the IL-10 SNPs in negative controls and either HIV or HCV infected patients. Nor was there a significant difference between the HIV/HCV co-infected group and the HIV mono-infected group when stratified by surrogate markers of immune control, nadir CD4 cell count level or zenith HIV viral load. However, there was a significant difference in distribution of the IL10 -592 variants between the HIV/HCV co-infected patients and the HIV mono-infected patients in a dominant model (p=0.02). The haplotypes of the variants in the dominant model were also associated with co-infection; (p=0.04) however, the significance was lost in regression model with other HIV/HCV covariants (p=0.08).
Conclusion: There is evidence suggesting that the IL10 -592 and IL10 -1082 gene variants are significantly associated with being co-infected with HIV/HCV, and may influence spontaneous HCV clearance.

**48**

**Host Genetic Markers in the New Millennium: Cytokines, Chemokines and Co-receptors Influencing Disease Progression in HIV, HCV and HIV/HCV Co-infection**

L.M. Bull(1), C.C. Aragaki(1), L.Y. Hwang(1)
(1)University of Texas School of Public Health, Houston, Texas

HIV and HCV both cause chronic immune activation which closely affects the rate of disease progression. Host genetic variation is strongly associated with the broad clinical spectrum of disease experienced. Here we provide a systematic review of the genetic factors that effect either HIV or HCV disease progression published from 2000–2007 and classify the results into: allelic variations in cytokines, chemokines, and co-receptors.
Discussion entails the gene's application in HIV/HCV co-infection research.

**49**

**Generalization to extended pedigrees of a latent class model with familial dependence for improved detection of linkage under heterogeneity**

A. Bureau(1,2), A. Tayeb(1,3), J. Croteau(1), C. Mérette(1,4), A. Labbe(1,3)
(1)Centre de recherche UL-Robert-Giffard, Dept. of (2)Social & preventive medicine, (3)Math. & stat. and (4)Psychiatry, Univ Laval, Canada

Clinical diagnoses of complex diseases often group together health problems that are genetically heterogeneous. This has led researchers to collect various measurements related to the diagnosis, such as detailed symptoms or endophenotypes. Latent class (LC) analysis can be applied to these measurements to define more homogeneous disease sub-types, influenced by a small number of genes that will thus be more easily detectable. We have previously developed a LC model allowing dependence between the latent disease class status of relatives within families, and have proposed strategies to incorporate the posterior probability of class membership in linkage analysis. Model fitting was implemented for nuclear families. In a simulation study, this approach was more powerful to detect disease genes than the standard heterogeneity approach of Smith and identity-by-descent sharing methods applied to the disease diagnosis. Here, we present an algorithm to perform computations under our LC model in extended pedigrees.
The algorithm recursively processes nuclear families like the Elston-Stewart algorithm, and its complexity is linear in pedigree size. We present simulations illustrating the behaviour of the LC approach on various pedigree structures. We also present the LC analysis of autism symptoms in pedigrees from the Autism Genetics Research Exchange, and a linkage analysis of phenotypes derived from the latent classes.

**50**

**Mapping SNPs on Chips to Gene Regions Using Linkage Disequilibrium**

William S. Bush, Eric S. Torstenson, and Marylyn D. Ritchie
Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN

Whole-genome association studies provide an unprecedented opportunity for new gene-centric analysis techniques. By mapping markers on a whole-genome genotyping platform to nearby genes, every gene becomes a candidate gene, allowing groups of genes to be analyzed together. Determining the genomic regions and genes a marker represents is a non-trivial task. While most platforms provide some information about marker-gene relationships, a uniform and user-defined process for assigning markers to genes is desired. We have implemented a set of algorithms for generating marker-to-region and marker-to-gene assignments based on patterns of linkage disequilibrium (LD) from the International HapMap Project. Two algorithms were implemented to define LD block boundaries: an LD-spline algorithm and an LD-stringency algorithm. Both approaches accept a user-specified $D'$ or $r^2$ threshold, allowing control over block definitions, and were developed as stored procedures for a MySQL database system. Using these methods, we have generated marker-to-region and marker-to-gene mappings for common whole-genome platforms using common LD thresholds. These approaches will provide the ability to explore whole-genome association data in a gene-centric manner using not only base pair locations to define "genes", but also known linkage disequilibrium patterns in the genome.

**51**

**A score statistic for linkage analysis of age at onset data**

A. Callegaro, J.C. van Houwelingen, J.J. Houwing-Duistermaat
Dept of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

Nonparametric linkage (NPL) analysis compares the identical by descent (IBD) sharing to the expected IBD sharing under the hypothesis of no linkage. For many diseases the population based cumulative hazards are known from registries and the correlation between age at onset of sibling pairs is known from twin studies. For example for breast cancer, this information is available. Our aim is to extend the NPL methods by taking into account the age at onset of selected sibling pairs using this information.
Li and Zhong (2002) proposed a likelihood ratio statistic based on an additive frailty model for genetic linkage analysis. We extend their model by including individual frailties. From this model we derive the score statistic. The new statistic appears to be a weighted NPL statistic with weights depending on the marginal cumulative hazards and the frailty parameter. By means of simulation we compare the new statistic to the NPL statistic.

Li H, Zhong X, 2002, Biostatistics 3, 57–75

**52**

**The heritable contribution to lower urinary tract symptoms (LUTS) in women**

L Cannon-Albright, P Norton, I Nygaard
University of Utah School of Medicine

We investigated the genetic contribution to LUTS using the Utah Population Database, a genealogy of 2.3 million individuals linked to Utah patient encounters. The relative risk (RR) of LUTS phenotypes were calculated by comparing the number of observed affected female relatives to population rates. We also performed a test for excessive relatedness using the Genealogical Index of Familiality (GIF).
First, second and third degree female relatives of women with different phenotypes of stress (SUI) and urge (UUI) urinary incontinence had significantly elevated RRs. However, only first degree relatives had an increased risk of nocturnal enuresis, primarily observed in individuals younger than age 20 years, suggesting common environmental, rather than genetic, influences. Nocturia demonstrated no evidence of familial or genetic clustering. The excess relatedness (GIF) statistic confirmed the increased RRs.
RRs from the UPDB are likely underestimated because controls include age- and parity-matched individuals from the entire database, including some censored phenotypes. This analysis avoids the problem of ascertainment bias. These findings support other reports of familial risk for UI; however, we found no evidence for a genetic contribution to nocturia or nocturnal enuresis, conditions reported to have a genetic contribution.

**53**

**An integrated analysis of genotype data from multiple platforms**

C. Berzuini (1,2,4), M. Tremelling(2) , L. Bernardinelli (1,3), M. Parkes(2)
(1) MRC Biostatistics Unit, Cambridge, UK, (2) IBD Research Group, Gastroenterology Dept., Addenbrooke's Hospital, Cambridge, UK, (3) Dip. Scienze Sanitarie Applicate, Univ. Pavia, I, (4) Dip. Informatica Sistemistica, Univ. Pavia, I

We consider problems of analysis of complex datasets from genetic association studies, in situations involving multiple genotyping platforms on not-necessarily identical sets of loci. Such situations arise when successive stages of study replication focus on narrower and narrower DNA regions, and are, therefore, conveniently performed using different platforms. Another reason may be the desire to incorporate available control data generated on a different platform in a related study. Our motivating study on Ulcerative Colitis (UC) cases involves multiple control datasets and multiple platforms. In such situations the data may often exhibit self-conflicting patterns, likely to be due to allele calling artifacts in one platform, or differential allele calling bias between cases and controls. An interpretation of the self-conflicting evidence is often critical to decisions on whether or not the study should be further pursued. We propose a Bayesian model/algorithm for an integrated analysis of the heterogeneous data, as an aid to the interpretation of the data pattern. The model represents the true haplotype sequences as generated by a first order Markov process along DNA. For each locus, the model also contains the observed imprecise, platform-specific genotypes which depend on unknown platform/locus specific errors. Application of the model allows a reconstruction of the underlying phased haplotype.

**54**

**A Standard XML Data Format for Genetic Epidemiology**

K.C. Cartier
Dept. of Epi & Biostat, Case Western Reserve University, USA

The abundance of computer programs for genetic data analysis represents a great resource for the field of genetic epidemiology. However, along with this resource comes an implicit and largely unaddressed problem: lack of a common interface for data sharing. Input formats are normally driven by the known or expected structure of the data to be studied, and output formats are often dictated by the types of reports that are needed. As a result, two genetic epidemiology programs chosen at random will not usually be able to read the same input file, nor can the output of one normally be used as input to the other. The solution to this problem is to create a standard data format for genetic epidemiology using eXtensible Markup Language (XML). XML is a technology that enables any data to be "self-describing", and therefore potentially independent of any particular application or design. Further, the extensible nature of XML as a specification language allows it to support an evolving standard. Under the XML system, an application simply scans a given input document for the fields it's interested in, and ignores the rest. During the past decade, XML has become something of a de facto standard data formatting method for World-Wide Web traffic, and is being used increasingly as the underlying data format for document-oriented applications such as Microsoft Office. This paper presents an overview of how XML works and could be adopted to the domain of genetic epidemiology, and concludes with a discussion on the expected advantages and disadvantages of this course of action.

**55**

**EFFECT OF GSTT1 GENE ON MELANOMA RISK IN PRESENCE OF MC1R VARIANTS & HOST FACTORS IN FRENCH FAMILIES WITH CDKN2A MUTATIONS**

V. Chaudru(1), MT. Lo(1), F. Lesueur(2),
H. Mohamdi(1), A. Chompret(2), MF. Avril(3),
F. Demenais(1), B. Bressac-de paillerets(2)
(1)INSERM U794,Univ.Evry,(2)Inst.Gustave Roussy,
Villejuif, (3)Hôpital Cochin, Univ. Paris 5, France

UV exposure is the major environmental risk factor of melanoma. Few studies have investigated effect of the glutathione S-transferase (GST) genes (implicated in detoxification of metabolites as generated by UV) on melanoma risk and have led to controversial results. Moreover, the influence of GST genes on the penetrance of CDKN2A, a major high-risk melanoma gene, has not yet been studied. We examined the effect of GSTP1, GSTM1 & GSTT1 genes on melanoma risk in 25 melanoma-prone families with CDKN2A mutations, in presence of other risk factors such as MC1R gene variants & nevi phenotypes. Logistic regression analyses were applied: 1) to the whole sample of 195 genotyped subjects; 2) to the subset of 96 CDKN2A mutation carriers. We used a generalized estimating equations approach to take into account family dependence. GSTM1 & GSTP1 genes had no significant effect on melanoma. However, risk of melanoma was significantly decreased in subjects with ≥1 null GSTT1 allele vs those with 2 active alleles, in presence of CDKN2A mutation & MC1R variants: odds-ratios (OR)=.24 (p=.001) in sample 1; OR=.30 (p=.03) in sample 2. ORs remain similar when nevi phenotypes, which also increased significantly melanoma risk, were included in the model. This GSTT1 protective role may be due to increasing cell death when the number of deleterious genetic events in melanocytes is accumulating. However, this finding needs to be confirmed.

**56**

**Estimating Genetic Relative Risk in Multistage Sampling Design: Application of Composite Likelihood Approach to a Study of Early-Onset Breast Cancer Families**

Y.-H. Choi and L. Briollais
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto

In genetic studies, multistage sampling permits the allocation of resources to families that are most informative for a given objective while allowing population-based inference. However, making inference about some parameters of interest (i.e. gene characteristics) under this design is particularly complex. We illustrate this problem by showing an application to a family study of early-onset breast cancer (<40 years). A total of 813 families were collected following a two-stage design in three population-based breast cancer registries (Australia, California, Ontario), as part of the NCI-funded Cancer Families Registries (CFRs) initiative. In the first stage, affected probands are randomly selected from the cancer registries and in the second stage, probands and their relatives were oversampled in strata defined by their genetic risk criteria, age and ethnicity. In this study, we make inference about the genetic relative risk associated with 1) BRCA1/2 mutation in BRCA1/2-carrier families, and with 2) a third unknown gene in BRCA1/2-non-carrier families, using a design-based approach. We first describe a composite likelihood formulation for time-to-onset data that can account for this particular design. We then compare the design-corrected and uncorrected estimators of the genetic relative risks (point estimate and variances). We finally discuss the optimality of such multistage design in light of our results.

**57**

**SumLINK: Localizing prostate cancer genes using the ICPCG pooled linkage resource**

GB Christensen, NJ Camp, and the ICPCG
Dept of Biomedical Informatics, U of Utah

We propose a new linkage-based statistic, sumLINK, to identify disease-susceptibility loci. Our approach focuses on linked pedigrees (pedigree-specific multipoint LOD> 0.588; p 0.05) to identify regions of consistency. The sumLINK statistic is the sum of multipoint LOD scores for linked pedigrees at a given point in the genome. To assess the significance of the sumLINK statistic we employ a unique shuffling method to simulate the null distribution. For each pedigree, we calculate the multipoint LOD values at 1-cM intervals across the genome. Data from all chromosomes are connected in a loop, which is then broken at a random point. Each pedigree is realigned to its new starting point and a null sumLINK statistic is calculated for each cM position in the shuffled data. This procedure maintains the linkage potential of each pedigree, but removes the consistency across linked pedigrees. The process is repeated 1000 times to determine the empirical distribution of the sumLINK for the given pedigree resource. We applied our sumLINK approach to genome-wide autosomal linkage data on 1,232 pedigrees from the International Consortium for Prostate Cancer Genetics (ICPCG). Several loci were identified as significant or suggestive at the genome-wide level. Several of these have not previously been implicated using standard HLOD analyses for the same resource. One advantage of loci identified with the sumLINK approach is that they have good potential for subsequent gene localization using statistical recombinant mapping, as, by definition, there exist multiple linked pedigrees contributing to each peak.

**58**

**Congenital Heart Defects, Maternal Genetic Susceptibility and Pro-oxidant Lifestyle Factors**

M.A. Cleves, M.A. Karim, S.L. Macleod, C.A. Hobbs
Arkansas Center for Birth Defects Research and Prevention, University of Arkansas for Medical Sciences, Arkansas Children's Hospital Research Institute

Background: Congenital heart defects (CHDs) are the most common birth defects. Women with CHD-affected pregnancies may be genetically susceptible to alterations in

folate-dependent metabolic pathways and this susceptibilities may be modified by pro-oxidant lifestyle factors.
Methods: We conducted a population-based case-control study of 620 women who had pregnancies affected by nonsyndromic CHDs, and 391 women without such a history. All pregnancies ended between January 1, 1999 and March 1, 2005. Maternal interviews were conducted and buccal cell samples collected. DNA samples were genotyped for methylenetetrahydrofolate reductase (MTHFR) 677 C>T, transcobalamin II (TCII) 776 C>G, and betaine homocysteine methyltransferase (BHMT) 742 G>A polymorphisms.
Results: Women who carried the TCII 776 G allele and either smoked or consumed alcohol periconceptionally had an increased risk of having an infant with a CHD. Women who were obese periconceptionally had an increased risk of having an infant with a CHD if they were homozygous for the MTHFR 677 C>T polymorphism or if they carried the BHMT 742 A allele.
Conclusions: Our findings indicate that pro-oxidant lifestyle factors modify the association between CHD-affected pregnancies and genetic susceptibilities. Our finding converge with evidence from previous studies to indicate that folate-dependent genetic and metabolic susceptibilities increase the risk of CHD-affected pregnancies.

**59**
**Logistic regression and Bayesian modelling of metabolic pathways**
N Cremer, L Beckmann, S Kropp, J Chang-Claude
Department of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany

The analysis of complex diseases requires powerful analytical tools to account for interactions and biological knowledge of disease etiology. Hierarchical Bayes model averaging (BMA) has been proposed for model selection and effect estimation, incorporating uncertainty of model choice. We compared BMA to conventional logistic regression, and to regression after backward selection.
We considered genetic variants in NAT1 and NAT2, and in CYP1A1, CYP1B1, GSTM1 and GSTT1, involved in the metabolism of aromatic amines and polycyclic aromatic hydrocarbons, respectively, in a case-control study of breast cancer. We adjusted for age, menopausal status, and breast cancer family history, and also stratified by the latter.
Results from logistic regression and BMA were consistent with respect to variable significance, as measured by p-values and Bayes factors (BFs) in the respective frameworks. Stepwise regression selected variables also highlighted by BMA in all analyses.
In spite of consistency with respect to significance, BMA underestimated effects compared to regression analysis, possibly due to prior specification. Prediction measured by the area under the curve ranged between 0.56 and 0.65 (BMA), and 0.58 and 0.73 (logistic regression) for adjustment and stratification scenarios, indicating that further variables should be included in the analysis.
In this dataset, regression analysis and BMA detected the same significant variables. Magnitude of effects differed, possibly due to choice of priors.

**60**
**Probabilities for Polymorphisms**
D.J. Crouch & C. Cannings
School of Medicine, University of Sheffield.

If genotype viabilies are drawn uniformly and independently for an autosomal locus with two alleles then the probability of a polymorphism (for an infinite, random mating population with discrete genetaions) is one third; we simply require the heterozygote to be the fittest of the three genotypes. If there are more than two alleles then the possible dynamics are more complex and the probability of a polymorphism with all three alleles is much lower. For the case of three alleles there are 13 permutationally distinct orderings of the heterozygote/corresponding homozygotes viabilities. Here we estimate the probabilities of various quantities of interest in the three allele cases using simulation, partitioned by the underlying orderings. The case of four alleles will also be considered.
For the case where the genotypes have different viabilities and with only two alleles there are six qualitatively distinct dynamics characterised by the ordering of the stability =S/instability=I of the possible equilibria (which can be ordered by allele frequencies). Thus we have SISIS, ISISI, SISI, SIS, ISI and SI). We shall provide estimates of the probabilities of these cases, and for the positions of the equilibria using simulation.

**61**
**Identification of susceptibility genes for Colon Cancer: Results from the Colon Neoplasia Sibling Study**
D. Daley(1,3) S. Lewis(1), P. Platzer(1), M. MacMillen(1), J. Willis(1), R.C. Elston(1), S.D. Markowitz(1,2), and G.L. Wiesner(1)
Case Western Reserve University (1) and Howard Hughes Medical Institute (2) and the University of British Columbia (3)

The Colon Neoplasia Sibling Study (CNSS) conducted a comprehensive genome-wide linkage scan, incorporating unaffected individuals, clinical information (histopathology, number, size of polyps, and other primary cancers), in conjunction with age at onset and family cancer history to classify 194 kindreds into five phenotypic groups (severe histopathology, oligopolyposis, multiple cancer, young, and colon/breast). The data were analyzed using Haseman-Elston regression analysis, with permutation testing and correction for multiple testing. Significant linkage was identified to chromosomes 15q14-q22, 17p13.3, 1p31.1, and 10q22.1-31. The most significant signal is on chromosome 1p31.1 (p-value=0.00007, multiple cancer group). We identified linkage to the HMPS/CRAC1 locus (15q14-q22) in the whole sample (p=0.018), young (p=0.0007) and oligopolyposis (p=0.003) groups, meeting thresholds for significant replication, indicating this locus may be involved in the development of colon polyps in patients of European descent diagnosed before the age of 51. In the breast/colon subgroup we found linkage to 17p13.3, identifying HIC1 as a novel candidate gene. Other signals include BMPR1A, ANXA1 and EXO1, demonstrating the use of clinical information, unaffected siblings, and family history, can increase the analytical power

of a linkage study. These methods can generally be applied in the analysis and dissection of complex trait.

**62**

**Candidate Genes for Asthma, Atopy and Allergic Disease**

D. Daley (1), P.D. Paré (1), A.J. Sandford (1), A.L. Kozyrskyj (2), C. Laprise (3), Y. Bosse (4), A. Montpetit (4), A. Becker (2), T.J. Hudson (4), and M. Lemire (4)
(1) University of British Columbia, (2) University of Manitoba, (3) University of Quebec at Chicoutimi, (4) McGill University and Genome Quebec Innovation Centre

To better understand the development of asthma and allergic diseases, we conducted a genetic association study using 3 Canadian populations: 1) a high risk birth cohort (Canadian Asthma Primary Prevention Study), 2) a population-based birth cohort of children from the Study of Asthma Genes and Environment, and 3) The Saguenay-Lac St. Jean Quebec family based sample from a French Canadian founder population. A primary objective of this study was to examine candidate genes with the same set of SNPs, a common genotyping platform, and stringent standardization of phenotypes. Our panel is comprised of candidate genes associated with asthma and allergic phenotypes (asthma, atopy, atopic asthma and airway hyperresponsiveness) with strong biologic plausibility and/or prior evidence for association. For a full list of genes and SNPs see http://genapha.icapture.ubc.ca/. For each gene a maximally informative set of common single-nucleotide polymorphisms (SNPs) was selected and genotyped using the Illumina GoldenGate assay. Genetic analysis was carried out using Family Based Tests of Association with correction for the number of SNPs and phenotypes tested. We have identified significant associations with IL13 for asthma, atopy, and atopic asthma, and association with IL18 and IFNGR2 for atopy. Complete results including phenotype and cohort specific findings to be presented.

**63**

**Region-specific p-values for genome-wide association studies**

O. De la Cruz(1), X. Wen(1,2), B. Ke(1), M. Song(1), D. Nicolae(1,3)
(1)Dept. of Statistics, University of Chicago, USA,
(2)Dept. of Human Genetics, University of Chicago, USA,
(3)Department of Medicine, Section of Genetic Medicine, University of Chicago, USA

In the setting of genome-wide association studies, we propose a method for assigning a measure of significance to regions of the genome. A high level of significance (a small p-value) will be evidence that genetic variation in the region (typed or untyped) might be associated with the disease under study.
This approach has at least three advantages: First, strong signals from different SNPs in one region that are in low linkage disequilibrium are combined to obtain a much stronger signal for the region, therefore increasing power. Second, it captures the information of untyped variation in each region. Third, it leads to results that are easy to interpret and can be readily combined across platforms.

**64**

**The applicability of the 4-gamete rule on HapMap data shows that linkage disequilibrium is primarily caused by lack of recombination and barely by genetic drift**

A.R. de Vries and G.J. te Meerman
Dept. of Genetics, University Medical Center Groningen and University Groningen, the Netherlands

The 4-gamete rule, which has been introduced more than 20 years ago [1], states that two nearby mutations lead to only three out of four possible haplotypes. The fourth haplotype can only be created by a recombination event, which becomes more likely with larger distance between the two Single Nucleotide Polymorphisms (SNPs). Data from the Encode regions of the HapMap project show that the 4-gamete rule holds for SNPs located at least 100 kb apart, up to 400 kb in regions with higher linkage disequilibrium. We show that current allele frequencies predict the missing haplotype as if allele frequencies have not substantially changed since the mutation of the most recent one among two mutations. This illustrates that in humans recombination has occurred infrequently for sometimes large genomic distances and that lack of recombination is still a major source of Linkage Disequilibrium, while genetic drift is not.

Reference:
[1] R.R. Hudson and N.L. Kaplan, 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

**65**

**Two New PPAR Gamma Gene Variants Confer Risk to Type 2 Diabetes in Khatri Sikhs from India: The Sikh Diabetes Study SDS**

D.K. Sanghera (1), J. Figgins (2), S.K. Nath (3), M. Gains (3), J. Singh (4), S.K. Ralhan (5), G.S. Wander (5), N.K. Mehra (6), M.I. Kamboh (2)
(1)Depart. of Pediatrics, Univ. of Oklahoma Health Sciences Center, USA, (2) Dept. of Human Genetics, Univ. of Pittsburgh, USA, (3) Oklahoma Medical Research Foundation, USA, (4) Guru Nanak Dev Univ., India, (5) Hero DMC Heart Inst., India, (6) All India Inst. of Medical Sciences, India

Type 2 diabetes mellitus T2DM has become a major public health problem in urban and sub-urban India. Several studies conducted recently have reported a common proline-to-alanine substitution Pro12Ala in the peroxisome proliferator-activated receptor gamma PPARG gene to be associated with obesity and T2DM. The present study was carried out to investigate the role of Pro12Ala along with six other tagging single nucleotide polymorphisms tagSNPs in the PPARG gene with T2DM in a family-based Khatri Sikh sample from India. We genotyped these variants in 990 Khatri Sikh subjects from 228 families 648 affected cases and 296 unaffected controls. Using the pedigree based disequilibrium test PDT, we evaluated each polymorphism for its association with T2DM. We have identified two additional new markers rs12490265; p=0.0004 and rs 2938395; p=0.0276 that showed significant distortion of transmission of the risk allele from

parents to affected offspring in addition to replicating the association of Pro12Ala polymorphism p=0.0195 and T2DM in Khatri Sikhs. It appears that these two markers contribute to the T2DM risk independent of Pro12Ala polymorphism since the LD between these SNPs and Pro12Ala is very weak. Our new findings suggest the possibility of occurrence of multiple functional sites in the PPARG gene that may influence T2DM susceptibility in Khatri Sikhs.

**66**

**Correcting for measurement error in individual ancestry estimates in structured association tests**

Jasmin Divers(1), Laura K. Vaughan(2), Miguel Padilla(2), José R. Fernandez(2), David B. Allison(2), David T. Redden(2)
(1) Section on Statistical Genetics and Bioinformatics, Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina,
(2)University of Alabama at Birmingham, Birmingham, Alabama

We present theoretical explanations and show through simulation that the individual admixture proportion estimates obtained by using ancestry informative markers should be seen as an error-contaminated measurement of the underlying individual ancestry proportion. These estimates can be used in structured association tests as a control variable to limit type I error inflation or reduce loss of power due to population stratification observed in studies of admixed populations. However, the inclusion of such error-containing variables as covariates in regression models can bias parameter estimates and reduce ability to control for the confounding effect of admixture in genetic association tests. Measurement error correction methods offer a way to overcome this problem but require an a priori estimate of the measurement error variance. We show how an upper bound of this variance can be obtained, present four measurement error correction methods that are applicable to this problem, and conduct a simulation study to compare their utility in the case where the admixed population results from the intermating between two ancestral populations. Our results show that the quadratic measurement error correccorrection method (QMEC) performs better than the other methods and maintain the type I error to its nominal level.

**67**

**Coronary Artery Disease in South Asians: Novel Apo A-1 Polymorphisms Associated with Low HDL**

R Kaur, Y Dong, H Zhi, G Reed, V George
Medical College of Georgia

Background- Coronary artery disease (CAD) is the leading cause of morbidity and mortality in the world. Even though its rates have decreased in the United States and other developed countries over the past 30 years, event rates are still high in South Asians. The presence of traditional CAD risk factors, increased diabetes, and metabolic syndrome in South Asians may not fully explain the excess CAD. HDL levels are among the most predictive risk factors for CAD and South Asians are known to have low HDL. The objective of this study was to identify Apo lipoprotein A-I (A-I) polymorphisms and explore its association with low high density lipoprotein (HDL) levels and other risk factors for CAD in South Asians living in the United States.
Methods and results- 30 South Asians immigrants between the ages of 40–65 years were included. 12- Hour fasting blood samples were collected for blood tests and for Apo A-I DNA sequencing. Apo A-I polymorphisms revealed six novel SNPs (G1-T319C,G2-T655C, G3-T756C, G4-C938T, G5-T1001C, and G6-C1149T), one (C938T) was found to be significantly associated with low (<40 mg/dl) HDL levels (p=0.004). The association was also seen with total cholesterol (p=0.26) and LDL levels (p=0.32).
Conclusion- We discovered 6 novel polymorphisms of Apo A-I in south Asian immigrants. Discovery of novel polymorphisms will help us to understand further the causes of increased CAD risk in South Asians so that preventative strategies targeted especially to this high risk group can be developed. Further larger studies are needed to explore their findings and assess the associations with CAD.

**68**

**Establishing an adjusted p-value threshold to control the family-wide Type 1 error in genome-wide association studies**

P. Duggal*, E.M. Gillanders*, T.N. Holmes, J.E. Bailey-Wilson.
*These authors contributed equally.
Statistical Genetics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD USA

Although it is widely recognized that some threshold or correction for multiple testing is necessary, in association studies it is not clear what method to utilize. One option is to perform a Bonferroni correction using all n SNPs, however this approach is highly conservative and for genome-wide association studies (GWAS) could result in a p-value threshold of 2x10-8 (0.01/500,000 SNPs). This extreme threshold would "overcorrect" for SNPs that are not truly independent, resulting in loss of power. Using data from the International HapMap, we evaluated the number of haplotype blocks and independent SNPs across the genome for the CEPH and Yoruban populations to identify the effective number of independent tests that can be used in the Bonferroni correction. Additionally, we will evaluate the number of haplotype blocks and independent SNPs for several Affymetrix and Illumina GWAS SNP panels. We will present adjusted Bonferroni p-value thresholds that are SNP panel- and population-specific to properly control the family-wide Type 1 error in a GWAS.

**69**

**Resolving the Power of Multifactor Dimensionality Reduction in the Presence of Many Noise Variables or Genetic Heterogeneity**

TL Edwards, SM Dudek, MD Ritchie
Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232

In human genetic studies of common complex disease a consideration is the detection of joint effects at several loci. The search space to find such multi-locus associations is very large relative to the number of single locus effects. Conventional parametric approaches which were not designed to screen these spaces suffer from low power due to correction for multiple comparisons. The Multifactor Dimensionality Reduction (MDR) algorithm exhaustively searches these spaces and has been shown to have power to detect interactions. Prior to this study, the performance of MDR to detect gene-gene interactions given large numbers of noise loci or varying degrees of genetic heterogeneity was unknown. Several 2-locus and 3-locus epistatic genetic models with a range of effect sizes were simulated. We explored increasing numbers of SNPs (100, 500, 1,000, 5,000, and 10,000) in datasets of 500 cases and 500 controls. The results show that MDR has power to detect these interactive effects in datasets that exceed the largest candidate gene studies when heritability and effect sizes are moderate to large. Three levels of heterogeneity and four sample sizes were also simulated. The results indicate that selection of a study population where heritability and effect size estimates are reasonable are more relevant to MDR performance than sample size. These results also demonstrate that MDR is robust to locus heterogeneity when the definition of power is liberal.

**70**

**genomeSIMLA: a data simulation package to explore the human genome**

TL Edwards, WS Bush, SD Turner, ES Torstenson, SM Dudek, MD Ritchie

Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232

In the quest for disease susceptibility loci many novel statistical and computational methods are in development. Data simulation is necessary to evaluate these methods before they can be applied to real data. It is difficult to emulate the properties of genetic data in human populations which arose from complex demographic history. Modeling all linkage disequilibrium (LD) parameters observed in real data is infeasible; also, synthetic models often lack the complexity of real data. Rather than modeling population history or LD we use a forward-time population simulator. We have developed a software package, genomeSIMLA, that simulates data on a genome-wide scale in both case-control or family-based study designs with LD resembling that observed in human populations. Random mating, population growth, and recombination are used to create a pool of chromosomes with complex LD. Marker positions can be used to estimate the frequency of recombinant gametes under Haldane or Kosambi models. These values are used to emulate LD observed in human populations. After a pool of chromosomes has developed suitable LD, datasets can be drawn by sampling chromosomes with replacement. Disease-susceptibility effects of genetic variables with any mode of inheritance and interactions may be modeled using a prospective logistic regression model. Purely epistatic models can also be simulated. genomeSIMLA is a data simulation package designed to evaluate novel analysis approaches in a realistic context on a large scale relevant to modern genetic epidemiology.

**71**

**Genetic variation in the hepatic lipase gene is associated with HDL-cholesterol levels in the NHLB Family Heart Study**

M.F. Feitosa (1), R.H. Myers (2), M.A. Province (1), I.B. Borecki (1)

(1) Washington University, St. Louis, MO; (2) Boston University, Boston, MA.

We previously reported linkage for high density lipoprotein cholesterol (HDL-C) on 15q21 (LOD=4.77). Hepatic lipase (HL), which has a major role in lipoprotein metabolism, resides within the linkage region and constitutes an obvious candidate gene. Although HL activity is widely known to influence HDL metabolism, the relationship between HL variants and HDL-C levels remains unclear. In the current study, we carried out family-based tests of association with both quantitative HDL-C and a dichotomous dyslipidemia trait (affected men: HDL $<$ 40 mg/dL and women: HDL $<$ 50 mg/dL). We genotyped 433 families (2,192 subjects) from the NHLBI Family Heart Study with 19 tag SNPs spanning 593 kb within the HL gene. While significant associations were found with several SNPs in the first intron (p=0.00067), some of the associations appeared to be stronger in women than in men (p=0.00018 vs. p=0.1965, respectively). Similar results were found using the dichotomous trait and TRANSMIT, despite allowance for sex-specific thresholds. By contrast, a SNP association also in intron 1 but ~81 kb away and in a different haplotype block (linkage disequilibrium $r < 0.04$), was nominally significant in men (p=0.03) but not in women (p=0.96). The linkage evidence was stronger in non-diabetic subjects, and our current data also suggest possible interactions with diabetic status. It appears likely that the several associated SNPs in intron 1 may lie on different haplotypic backgrounds, with sex-specific effects on HDL-C variation.

**72**

**A genetic epidemiologist's view on the human pseudoautosomal regions**

C. Fischer (1), T. Wienker (2), A. Flaquer (2)

(1) Institute of Human Genetics, University of Heidelberg, Germany

(2) Institute of Medical Biometry, Informatics and Eopidemiology, University of Bonn, Germany

Two small intervals of sequence identity on the telomeres of X- and Y-chromosomes, the human pseudoautosomal regions, have drawn considerable interest in genetics and evolution biology but are mostly ignored in linkage and association studies. The pseudoautosomal regions behave like autosomes in the sense that they pair and crossover during meiosis. However, in contrast to autosomes, the recombination activity in PAR1 is extremely different between male and female. Additionally, in male it exhibits one of the highest recombination frequencies throughout the entire genome.

Partly conflicting genetic maps have been estimated by using data from three-generation-pedigrees, by sperm typing and, very recently, using haplotypes from surveys of single nucleotide polymorphism variation. We review the existing tools like genetic and physical maps, linkage disequilibrium, linkage and association analysis, implemented statistical methods, and whether they are appropriate for the loci in PARs. For multipoint linkage analysis sex specificity has to be taken into account twice, firstly by using sex-specific genetic maps, and secondly by taking into account the sex-specific pseudoautosomal inheritance pattern. Up to now, microsatellite panels and SNP-chips do not contain sufficiently many markers in PAR1/2. The number of markers in PAR1, needed in indirect association studies, has to be much larger than for autosomal regions of the same size since linkage disequilibrium is very low. For genome wide studies it is absolutely necessary to include the pseudoautosomal regions since such expensive studies cannot afford to oversee pseudoautosomal linkage or association. A sufficient number of markers, tailored statistical methods, and their integration in software tools will facilitate this purpose.

**73**

**Meta analysis of genome wide linkage studies for autoimmune diseases**

P. Forabosco (1,2), E. Bouzigon (1,3), M.Y. Ng (1), J. Hermanowski (1), S.A. Fisher, L.A. Criswell (4), C.M. Lewis (1)
(1) Med. & Mol. Genet., King's College London, UK
(2) IGP-CNR, Italy
(3) INSERM U794, France
(4) Dept. of Med., UCSF, USA

Strong evidence of familial clustering of autoimmune diseases (AIDs) has been observed, and loci identified for specific AID often overlap with loci implicated in other AIDs, suggesting the presence of pleiotropic genes.
We used the genome scan meta-analysis (GSMA) method to combine linkage results from non-overlapping genome screens for AIDs. In total, 57 independent studies were identified, 43 with complete genome-wide results for the following disease groups (# studies): rheumatoid arthritis (6), systemic lupus erythematosus (3), seronegative spondyloarthropathies (18), autoimmune thyroid disease (3), vitiligo (1), type 1 diabetes (1), multiple sclerosis (7), and celiac disease (4). Together, the data consist of 7,829 families with 19,347 affected individuals.
The GSMA assesses the strongest evidence for linkage within bins of traditionally 30 cM width. To explore how results could be affected by bin definition, we considered different bin widths (20cM and 40cM) and shifted 30 cM bins obtained by moving bin boundaries. Weighted analysis controlled for the number of affecteds in each study, and the number of studies in each disease, with analysis of various AID clusters also performed.
In addition to the HLA region, which was confirmed with genome-wide evidence for linkage ($p < 0.00001$), we consistently observed suggestive evidence for linkage on chromosome 16p (p=0.0013, 30cM bins). This region may harbour a gene/s conferring risk for several AIDs.

**74**

**Genotype-by-smoking interaction on blood pressure traits: the Strong Heart Family Study**

N Franceschini (1), JW MacCluer (2), HHH Göring (2), VP Diego (2), SA Cole (2), S Laston (2), LG Best (3), RR Fabsitz (4), ET Lee (5), M Russell (6), KE North (1)
(1) UNC, Chapel Hill, NC, (2) SFBR, San Antonio, TX, (3) MBIRI, SD, (4) NHLBI, Bethesda, MD, (5)U of OK, Oklahoma City, OK, (6)MedStar, Washington, DC.

Smoking is associated with increased blood pressure in populations. We evaluated the genotype-by-smoking interaction across quantitative blood pressure traits in 3,634 American Indian participants of the Strong Heart Family Study (SHFS), recruited in Arizona (AZ), North and South Dakota (DK), and Oklahoma (OK). Smoking exposure (ever versus never), mean systolic (SBP) and diastolic (DBP) blood pressures were obtained at a study exam. Traits with a skewed distribution were rank-normalized, and adjusted for age, age$^2$, sex and hypertension treatment within study center, using linear regression models (SAS 9.1). Variance component linkage analysis (*SOLAR*) was performed using marker allele frequencies derived from all individuals and multipoint IBDs calculated in Loki. Additive genotype-by-smoking interaction was estimated for SBP, DBP and pulse pressure using a maximum likelihood variance decomposition technique implemented in SOLAR. Hypertension prevalence was 38% in AZ and OK but only 26% in DK, and approximately 65% of individuals with hypertension in all three centers were on drug treatment. Mean blood pressures were 121/77 mm Hg (AZ), 120/72 mm Hg (DK) and 127/77 mm Hg (OK). The prevalence of ever smoking was 50% in AZ, 66% in DK and 58% in OK. We found that the genes influencing SBP (P=0.01), among DK participants, and DBP (P=0.006), for OK participants, were differentially expressed in individuals with ever smoking exposure compared to non-smokers. These results suggest that smoking exposure may modify the effects of genes influencing blood pressure traits.

**75**

**Bayesian hierarchical nonlinear models for analysis of pharmacogenomic cytotoxicity data**

B.L. Fridley(1), D. Schaid(1), R. Weinshilboum(2), L. Wang(2)
(1) Div. of Biostatistics, Mayo Clinic, USA
(2) Dept. of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, USA

Pharmacogenomic research has recently incorporated cell-based model systems. Hypotheses generated with the cell-based model system are then tested in individuals treated with the agent ("translational medicine"). A large number of statistical methods have been developed to evaluate non-linear dose-response curves in order to model how response curve parameters are influences by subject-specific characteristics, but few efforts have been made to tie these types of models to genomic studies. We will illustrate the use of Bayesian nonlinear hierarchical models for analysis of pharmacogenomic data with concentration-effect endpoints. Concentration-effect endpoints are any measurable cellular

phenotype that is related to drug concentration, one example being cytotoxicity. Currently, cytotoxicity endpoints are analyzed one drug concentration at a time or dose-response curves are fit to the cytotoxicity endpoints from which summary endpoints (e.g., GI50) are used as the phenotype in the analysis. A more complete analysis of the impact of genetic variation on the entire dose-response curve might better reflect the "true" relationship between drug response and genetic variation and lead to insight into the understanding of the pharmacogenomics of a particular drug/agent. The model will be illustrated utilizing data from a study of the pharmacogenomics of gemcitabine using data collected on the Coriell Human Variation Panel comprised of 203 cell lines by the Mayo-NIH PGRN and with simulated data.

**76**

**Ascertaining families on a major gene and disease status: Factor V Leiden (FVL) thrombophilia as a model to tackle the genetic architecture of complex diseases**

F. Gagnon, D.E. Bulman, P.S. Wells
Department of Public Health, University of Toronto

Venous thromboembolism (VTE) is a common complex disease with known environmental risk factors and a well-characterized major gene variant, FVL. FVL thrombophilia is associated to a single point mutation in the factor V gene leading to the Activated Protein C Resistance phenotype. This disorder has an autosomal dominant inheritance and a population frequency of 2% to 15%, and up to 60% in VTE cases. The predictive clinical value of FVL is limited since only 20–50% of heterozygous individuals develop VTE despite accounting for other known risk factors. Experimental evidence suggests that this variability is more likely due to modifier genes than unknown environmental factors. Thus, we are taking advantage of this well-characterized genetic variant to identify genes implicated in VTE. Our major objective is to identify the modifier genes in FVL thrombophilia, and our main strategy is to capitalize on the several genetically determined plasma hemostatic-related proteins associated to VTE. We have ascertained 7 large French-Canadian families (n=306) through a single proband with both VTE and FVL. Detailed phenotyping of hemostatic-related quantitative traits (QT) (~30), as well as several putative environmental (e.g. hormonal therapy, smoking, physical activity) covariates have been collected on all family members. The specific aims of this paper are to phenotypically characterize these families; to compare FVL carriers vs. non-carriers; and to present results from Bayesian Markov chain Monte Carlo (MCMC)-based oligogenic segregation analyses; e.g. Plasma factor XIII activity was significantly higher in FVL carriers vs. non-carriers. Factor XIII was correlated to several lipid-related QT (e.g. plasma cholesterol), as well as other hemostatic QT including plasma levels of factor V. In a large meta-analysis, we recently reported that a factor XIII A-subunit variant has a protective effect against VTE (Wells et al. 2006). Genotyping for genome-wide linkage scan is currently under way. Future work includes genome-wide linkage analyses based on MCMC approaches, modeling pleiotropy and epistasis, while accounting for FVL effect.

**77**

**Identifying Genetic Risk Factors Underlying Alcoholism Using A Multivariate Phenotype Approach: The COGA Experience**

S. Ghosh(1), L.J. Bierut(2)
(1)Indian Statistical Institute, Kolkata, India (2)Wash U School of Medicine, St. Louis, MO, USA

Using the clinical end-point of alcoholism as a linkage phenotype in the Collaborative Study On The Genetics Of Alcoholism (COGA) project has failed to identify chromosomal regions harboring potential genes for the complex trait. This has motivated analyses of quantitative endophenotypes correlated with alcoholism, which capture greater variation within trait genotypes than the end-point binary trait. However, a single quantitative trait is unlikely to be a powerful surrogate of the end-point trait and it may be more prudent to use a correlated multivariate phenotype for linkage analyses. We propose, along the lines of Sham et al. (2002), a linear regression formulation in which the traditional response variable, that is, some function of the multivariate phenotype and the explanatory variable, that is, the genetic similarity between sib-pairs are interchanged. Based on simulations, we find that the proposed method is more powerful than analyses based on the first principal component of the correlated phenotypes and the end-point binary trait. We perform a genome-wide linkage scan of a multivariate phenotype comprising three alcohol related quantitative traits: number of drinks in a day, electroencephalogram and count of externalizing symptoms associated with anti-social behavior using the proposed regression approach and identify chromosomal regions on Chromosomes 1,4,5 and 15 exhibiting significant linkage. We find that these regions harbor potential genes for alcoholism like the GABRA receptor, the ADH cluster and CHRNA7.

**78**

**Comparison of Methods for Combining Case-Control with Family-Based Association and Linkage Samples**

B. Glaser, P. Holmans
Biostatistics and Bioinformatics Unit, Cardiff University, School of Medicine, UK

The power of genetic association studies can be enhanced by combining the analysis of case-control with family-based association samples. Various methods for this have been developed; as yet, there have been no comparisons of their power. We investigate the power of a number of methods for combination of case-control with parent-offspring trio samples, as well as a simple combination of chi-squared statistics from individual samples. The performance of the methods for analysing multiplex sibships is also studied.
Power and Type I error of the methods are investigated by simulation under a variety of disease models, allele frequencies, genetic effects and sample sizes.
Our results show that the majority of methods for combined case-control and parent-offspring trio analysis give better power than single sample statistics and simple combination of chi-squared statistics, when effect sizes and allele frequencies in the individual samples are similar. However, the simple

combination of chi-squares can be more powerful than combined statistics, when allele frequencies and effect sizes differ between samples. In addition, we identify disease models for which the power of the investigated methods can be compromised. These results highlight the importance of studies investigating the performance of novel methods, in order to indicate the situations where they will be most useful.

**79**

**A test for imprinting derived from a regression-based linkage method**

O.Y. Gorlova(1), L. Lei(1,2), D. Zhu(1), S.-F. Weng(1), S. Shete(1), Y. Zhang(1), C.I. Amos(1)
(1)Dept of Epidemiology, MD Anderson Cancer Center, Univ. of Texas, Houston, TX
(2) Amgen Inc.

In an extension of a regression-based quantitative-trait linkage method to incorporate parent-of–origin effects, we separately regress total, paternal, and maternal IBD sharing on traits' squared sums and differences. We developed a test for imprinting that indicates whether there is any difference between the paternal and maternal regression coefficients. This test asymptotically follows the standard normal distribution inder the null. Since this method treats the identity-by-descent information as the dependent variable that is conditioned on the trait, it can be readily applied to data from complex ascertainment processes. We performed a simulation study to examine the properties of the test for imprinting. We evaluated effects of missing parental genotypes, misspecified population mean for the trait, different major gene variance, selected samples, and samples consisting of mixture of families with different size on the type 1 error rate and power of the proposed test for imprinting. We present a permutation algorithm that allows deriving critical values and empirical p-values when the test statistic deviates from the standard normal distribution. We found that when using empirical critical values, the method shows identical or higher power compared to existing methods for evaluation of parent-of-origin effect for quantitative traits.

**80**

**Genetic determinants of cardiac mass and structure – Preliminary results of the first GWA study**

A. Götz(1,2), W. Lieb(1), I.R. König(2), D.F. Schwarz(2), F. Pahlke(2), S. Szymczak(2), C. Gieger(3), I. Heid(3), T. Meitinger(4), A. Ziegler(2), H. Schunkert(1), J. Erdmann(1)
(1) Medizinische Klinik II, UK-SH Lübeck; (2) Institute of Medical Biometry and Statistics, University of Lübeck; (3) Institute of Epidemiology, GSF Neuherberg; (4) Institute of Human Genetics, TU München

Alterations in left ventricular mass (LV) are associated with increased cardiovascular morbidity and mortality. There is evidence that LV structure and function are in part determined by genetic factors.
A total of 591 probands from the population-based KORA/MONICA Augsburg survey were phenotyped. All individuals were genotyped with the GeneChip® Human Mapping 500K Array from Affymetrix. We performed several quality checks (MAF, Mif per SNP, test for deviation from HWE, visual inspection of signal intensity plots for interesting SNPs).
We calculated residuals on several phenotypes that explain LV mass structure and function adjusting for known covariates. SNPs were tested for association with residuals using parametric (Kruskal-Wallis, Jonckheere-Terpstra) and non-parametric tests (linear regression, F test). To extract SNPs for validation, we inspected regions around the significant SNPs with LD plots and LDU maps. Altogether 85 significant SNPs were identified and further 115 SNPs from the flanking regions were selected for replication.
To our knowledge this is the first GWA study which identified interesting genomic regions with association to LV structure and function. The further strategy includes validation of most promising regions in two larger independent population-based samples.

**81**

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

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

Mutations in the LMNA gene cause autosomal dominant Emery-Dreifuss muscular dystrophy (AD-EDMD), characterized by skeletal and cardiac involvements. We observed intrafamilial variability in a single family carrying LMNA mutation characterized by a wide range of age of onset of myopathic symptoms. This phenotypic heterogeneity suggests the contribution of a likely modifier locus. The aim of the study is to identify the secondary modifier locus of the muscular involvement linked with the major LMNA mutation. The sample is constituted of a French family of 102 individuals. We performed a systematic genome scan of 59 individual, with 280 highly informative microsatellite markers. Thirty of them are affected by the LMNA mutation of which 11 have skeletal muscle and heart disease, 17 only heart disease and 2 are still asymptomatic. The main criterion is the age of onset of skeletal muscle symptoms.
In order to test linkage and association, we will use a family based approach of linkage disequilibrium (QTDC) and, in order to test linkage and segregation, we will use a Monte Carlo Markov chain method as implemented in Loki. The results of these two analyses will be compared.
Analyse of the data is currently in progress. The results will be presented later. Identification of such modifier loci provides additional insights into disease mechanisms; genetic counselling and patient follow up.

**82**

**Direct Testing of Untyped SNPs Using Multimarker Tags**

S. Griffiths, F. Dudbridge
MRC Biostatistics Unit, Cambridge, UK

In association studies it is generally too expensive to genotype all variants in all subjects. We can exploit linkage disequilibrium between SNPs to select a subset that captures the variation in a training data set obtained either through direct resequencing or a public resource such as the HapMap. These tag SNPs are then genotyped in the whole sample. Multimarker tagging is a more aggressive adaptation of pairwise tagging that allows for combinations of tag SNPs to predict an untyped SNP. Here we describe a new method for directly testing the association of an untyped SNP using a multimarker tag.
Previously, other investigators have suggested testing a specific tag haplotype, or performing a weighted analysis using weights derived from the training data. However these approaches do not properly account for the imperfect correlation between the tag haplotype and the untyped SNP. Here we describe a straightforward approach to testing untyped SNPs using a missing-data likelihood analysis, including the tag markers as nuisance parameters. The training data is stacked on top of the main body of genotype data so there is information on how the tag markers predict the genotype of the untyped SNP. The uncertainty in this prediction is automatically taken into account in the likelihood analysis.
We compare our approach with testing specific tag haplotypes and separately with WHAP, a method described recently by Zaitlen et al. We show that our approach yields more power than single haplotype imputation and similar power to WHAP, yet it has the advantages that it takes into account training set phenotypes and we may obtain an estimate of the odds ratio.

## 83
## Genome-wide evidence for linkage of a new region (11p14) to Atopic Dermatitis and Allergic Diseases

M. Guilloud-Bataille(1), E. Bouzigon(2), I. Annesi-Maesano(3), F. Kauffmann(4), M. Lathrop(5), F. Demenais(2), M.-H. Dizier(1) on behalf of the EGEA cooperative group.
(1) INSERM U535, (2) INSERM U794, (3) INSERM U707, (4) INSERM U780, (5) CNG, France.

Atopic dermatitis (AD), asthma and allergic rhinitis (AR) are allergic co-morbidites which are likely to depend on both common and specific genetic factors. After a previous genome-wide linkage screen conducted for asthma and AR in a sample of 295 French EGEA families with at least one asthmatic subject, our aim was to search for genetic factors involved in AD as well as those ones shared by the three allergic diseases using the same EGEA data. AD and the phenotype of 'allergic disease' defined by the presence of at least one of the 3 diseases (asthma, AR, AD) were examined by linkage analyses using the Maximum Likelihood Binomial (MLB) method. A fine mapping was carried out in regions detected for potential linkage, followed by association studies using the Family Based Association Test (FBAT). Evidence for linkage to 11p14 region was shown for both AD and 'allergic disease' (genome-wide p value $< 0.05$). Linkage was also indicated between AD and 5q13 and between 'allergic disease' and both 5p13 and 17q21 regions. Fine mapping supported the evidence of linkage to 11p14 and FBAT analyses showed association between 'allergic disease' and a marker located at the linkage peak on 11p14. Further linkage disequilibirum mapping in this region will allow identification of genes involved in AD and/or in the 'allergic disease' phenotype.

## 84
## Estimating significance threshold for genomewide association scans

A. Gusnanto, F. Dudbridge
MRC Biostatistics Unit, Cambridge, United Kingdom

The question of what significance level is appropriate for genomewide association studies is somewhat unresolved. Permutation testing is advocated, but does not resolve the difference between the genomewide multiplicity of the experiment, and the subset of markers actually tested. Bayesian arguments, however, can struggle to distinguish direct associations from those due to LD. A standard significance level would facilitate reporting of results and reduce the need for permutation tests.We used genotypes from the Wellcome Trust Case-Control Consortium to estimate a genomewide significance level. We sub-sampled the genotypes at increasing densities, using permutation to estimate the nomimal p-value for 5% family-wise error. By extrapolating to infinite density, we estimated the genomewide significance level to be about 10-7. We compared this to two estimators of the effective number of tests. The first fits a beta distribution to permutation replicates, and the second is based on an eigenvector decomposition of the genotype data. The beta distribution is not exact, but we found that it provides a workable approximation for calculating genomewide significance. Patterson's eigenvalue estimator requires less computation but was found to be an order of magnitude too low, leading to increased type-1 errors. We conclude that permutation is still needed to obtain genomewide significance levels, but with sub-sampling, extrapolation and estimation of an effective number of tests, the significance level can be standardized for all studies of the same population.

## 85
## MDR-Bagging: The improvement of Multifactor Dimensionality Reduction using Bagging Predictors with Out-Of-Bag Estimation

Min-Jin Ha, Eun-Kyung Lee and Taesung Park
Department of Statistics, Seoul National University

The multifactor dimensionality reduction (MDR) is a nonparametric method that can identify gene-gene interactions in case-control studies. MDR uses consistency, prediction error, and balanced accuracy to select the best combination of genetic factors through cross validation. However, MDR tends to perform poorly in many real applications especially when the data are sparse and unbalanced between cases and controls. We propose an improvement, MDR with Bagging predictors (MDR-Bagging). MDR-Bagging makes MDR procedure more stable by using

aggregated predictors from bootstrap samples. We expect that MDR-Bagging has a potential to detect interactions with high accuracy even in small-size samples.

**86**

**Power comparison of model-free linkage methods using covariates**

M.L. Hamshere, P.A. Holmans
Biostatistics and Bioinformatics Unit, Cardiff University School of Medicine, UK

The power of model-free analyses to detect linkage to complex traits can be greatly improved by the inclusion of covariates, by reducing phenotypic (and thereby genetic) heterogeneity as well as allowing gene-environment interactions to be modelled. Tsai & Weeks (2006, Genet Epidemiol 30:77–93) compared the power of several analysis methods, including mixture models, logistic regression methods and ordered subsets analysis. In their study, trait susceptibility was modelled by a threshold model with underlying liability determined by various types of interaction between genotype and a quantitative environmental risk factor. The mean value of the risk factor among affected individuals was used as the covariate in the linkage analyses.
We extend the comparison to other models of disease susceptibility, such as pleiotropy (one locus influencing both trait risk and quantitative covariate value), modifier loci (where the locus influences covariate value but not trait risk) and heterogeneity (where trait susceptibility is determined by posessing risk alleles at either of two loci, with these loci manifesting different covariate values). We also investigate the use of different measures of the individuals' covariates in the analyses (e.g. trait differences).
We find that the optimal choice of covariate measure depends on the disease model, and this in turn influences the relative power of the analysis methods.

**87**

**Influence of linkage disequilibrium and marker allele frequency on the sample size needed to detect gene-environment interaction in indirect association mapping**

R. Hein, L. Beckmann, J. Chang-Claude
Department of Cancer Epidemiology, German Cancer Research Center DKFZ, Heidelberg, Germany

Association studies accounting for gene-environment interactions (GxE) may be useful for detecting genetic effects and identifying important environmental effect modifiers. Current technology facilitates very dense marker spacing in genetic association studies. In this case, the true disease variants may not be genotyped, so that causal genes are searched for by indirect association using genetic polymorphisms associated with the true disease variants. Zondervan and Cardon [Nat Rev Genet, 2004, 5:89–100] showed that sample sizes needed to detect markers which are associated with disease variants highly depend on the effect size at the true disease loci, the linkage disequilibrium (LD) between variants and markers, and whether allele frequencies match. We examined sample sizes needed to detect GxE in indirect association studies and provide an algorithm for power and sample size estimations. Besides LD, allele frequencies, marginal effects of true disease loci, sample sizes depend on marginal effect of environmental exposures, prevalence of environmental exposures, and magnitude of interactions. For discordant allele frequencies and incomplete LD, sample sizes can be unfeasibly large. Extent of LD and discordant allele frequencies have equally strong influence on sample size requirements when considering r2 instead of D'. The impact of both factors increases for disease loci with lower allele frequencies. Given small genetic marginal effects, large interaction effects can be detected using smaller sample sizes than those needed for the detection of main effects, depending on LD and marker allele frequencies.
A software program that implements the method will be made freely available.

**88**

**Inference from Genome-Wide Association Studies using a Novel Markov Model**

F.J. Hosking(1), J.A.C. Sterne(2), G. Davey Smith(2), P.J. Green(1)
(1) Department of Mathematics, University of Bristol, UK, (2) Department of Social Medicine, University of Bristol, UK

We propose a Bayesian modelling approach to the analysis of genome-wide association studies (GWAS) based on single nucleotide polymorphism (SNP) data. Our model combines various aspects of k-means clustering, hidden Markov models (HMMs) and logistic regression into a fully Bayesian model. It is fitted using the Markov chain Monte Carlo (MCMC) stochastic simulation method, with Metropolis-Hastings update steps. The approach is flexible, both in allowing different types of genetic models, and because it can be easily extended while remaining computationally feasible due to the use of fast algorithms for HMMs. It allows for inference primarily on the location of the causal locus but also on other parameters of interest. The model is used here to analyse three data sets, using both synthetic and real disease phenotypes with real SNP data, and shows promising results.

**89**

**Selection of informative SNPs for sibling pair linkage analysis from large SNP arrays**

J.J. Houwing-Duistermaat, Q. Helmer, B.T. Heijmans, M. Beekman, P.E. Slagboom.
Dept of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

With the availability of large dense single nucleotide polymorphism (SNP) arrays, it is now possible to select a study specific subset of SNPs for linkage analysis. For example when one of the affected sibling pair is part of a genome wide association study, the other sibling may be typed for a part of the SNPs to be able to perform sibling pair linkage analysis. We developed a selection procedure to obtain the most informative set of a certain size in the case of no parental genotypes. The marker information, i.e. the

information on identical by descent status of sibling pairs depends on both the marker spacing as well as on the allele frequency. This makes it hard to derive a formula for the best marker set. We propose to use Merlin to compute the marker information and drop the least informative marker in the region with highest marker information. We repeat dropping markers until the desired number of markers is reached.
We applied this procedure to selected 5K SNPs from 300K SNPs in the Leiden Longevity Study. The allele frequencies of the 300K SNPs were available for our population and we selected SNPs with minor allele frequency above 0.2. We mapped these SNPs on the integrated genetic map constructed by David Duffy (http://www2.qimr.edu.au/davidD/) by interpolation. Then we applied our method. After genotyping the selected SNPs appeared to have an average information of 60%, which was also expected. The method outperforms selection just based on spacing.

**90**

**Age Specific Effect of TGFβ1 Variants on Pulmonary Function in Cystic Fibrosis Patients**

B. Huang (1,2), C. Taylor (1,2), R. Dorfman (1), A. Sandford (3), P. Parè (3), K. Shumansky (3), J. Zielenski (1), P. Durie (1,2), M. Corey(1,2).
(1) The Hospital for Sick Children, Toronto, Canada
(2) University of Toronto, Canada
(3) University of British Columbia, Canada

Transforming growth factor beta1 (TGFβ1) has been reported as one of the major genetic modifiers of cystic fibrosis (CF), with varying significance. In our Canadian CF Modifier Gene Study we ascertained a cohort of n=1583 pancreatic insufficient CF patients to study the effects of TGFβ1. The L10P polymorphism in TGFβ1 was selected as a surrogate of the 4 SNPs genotyped within the TGFβ1 gene. Analyses of pulmonary function in different age groups indicated that the C allele has a significant dominant protective effect only in the subgroup of CF patients aging 15–30 (p=0.008). The effect is not significant in younger or older patients. The age groups were defined based on the plot of local average FEV1 score against age by different genotype groups. In a previous study, we did not see any effect of TGFβ1 on the time-to-onset of *Pseudomonas aeruginosa* infection. However, in this study, the CC polymorphism of codon10 was significantly associated with a higher proportion of sporadic than chronic infection in adult CF patients aging 18–36 (p=0.013), suggesting that CC may delay the transition to chronic infection. We conclude that the effect of TGFβ1 and other CF modifier genes will depend on age, reflecting the interaction of infection and inflammatory processes involved in the long progression of CF lung disease.

**91**

**The contribution of genetic and other factors to von Willebrand factor (VWF) levels in the blood**

L.O. Hudson, M.D. Teare, A. Goodeve,
D. Hampshire
Division of Genomic Medicine, School of Medicine and Biomedical Sciences, University of Sheffield, UK

von Willebrand disease (VWD) is a common inherited bleeding disorder, with three distinct types, affecting both males and females. Type 1 VWD results from a partial quantitative deficiency in von Willebrand factor (VWF) plasma levels, and can be characterised by a personal and family history of mucocutaneous bleeding. Type 1 is also the most common, occurring in approximately 50–75% of patients, however type 1 VWD can be difficult to diagnose, with various factors known to affect VWF plasma levels including age and stress. There is also evidence to indicate an association between type 1 VWD and ABO blood group. Despite this, there is very little known about how much these various factors affect VWF plasma levels or whether any other factors also play a role. Recently, a study conducted within the EU attempted to determine the molecular basis of type 1 VWD, investigating a large cohort of patients diagnosed with the disease, plus their affected and unaffected family members (the 'Molecular and clinical markers for the diagnosis and management of type 1 von Willebrand disease' (MCMDM-1VWD) study). Extensive phenotypic and genotypic data has been collected on these families and also on a series of healthy controls.
Using the range of information available in the EU study, we examine the effect that different factors have on VWF plasma levels and attempt to identify the extent to which variation in VWF plasma levels is due to measured and unobserved genetic and environmental effects.

**92**

**Testing association in the presence of linkage - a powerful score for binary traits**

K. Humphreys (1), G. Jonasdottir (2), J. Palmgren (1,2)
(1) Dept. of Medical Epi. and Biostatistics, Karolinska Institute, Stockholm, Sweden (2) Dept. of Mathematics, Stockholm University, Sweden.

We describe a score for testing association in the presence of linkage for binary traits. The score is robust to varying degrees of linkage, and it is valid under any ascertainment scheme based on trait values as well as under population stratification. The score test is derived from a mixed effects model where population level association is modeled using a fixed effect and where correlation among related individuals is allowed for using log-gamma random effects. The score does not assume full information about the inheritance pattern in families or parental genotypes. We compare the score to the semi-parametric family-based association test (FBAT), which has won ground because of its flexible and simple form. We show that a random effects formulation of the co-inheritance can improve the power substantially.

**93**

**Genomewide scan of African American and caucasian populations for linkage to myopia**

G. Ibay(1), R. Wojciechowski(1), E. Ciner(2), D. Stambolian(3), J.E. Bailey-Wilson(1)
(1)NHGRI/NIH, Balto, MD, (2)Pennsylvania College of Optometry, Phil., PA, (3)Dept. of Ophthalmology, Univ. of Pennsylvania, Phil., PA.

Worldwide, myopia and associated refractive errors are the leading cause of visual impairment and are a disease focus of VISION 2020 whose goal is to eliminate global blindness. Our goal is to find evidence for a myopia susceptibility gene causing common myopia (at least −1.00 diopter or lower in each meridian of both eyes). Cycloplegic and manifest refraction were measured on 94 African American and 36 Caucasian families, each with a myopic proband and at least one other affected member. A genomewide linkage scan using 387 markers was performed on 393 African Americans and 184 Caucasians. Multipoint nonparametric linkage (NPL) in the African American families identified two loci with nominal evidence for linkage: on chromosome 6 (D6S1035, 164.8 cM, NPL=2.59, p-value=0.005) and chromosome 7 (D7S817, 50.3 cM, NPL=2.57, p-value=0.005). The Caucasian families showed some evidence of linkage on chromosome 20 (D20S481, 62.3 cM, NPL=2.52, p-value= .008). Multipoint NPL scores from both genome screens were combined by calculating a combined NPL score while still using ethnically appropriate allele frequencies in each population. The best evidence for linkage combining the two populations was in the 6.6-cM interval around D20S478 with a combined NPL of 2.48, and a p-value=0.008. Genome-wide significant linkage of refractive error to the same region of chromosome 7 supports our results in the African American (see Wojciechowski et al). Further study of these regions is ongoing.

**94**

**Using haplotype-dropping to quantify and correct bias in tagging SNPs caused by insufficient sample size and marker density**

M.M. Iles
Leeds Institute of Molecular Medicine, University of Leeds, UK

Tagging SNPs (tSNPs) are commonly used to capture genetic diversity cost-effectively. However, it is important that the efficacy of tSNPs is correctly estimated; otherwise coverage may be insufficient. If the pilot sample from which tSNPs are chosen is too small or the initial marker map too sparse, tSNP efficacy may be overestimated.

An existing estimation method using bootstrapping goes some way to correct for insufficient sample size and overfitting, but does not completely solve the problem. We describe a novel method, based on exclusion of haplotypes, that improves on the bootstrap approach. Using simulated data, the extent of the sample size problem is investigated and the performances of the bootstrap and novel method compared. We incorporate an existing method adjusting for marker density by 'SNP-dropping'.

We find that insufficient sample size can cause large overestimates in tSNP efficacy, even with as many as 100 individuals, and the problem worsens as the region studied increases in size. Both the bootstrap and novel method correct much of this overestimate, with our novel method consistently outperforming the bootstrap method.

We conclude that a combination of insufficient sample size and overfitting may lead to overestimation of tSNP efficacy and underpowering of studies based on tSNPs. Our novel approach corrects for much of this bias and is superior to the previous method. Sample sizes larger than previously suggested may still be required for accurate estimation of tSNP efficacy. This has obvious ramifications for the selection of tSNPs from HapMap data.

**95**

**A method for selecting a minimum number SNPs for linkage analysis from very high (500+K) SNP sets, while maintaining sufficient informativeness**

J.M. Farnham, N.J. Camp, A. Thomas, L.A. Cannon-Albright
Department of Biomedical Informatics, University of Utah

Multipoint linkage analysis requires genetic markers in linkage equilibrium (LE). With the advent of very high resolution SNP genotyping (500+K markers), selecting an appropriate subset of SNPs becomes an important consideration. Removing linkage disequilibrium (LD) is essential; however, with more than 20,000 SNPs per chromosome, even when markers in LD have been eliminated, thousands of markers remain. While it is important that sufficient markers are retained to reconstruct segregating haplotypes successfully, the density of a LE SNP map may be beyond the point where marginal additional informativeness is outweighed by the cost of analysis resources. Once markers in LD have been removed, then selection of a set of SNPs dense enough to maintain high informativeness is needed. We present a method for the second step. For each pedigree, the expected length of non-recombinant sharing among the affecteds and the number of chromosomes entering the pedigree are calculated. This information is used to determine how many LE SNPs are required per non-recombinant region to produce haplotypes with sufficient heterozygosity such that there is high probability that each haplotype entering is unique (ie. high informativeness). Thus, the required number of SNPs per cM is estimated and a decision on the required resolution can be made. The method was applied to a set of extended pedigrees which were genotyped with 550K SNPs. The resolution suggested by the method is compared to the 'information content' provided by Merlin.

**96**

**Application of Bayesian Graphical Model to Study the Genetic Susceptibility to Breast Cancer**

S. Kang (1), H. Ozcelik(1), H. Jarjanazi(1), H. Massam(2), and L. Briollais(1)
(1) Samuel Lunenfeld Research. Inst., Mt. Sinai Hospital, Canada
(2) Dept. of Math. & Stat., York Univ., Canada

Graphical models have been recently applied in association studies to model complex patterns of linkage disequilibrium either in the context of candidate gene studies (Thomas 2004) or genome-wide association studies (Verzilli 2006). These

two approaches used prior information based on the physical distance between the genetic markers. In this study, we propose a new Bayesian graphical modeling framework where prior information is specified in terms on functional distance between genetic markers. We applied this new methodology to a population-based study of breast cancer including 398 cases selected from the Ontario Familial Breast Cancer Registry (OFBCR) and 372 age-matched controls. We studied the association between 17 SNPs, selected from 15 genes in the DNA repair pathway, and breast cancer. Our prior information is based on protein-protein interaction inferred from BIND and other databases as well as the predicted functionality of the 17 SNPs on the protein activity. We first present the theoretical foundation of our graphical model with Bayesian and non-Bayesian perspectives and then discuss different settings of priors (uninformative and informative), the effect of hyper-parameters on priors and the use of different fitting algorithms (MCMC or RJMCMC). We finally show how this methodological approach can help to reduce false positive results, improve the detection of real association and lead to a better modeling of complex gene-gene interactions in breast cancer susceptibility.

**97**

**Detection of gene-environment interaction: a new test based on sibling recurrence risks**

R. Kazma(1,2), C. Bonaïti-Pellié(2,1), J.M. Norris(3), E. Génin(2,1)
(1) Univ Paris-Sud, UMR-S535, Villejuif, France, (2) INSERM, UMR-S 535, Villejuif, France, (3) Dept. of Preventive Medicine and Biometrics, University of Colorado at Denver and Health Sciences Center, USA.

Gene-environment interactions (GEI) may play important roles in complex disease susceptibility but their detection is often difficult. In order to investigate GEI, we propose a method based on the degree of familial aggregation according to the exposure of the index. In case of GEI, the distribution of genotypes of affected individuals, and consequently the risk in relatives, depends on their exposure. We developed a test comparing the risks in sibs according to the index exposure. To evaluate the power of this new test for various sample sizes of affected individuals, we derived the formulas for calculating the expected risks in sibs according to the exposure of indexes for various values of exposure frequency, relative risk due to exposure alone, frequencies of genotypes, genetic relative risks and interaction coefficients. We conclude that this test is valuable for diseases with moderate familial aggregation, only when the role of the exposure has been clearly evidenced.
Since a correlation for exposure among sibs might lead to a difference in risks in the different index exposure strata, we also added an exposure correlation coefficient in the model. Interestingly, we find that when this correlation is correctly accounted for, the power of the test is not decreased and might even be significantly increased. We used this method on type 2 diabetes data and compared it to other methods to detect GEI.

**98**

**Whole genome association studies of rheumatoid arthritis and replication of identified susceptibility loci**

X Ke(1), W Thomson(1), A Barton(1), S Eyre(1), A Hinks(1), J Bowes(1), R Donn(1), S Hider(1), I.N. Bruce(1), A.G. Wilson(2), A Morgan(3), P Emery(3), YEAR consortium, A Carter(4), S Steer(5), L Hocking(6), D.M. Reid(6), D Strachan(7), P Wordsworth(8), J Worthington(1)
(1)ARC-EU, University of Manchester, UK, (2)School of Medicine & Biomedical Sciences, University of Sheffield, (3)Academic Unit of Musculoskeletal Disease, Chapel Allerton Hospital, Leeds, (4)Academic Unit of Molecular Vascular Medicine, University of Leeds, (5)Clinical and Academic Rheumatology, Kings College Hospital, London, (6)Department of Medicine & Therapeutics, University of Aberdeen, (7)Division of Community Health Sciences, St George's, University of London, (8)Nuffield Department of Orthopaedic Surgery Nuffield Orthopaedic Centre, Oxford, UK.

ARC-EU is part of the Wellcome Trust Case Control Consortium (WTCCC), which has conducted a WGA scan on 7 common diseases including rheumatoid arthritis (RA). The study identified more than 50 SNPs at $p < 1x10-4$ and 9 SNPs at p=5x10−5 − 1x10−7, associated with RA, excluding known variants from HLA and PTPN22 regions. In the replication cohort of 4,373 RA cases and 2,365 controls, one of the SNPs (OR=1.25, 95% CI 1.13–1.37, trend $p < 5x10-6$) was found to be significantly associated with RA (OR=1.26, 95% CI 1.16-1.38, trend $p < 1.3x10-7$). Stratified analysis revealed patients positive for rheumatoid factor and anti-CCP had a much stronger association with this SNP. This SNP was however not linked to any obvious candidate gene. Several other SNPs were also replicated with borderline significance ($p < 0.05$) and further replications are underway.

**99**

**IRF5 and Lupus Risk in Multiple Races**

JA Kelly(1), JC Edberg(2), KM Kaufman(1), J Merrill(1), JA James(1), MC Marion(3), CD Langefeld(3), MA Petri(4), JD Reveille,(5) R Ramsey-Goldman(6), LM Vilá(7), GS Alarcón(2), RP Kimberly(2), JB Harley(1)
(1)OMRF, OK, (2)U of Alabama at Birmingham, AL, (3)Wake Forest U, NC, (4)Johns Hopkins, MD, (5)U of Texas HSC, TX, (6)Northwestern U, IL, USA; (7)U of Puerto Rico, PR

We evaluated interferon regulatory factor 5 (IRF5) SNPs in a large collection of lupus cases and controls from minority (African-Americans (AA), Hispanics (HI), and Puerto Ricans (HI-PR)) and European-American (EA) populations. A total of 6020 samples (2683 cases and 3337 controls) from two independent cohorts were evaluated. Case-control association tests were obtained using Pearson chi-square statistics and conditional haplotype analyses were conducted using WHAP. Significant associations with rs2004640 and rs3807306 were observed in the AA and HI cohorts. Both loci also demonstrated strong association with lupus in the EA cohort. Suggestive association between rs3807306 and lupus in the HI-PR cohort was also observed. We identified a 5-marker

risk haplotype in the AA cohort, with rs3807306 accounting for the majority of the observed statistical effect. This outcome was also observed in the HI samples. The risk haplotype was again observed in the EA cohort, and though its effect was not significantly different than a three-marker haplotype previously reported, haplotypes containing the common A risk allele at rs3807306 were predictive of lupus risk. In summary, we establish association with IRF5 in AA and HI lupus patients, providing evidence that IR5 is likely to be a crucial component in lupus pathogenesis in multiple ethnic groups.

**100**

**Type I error rate when using high-density SNP panels for nonparametric multipoint linkage analysis of two-generation and multigenerational pedigrees**

Y. Kim (1,2), P. Duggal (2), E.M. Gillanders (2), H. Kim (1), J.E. Bailey-Wilson (2)
(1) Dept. of Epi & Biostat, Seoul National Univ., Seoul, Republic of Korea
(2) Inherited Disease Research Branch, NHGRI/ NIH, MD, U.S.A

High-density single nucleotide polymorphism (SNP) panels have become major resources for linkage analyses. There is often linkage disequilibrium (LD) between these markers; however, most analysis programs currently assume linkage equilibrium when inferring parental haplotypes. Failure of this assumption can result in inflation of Type I error rates, which is worse when parental genotypes are unavailable. We investigated the effect of LD on the Type I error rate of nonparametric multipoint linkage analysis of two-generation and multigeneration multiplex families in the presence of missing genotype data. Using genome wide SNP data from the Collaborative Study of the Genetics of Alcoholism, we modified the original dataset into 30 data sets: 6 patterns of missing data for 5 levels of SNP density. To assess the Type I error rate, we simulated 1,000 qualitative traits from random distributions, unlinked to the marker data. Large increases in Type I error rates were not observed for most of the missing data patterns if the SNP markers were more than 0.3 cM apart. However, in a dense 0.25 cM map, removing genotypes on founders and/or founders and parents in the middle generation caused substantial inflation of the Type I error rate. Substantial intermarker LD existed in the 0.25cM map whereas the 0.3cM and less dense maps exhibited very little intermarker LD.

**101**

**Deep resequencing identifies an APOE hepatic control region variant that accounts for ApoE linkage beyond the ε2/3/4 polymorphism**

K.L.E. Klos, L.C. Shimmin, J.E. Hixson, E. Boerwinkle
University of Texas Health Science Center, Houston, TX

Resequencing a 102Kb region around the apolipoprotein E (APOE) gene in 20 individuals each from three ethnic groups identified regulatory region variations subsequently associated with plasma lipids in large population-based samples. To evaluate their ability to account for a plasma ApoE QTL on chr. 19q in the GENOA study, three SNPs and the APOE ε2/3/4 polymorphism were genotyped. Linear adjustment of ApoE levels for APOE genotypes were used to remove evidence of linkage due to measured variation. Separately, APOE polymorphisms were added to the markers on chr. 19, and the proportion of variance attributed to IBD sharing was estimated in order to increase power to detect APOE region effects. Linkage analysis of apoE adjusted for age, BMI and gender was performed for both methods. As previously reported, the chr. 19 QTL is partially accounted for by APOE ε2/3/4. In African-Americans, inclusion of both ε2/3/4 and a far-downstream hepatic control region SNP increased the LOD score from 3.99 to 7.81. Adjusting for these polymorphisms reduced the LOD score to 1.40. Similarly, in European-Americans LOD scores increased from 1.79 to 2.64 after inclusion of APOE region markers and decreasing to 0.76 with linear adjustment. In Mexican-Americans, where evidence of linkage was least (LOD=0.88), addition of markers and linear adjustment both resulted in slightly lower LODs. SNPs discovered through deep resequencing and fine-scale association may be applicable to explaining linkage results in pedigree-based samples.

**102**

**CLUMPHAP: A simple tool for performing haplotype based association analysis**

J Knight (1), D Curtis (2), PC Sham (1,3)
(1)Social Genetic and Developmental Psychiatry, Institute of Psychiatry, Kings College London. (2)Academic Centre for Psychiatry, Royal London Hospital. (3)Department of Psychiatry & Genome Research Centre, Hong Kong University.

Large numbers of single nucleotide polymorphisms (SNPS) have been characterised and it is believed this progress will increase our chances of identifying the genes that influence complex traits. In principle, association analysis of haplotypes rather than single SNPs may better capture an underlying causal variant, but the large number of haplotypes can lead to reduced power because of the need to adjust for multiple testing. This paper presents a novel method based on clustering similar haplotypes to address this issue. Using simulation studies we compare its power to identify untyped susceptibility locus with the power of more widely used approaches. We demonstrate its advantage over the omnibus haplotype test and its comparability with multiple regression locus-coding approaches. We have implemented this technique in a program called CLUMPHAP. This represents an extension of the basic methodology used in CLUMP, a program designed to analyse multi-allelic markers (1). CLUMPHAP groups haplotypes into two groups (based on a distance matrix) and determines which grouping yields the strongest evidence for association. Overall significance is determined using permutation testing. The results are easy to interpret, a significant result suggests that a disease causing variant is present on haplotypes in the group which has an increased overall frequency in cases.

1. Sham & Curtis (1995) Ann Hum Genet 59:97–105

**103**

**Genotype relative risk estimation using logistic regression methods in family based data**

S. Kotti (1,2), F. Clerget-Darpoux (1,2), H. Bickeböller (3)
(1) Univ Paris-Sud, UMR-S535, Villejuif, France
(2) Inserm U535, Villejuif, France
(3) Department of Genetic Epidemiology, Medical School, Georg-August-University of Göttingen, Göttingen, Germany

Logistic regression methods using internal controls have become increasingly popular for detecting the effect of a genetic factor and for estimating the genotype relative risks (GRRs). Different internal controls may be used. One pseudocontrol formed by the non-transmitted alleles from parents to the affected offspring is often compared in a 1:1 matching to the genotype of the case. However, three pseudocontrols formed by the parental alleles except the genotype of the case can be used in a 1:3 matching. We compare the GRR estimations under the conditional logistic regression (CLR) and the unconditional logistic regression (ULR) with 1 and 3 pseudocontrols, respectively. The main results are that the GRR estimators are unbiased when using the CLR approach, with smaller variances using 3 pseudocontrols. When samples are analyzed without conditioning on parents, the GRR estimators are strongly biased using the ULR approach with 3 pseudocontrols. In addition, the GRRs estimates using ULR with 1 pseudocontrol is more efficient than those using the CLR with 3 pseudocontrols.
In principle, family based procedures have been developed to adjust for population stratification by conditioning on the parents. However, there is now a recent interest in combining family based and case-control samples. In this context, we show that the use of the ULR approach with one pseudocontrol is the only appropriate one.

**104**

**Application of Bayesian Classification with Singular Value Decomposition Method in Genome-Wide Association Studies**

Soonil Kwon, Xiuqing Guo
Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA

Genome-wide association studies usually involve several hundred thousand of single nucleotide polymorphisms (SNPs). Conventional statistical approaches face challenges when dealing with such an enormous number of SNPs (p) with a relatively small sample size (n). Bayesian classification methods have been introduced for studying binary disease status. By utilizing singular value decomposition (SVD), we extended the Bayesian classification method in both probit and logit models to multinomial ordinal responses data here. For various types of prior information (e.g., vague priors, convenient priors), we showed how SVD influences posterior probabilities. We developed a Markov Chain Monte Carlo (MCMC) based computation algorithm to realize the Bayesian classification with SVD method (BCSVD). To evaluate our method, we simulated 20 samples and 1000 SNPs, with 9 out of the 1000 SNPs (every 100th SNP, except the last one) contribute to disease status. We assumed that there are 3 disease development stages and an additive genetic model for each of the 9 disease associated SNPs. Using the BCSVD method and MCMC algorithm, we successfully identified 8 out of the 9 associated SNPs, while the other one also showed some small effect. The final model with these 9 SNPs can classify people into the correct disease status category with >80% probability. Our results demonstrated that the Bayesian classification with SVD method can be reliably used to analyze genome wide association data when p>>n for multinomial ordinal responses.

**105**

**Polymorphisms in the One-Carbon Metabolic Pathway, and the Risk and Survival of Colorectal Cancer**

C Kyte, J Chen, M Valcin, W Chan, JG Wetmur, J Selhub, DJ Hunter, J Ma
Departments of Community and Preventive Medicine [J. C., C. K., W. C.] and Microbiology [J. G. W.], Mount Sinai School of Medicine, New York, New York 10029; Channing Laboratory, Department of Medicine, Brigham and Women's Hospital [D. J. H., M. J. S., J. M.] and Harvard Medical School, Departments of Epidemiology [D. J. H., M. J. S.] and Nutrition [M. J. S.], Harvard School of Public Health, Boston, Massachusetts; and Jean Mayer United States Department of Agriculture Human Nutrition Center on Aging at Tufts University, Boston, Massachusetts [J. S.]

Polymorphisms in one-carbon metabolizing genes cytosolic serine hydroxymethyltransferase (cSHMT), methylenetetrahydrofolate dehydrogenase (MTHFD1), glutamate carboxypeptidase II (GCPII), and thymidylate synthase (TS), have been identified. There is evidence implicating their functionality in modifying folate status and related CRC risk. Molecular and genetic epidemiologic approaches were used to investigate (1) whether dietary folate is a micronutrient that is protective against colorectal cancer; (2) whether the sub-population carrying genetic polymorphisms in folate-metabolizing genes is at increased risk of colorectal cancer; (3) how folate may interact with these polymorphisms to modify risk of colorectal cancer, and (4) how folate may interact with folate antagonists in contributing to the risk of colorectal cancer. The significance of this research lies not only in its potential to clarify etiology of colorectal cancer but also to guide the prevention of colorectal cancer through dietary modification.

**106**

**Trans- and long-range cis- associated SNPs - potential inference errors for genome-wide association studies**

R.W. Lawrence, L.R. Cardon, E. Zeggini.

Wellcome Trust Centre for Human Genetics, Oxford, UK.

Recent advances in high-throughput genotyping and a better understanding of human genome sequence variation have now made genome-wide association scans possible. However, exhaustive screening of common variation is not yet feasible. Therefore, inferences about the localisation of disease variants have to be made on the basis of genome-wide association scan results. Incomplete surveys of the local linkage disequilibrium architecture could conceivably lead to misinterpretation of findings. We have calculated linkage disequilibrium between every SNP pair (with MAF > 5%) from all three HapMap phase II samples (CEU, YRI, and JPT/CHB combined). We observe that a number of SNPs have at least one strongly associated (r2 > 0.7) marker on a different chromosome or at a distance greater than 1Mb on the same chromosome. 1.4% and 1.3% of common SNPs were strongly correlated (r2 > 0.7) with another variant at a distance greater than 500kb in CEU and YRI samples respectively. Although relatively rare (11,770 out of 2 million SNPs from the CEU sample), trans-chromosomal associations could lead to inference errors in the downstream interpretation of genome-wide association study results. We are developing a resource enabling researchers to quickly retrieve information on these long-range associations for any given common HapMap SNP. This will conceivably help gene-hunters localise strong signals emerging from disease association studies, plan targeted replication strategies and delineate appropriate intervals for fine-mapping and resequencing.

## 107
## Ignoring intermarker linkage disequilibrium induces false-positive evidence of linkage for consanguineous pedigrees when genotype data is missing for any pedigree member

S.M. Leal, B. Li
Dept. of Molecular and Human Genetics, Baylor College of Medicine, USA

Missing genotype data can increase false-positive evidence for linkage when either parametric or nonparametric analysis is carried out ignoring intermarker linkage disequilibrium (LD). Previously it was demonstrated by Huang et al (2005) that no bias occurs in this situation for affected sib-pairs with unrelated parents when either both parents are genotyped or genotype data is available for two additional unaffected siblings when parental genotypes are missing. However, this is not the case for consanguineous pedigrees, where missing genotype data for any pedigree member within a consanguinity loop can increase false-positive evidence of linkage. The amount of false-positive evidence for linkage and which family members aid in the reduction of false-positive evidence of linkage is highly dependent on which family members are genotyped. When parental genotype data is available, the false-positive evidence for linkage is usually not as strong as when parental genotype data is unavailable. For a pedigree with an affected proband whose first-cousin parents have been genotyped, further reduction in the false-positive evidence of linkage can be obtained by including genotype data from additional affected siblings of the proband or genotype data from the proband's sibling-grandparents. When parental genotypes are unavailable, false-positive evidence for linkage can be reduced by including in the analysis genotype data from either unaffected siblings of the proband or the proband's married-in-grandparents.

## 108
## Integrating Pathway and Linkage Information towards Candidate Gene Prioritization

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

Pinpointing susceptibility genes among the myriad of genes surrounding loci identified by linkage analysis is a major challenge. Complex genetic disorders are thought to be the result of many weakly (inter)acting genes. Nonetheless, for most disorders, only a few molecular pathways will influence the disease. Each of the few thousands human genes may be involved or not into those few pathways and this constantly updated information can now easily be accessed from publicly available databases. The fact that several genes residing at different loci are involved in the same pathway has implications as to the behavior of the genome-wide linkage curve. We propose to formally incorporate this extra source of gene-pathway information into the process leading to candidate gene prioritization. Prior to the linkage study and in absence of any biological information, all genes have an equal chance of being a susceptibility gene. Following a linkage study for the disease of interest, the probability for each gene to influence the disease is updated in the light of the linkage signal and its consistency with the gene-pathway information available. Technically, we use a convenient Gaussian model to describe the gene effects and the linkage curve. This allows simplified computations whose end-result is a probability for each gene to be linked to the disorder. We have applied our methodology to data in rheumatoid arthritis using 47 relevant pathways.

## 109
## Risk estimation for common complex genetic disorders: using genetic, environmental and family history information in Crohn's disease

C.M. Lewis(1), S.C.L. Whitwell(2), A. Forbes(3), J. Sanderson(4), C.G. Mathew(1), T.M. Marteau(2)
(1) Medical and Molecular Genetics, (2) Psychology (3) Gastroenterology, UCL, (4) Gastroenterology, King's College London

Progress has been made in identifying mutations that confer susceptibility to complex diseases, with the potential to use these mutations in determining individual-specific disease risk. We develop methods to estimate disease risk based on (1) genotype relative risks, (2) exposure to an environmental factor and (3) residual family history unaccounted for by the genes modelled. For Crohn's disease (CD), this is applied to variants in the genes CARD15, IL23R and ATG16L1 and smoking. The sibling relative risk ($\lambda_S$) for CD is estimated to be 27. Assuming a multiplicative relationship, the component of this due to CARD15 mutations is only $\lambda_{S,CARD15}$=1.16,

with smaller contributions from IL23R and ATG16L1. We estimate CD risk using CARD15 mutation status (genotype relative risks for heterozygous and homozygous mutation carriers of 2.25 and 9.25 respectively), and smoking (two-fold increased risk, independent of CARD15 mutations). Siblings of a CD case who smoke, and who carry 0, 1, or 2 CARD15 mutations are at approximately 40-fold, 90-fold or 350-fold increased risk of developing CD, respectively. The risk estimates will be used in a trial to test whether CARD15 genotype information increases smoking cessation behaviour in a cohort of smokers with a CD family history.

**110**

**Issues of design and participation in family based studies**

Adam Linke, Penella J Woll, Janet Horsman, M Dawn Teare
University of Sheffield School of Medicine and Biomedical Sciences, Sheffield, UK

Population based family study designs have the potential to facilitate the analysis of the effects of both genes and environment. These types of studies integrate the population based approaches of environmental epidemiology and the methods enabling the analysis of correlations between relatives sharing both genes and environment. Though these resources can be used for gene discovery and genetic association, the real strength of these studies will be in the characterisation of candidate genes and models of gene/gene and gene/environment interaction. The merits of variations of this design have been well explored, however the impact of missing data and the systematic handling of multiple reports on unobserved relatives may place new demands on the statistical analysis. The extent to which these studies are feasible will depend upon the disease under investigation and population or country specific ethics and data protection.

This work will investigate the implications of a) participation rates and, b) choice of control (genetically related / related through marriage/ population based), c) multiple reports of possibly contradictory family history, in family based-designs. It is likely that This will be explored through simulation and from preliminary data from the ReSoLuCENT study. This ongoing North Trent study recruits early-onset lung cancer cases and family based controls.

**111**

**Selection of the most informative individuals from families with multiple siblings for association study**

C Liu, L. A. Cupples, Q. Yang, J. Dupuis
Dept. of Biostatistics, Boston Univ. School of Public Health, USA

An association study often follows an initial linkage analysis for mapping and identifying genes for complex human traits. Many human traits are measured on continuous scales and thus, association analyses may be applied to such traits to understand the genetics underlying these traits that are frequently correlated with disease phenotypes. Follow-up association studies may be conducted on unrelated subsets of individuals where only one member of a family is included. We propose two methods to select one sib per sibship with multiple siblings: 1) one sib with the most extreme trait value; and 2) one sib using a combination score statistic based on extreme trait values and identity-by-descent sharing information. We compare the Type I error and power for these strategies. In addition, we compare these selection strategies with a strategy that randomly selects one sib per sibship, and also with an approach that includes all sibs. We find that using the combination score statistic can increase power by 30 to 40% compared to a random selection strategy. The selection strategy based on both identity-by-descent sharing information and extreme trait values loses only 8 $\sim\sim$ 13% of power compared to an approach that uses all sibs, across all additive models considered, but offers at least 50% genotyping cost saving.

**112**

**Segregation Analyses of 741 Population-Based UK Families Affected by Prostate Cancer**

R.J. MacInnis(1,2), A.C. Antoniou(1), R.A. Eeles(3), D.F. Easton(1)
(1)CR-UK Genetic Epidemiology Unit, University of Cambridge, UK, (2)Centre for MEGA Epidemiology, University of Melbourne, Australia, (3) Cancer Genetics, Institute of Cancer Research and Royal Marsden Hospital, London, UK.

A few studies have proposed genetic models for prostate cancer predisposition, but the results are inconsistent. Some suggest that familial prostate cancer is due to a rare, highly penetrant dominant gene (or genes), while others support a recessive inheritance. We have investigated genetic models of susceptibility to prostate cancer using segregation analysis of occurrence in families ascertained through a population based series of 741 incident cases. We investigated major gene models (dominant, recessive, co-dominant, x-linked), polygenic models, and mixed models of susceptibility using the pedigree analysis software MENDEL. The hypergeometric model was used to approximate polygenic inheritance. We constrained the overall incidence of prostate cancer to agree with the national incidence rates for England and Wales (1960–1993), stratified by age and birth cohort. The best fitting model for the familial aggregation of prostate cancer was the recessive. The frequency of the susceptibility allele in the population was estimated to be 0.15 (95% CI 0.09–0.23). The risk of prostate cancer by age 80 among the rare homozygote carriers born after 1918 was 83%. These analyses will be strengthened using an additional 3000 families from the UK and Australia.

**113**

**Impact of genotyping errors on the type I error and the power of haplotype-based association methods**

V Marquard, L Beckmann, I Heid, J Chang-Claude
(1)Dept. of Cancer Epi, German Cancer Research Center, Heidelberg, Germany (2)GSF, Nat. Res. Center of Environment & Health, Inst. of Epi, Neuherberg, Germany

It is known that genotyping errors (GEs) may have an impact on the type I error and the power of statistical methods in

genetic association analysis. We investigated these values in a case-control study, when differential and non-differential GEs were introduced in realistic scenarios. We simulated 100 data sets, where individual genotypes were drawn from a haplotype distribution of 18 haplotypes with 15 markers in the APM1 gene. For power analysis, marker 13 was assigned as disease locus (DL). GEs were introduced following the unrestricted and the symmetric with 0 edges error models described by Heid et al. (2006). In six scenarios, errors resulted from changes of one allele to another with probabilities of 1, 2.5 or 10%, respectively. Multiple errors per haplotype were possible.
We examined three association methods: Mantel statistics using haplotype-sharing; a haplotype-specific score test; and Armitage trend test for single markers.
Only for high and differential error rates (2.5 and 10%), the type I error of the Mantel statistic was slightly, that of the Armitage trend test moderately increased and that of the score test highly increased. At high GE rates, the percent significant replications of the Armitage trend test and the Mantel statistics was reduced for non-differential errors and inflated for differential errors. Localization of the DL may thus be hampered by highly significant results at neighbouring markers.

**114**

**Heritability Of Quantitative Traits Associated With Type 2 Diabetes In Families From South India**

RA Mathias (1), M Deepa (2), D Raj (2), AF Wilson (1), V Mohan (2)
(1) National Human Genome Research Institute, NIH, Baltimore, USA, (2) Madras Diabetes Research Foundation & Dr. Mohans Diabetes Specialities Centre Gopalapuram, Chennai, India

India is emerging as a major contributor to the global public health burden of diabetes. We have undertaken a family study in Chennai, South India and report on the familial aggregation of quantitative traits associated with Type 2 Diabetes in these families. The 26 families comprise 524 individuals over the age of 19 totaling 2,362 relative pairs. Detailed questionnaires and trait-phenotype data were obtained on all participating individuals including fasting blood glucose, fasting insulin, lipid profiles, height, weight and other anthropometric and clinical measures. Heritability estimates were determined for all quantitative traits at the univariate level, and bivariate analyses were performed across these quantitative traits. Heritability estimates were greater than or equal to 0.21 for all traits (average=0.37). Heritability estimates for traits directly related to the Type 2 Diabetes phenotype, fasting blood glucose and fasting insulin, were $0.24 \pm 0.08$ and $0.41 \pm 0.09$, respectively. Bivariate analyses suggested common genetic and environmental control for several traits, including: fasting insulin and central obesity measures (BMI, waist and hip), with complete genetic correlation between fasting insulin and waist. The evidence for pleiotropic control of insulin and central obesity-related phenotypes supports the presence of an insulin resistance syndrome in South Asians along with a tendency for central obesity.

**115**

**Genome-Wide Association Study Of Asthma-Related Quantitative Traits In Populations Of African Descent**

RA Mathias (1), A Grant (2), AF Wilson (1), T Beaty (2), K Barnes (2)
(1) National Human Genome Research Institute, NIH, USA; (2) Johns Hopkins University, USA

Asthma constitutes a substantial public health burden, yet its underlying etiology remains unclear and few genetic loci have been identified with genome-wide linkage studies. Asthmatics of African descent have more severe asthma, higher IgE levels, higher degree of steroid dependency, and more severe clinical symptoms than white Americans. In this study we have undertaken a collaborative effort across five institutions to perform genome-wide association (GWA) analysis in 2,000 subjects of African descent: 1000 Afro-Caribbean subjects in 171 families from Barbados and 1000 African American subjects (500 cases and 500 controls) from the Baltimore-Washington, D.C. area. The availability of two populations allows for the possibility of the replication of results within a single study. Although the ascertainment schemes for both samples focused on asthma as the primary phenotype, data are available on associated quantitative traits on all (total serum IgE) or a subset (pulmonary function data, specific IgE) of individuals. The Illumina HumanHap650Y BeadChip comprising 655,352 SNPs was used to genotype these populations. This set of genotypes contains all SNPs from the HumanHap550chip plus an additional 100,000 tagSNPs for the Yoruba population. Genotypes are currently available on the case-control data and genotyping is underway on the family data at The Center for Inherited Disease Research. We will present results from the GWA analysis of quantitative traits, with particular attention to total serum IgE.

**116**

**A Logistic Regression Model for Combined Individual- and Family-level Association Analyses of Binary Traits**

L. Mirea(1,2), S.B. Bull(1,2), J.E. Stafford(1), L. Sun(1)
(1) Dept. of Public Health Sciences, Univ. of Toronto, Canada
(2) Samuel Lunenfeld Research Inst., Mt. Sinai Hospital, Toronto, Canada

Subjects available for genetic association studies may include either singletons or trios for population- (PA) or family-based (FBA) analyses, respectively. To analyze a binary trait in both singletons and trios we consider a logistic regression framework and investigate separate and joint models for PA and FBA genotype components. Singleton genotypes contribute only to the PA component. In trios, PA and FBA genotype scores are computed conditional on the parental genotypes as described by Lange et al. (2003 Am J Hum Genet 73: 801–811). Given a fixed number of singletons and trios, we simulated a binary trait under a range of genetic models with and without population stratification (PopStrat) and examined PA, FBA and joint PA+FBA models. The PA and FBA components are orthogonal under a linear but not logistic model. In the simulated data, logistic regression analyses indicate negligible covariance ($<0.02$) between the PA and

FBA components, and thus each component provides an independent test for association. In the presence of PopStrat, the PA estimate may be biased, whereas the FBA estimate is robust. The type I error and power of PA and FBA tests from separate or joint models are comparable. Finally, we consider tests that combine PA and FBA information, and find that 1df-tests based on a weighted combination of parameter estimates are more powerful that 2df-tests, but the latter are more robust to PopStrat bias toward the null.

**117**

**Genetic association studies using samples ascertained on the basis of a correlated trait**

G.M. Monsees (1), P. Kraft (1,2)
(1) Dept. of Epi (2) Dept. of Biostat, Harvard School of Public Health, USA

Large cohorts with prospectively-measured biomarkers are very expensive. Utilizing existing data from nested case-control studies may reduce genotyping or biomarker ascertainment expenses when investigating gene-biomarker relationships. However, improperly accounting for selection based on case/control status could introduce bias in the estimates of gene-biomarker association. We use simulation to compare the relative bias and efficiency of six methods for testing association between a dichotomous genetic factor (G) and a continuous biomarker (X), within the framework of a nested case-control study. We compare the following analysis methods, each in the form of a linear regression: regressing X on G (ignoring case-control status D); regressing X on G adjusting for D; regressing X on G adjusting for D and allowing for G-D interaction; regressing X on G restricted to controls (or cases); and inverse probability-of-sampling weighted (IPW) linear regression. This latter method assumes the sampling fractions for cases and controls are known or estimable (as will generally be the case for nested case-control studies). All methods appear unbiased and have appropriate type I error rates when either G or X is independent of D. IPW is the only unbiased method when both G and X are associated with D; however, it has lower power than other methods. The most appropriate analysis will depend on the goal of the study: gene characterization studies may favor IPW for its lack of bias, whereas large genomic screens may necessitate a more powerful approach.

**118**

**Multifactor Dimensionality Reduction 1.0**

J.H. Moore, B.C. White, N. Barney
Dept. of Genetics, Dartmouth Medical School, USA

Multifactor dimensionality reduction (MDR) was developed as a computational alternative to parametric statistical methods for detecting, characterizing, and interpreting epistasis in genetic studies of common human diseases. MDR uses a constructive induction approach to change the representation of the data to make interactions easier to detect, especially when statistically significant main effects are not present. Our goal was to make MDR available to the genetic epidemiology community through a software package that is open-source, freely-available, user-friendly, and platform-independent. The resulting software includes an intuitive graphic user interface (GUI) for general users but can be run from the command line on a parallel computer, for example. We released the first beta version of MDR in February of 2005 and made it freely available for download via sourceforge.net and www.epistasis.org. We describe here a mature version 1.0 of the MDR software that is programmed entirely in Java and is the result of more than three years of development and testing. The MDR software and its components have been downloaded more than 10,000 times since its first release placing it in the top 40 among more than 1,000 bioinformatics software packages maintained and distributed via sourceforge.net. PubMed currently lists more than 70 publications with "multifactor dimensionality reduction" in the abstract or title making it of the most commonly used data mining methods for modeling interactions. Future versions will include additional tools for modeling epistasis in genome-wide association studies.

**119**

**Estimation of trait parameters in human QTL mapping under different ascertainment schemes**

I. Mukhopadhyay (1), E. Fiengold (2), D.E. Weeks(2)
(1) Department of Statistics, University of Burdwan, India
(2) Department of Human Genetics and Biostatistics, University of Pittsburgh, USA

Abstract: In almost all the regression-based statistics and the score statistics methods of QTL mapping, usually it is assumed that the trait parameters (segregation parameters) $\Psi$ (=($\mu$, $\sigma^2$, $\rho$)) are known. This is rarely so unless one has a very good idea about $\psi$ from a previous study. Through simulation, Sham et al (2002), T. Cuenco et al (2003a, 2003b) and Szatkiewicz and Feingold (2005) studied the effect of misspecification of the segregation parameters considering a few values, which differs from the true values of the parameters, for population samples and selected samples. They showed that departure from the true values of $\psi$ results in substantial loss of power. In view of this it becomes important to develop a proper estimation procedure of the segregation parameters involved in regression-based statistics and score statistics. In this work, we propose some estimation methods for $\psi$ based on population samples as well as selected samples. We discuss our proposed methods of estimation under several ascertainment schemes e.g. affected concordant sib-pair, discordant sib-pair, discordant and concordant sib-pair etc. Through simulation, we evaluated the performance of several test statistics by calculating probability of Type I error and power under several genetic models. Keeping the probability of Type I error at the desired level, the power of the statistics using our proposed method is very close to the power that can be obtained assuming the true values of the segregation parameters.

**120**

**Meta-analysis of genome-wide linkage studies: optimal bin width for the GSMA**

M.Y. Ng, C.M. Lewis
Department of Medical and Molecular Genetics, King's College London School of Medicine, UK

Linkage studies of common complex disorders have low power to detect linked regions. Genome search meta-analysis method (GSMA) pools results from different linkage scans to identify novel linked regions or confirm susceptibility regions. GSMA divides the genome into n bins of approximately equal width. For each study, the strongest linkage evidence within each bin is identified, and bins are subsequently ranked in descending order. The rank of each bin is then summed across studies and assessed by Monte Carlo simulation. We compared the power to detect linkage of the original 30cM bin width to bins of 20cM and 40cM. A dominant disease locus model with sibling relative risk ($\lambda_s$) of 1.15 and 1.3 located at either 10cM or 80cM on a 180cM chromosome was simulated. Each chromosome has 18 markers with 10cM inter-marker spacing, and the simulation comprised one linked and 19 unlinked chromosomes. We simulated 200 affected sib pair families, with different parental genotype availability. We determined power to detect suggestive evidence for linkage, which controls for different numbers of bins across the genome. For the 80cM locus, the average power of the 20, 30 and 40 cM bin widths at $\lambda_s$=1.15 were 0.82, 0.85 and 0.68, respectively at the suggestive threshold and the same pattern was observed at $\lambda_s$=1.3. The bin widths performed equivalently for the 10cM locus. The results imply that bin widths of 20 and 30cM have similar power to detect linkage, but larger bins may be less effective when the disease locus is not telomeric.

## 121
## Family-based association analysis of polymorphisms in CCAAT/enhancer binding protein genes in relation to quantitative cardiovascular disease risk factors

J. Nsengimana (1), J.H. Barrett (1), C.E. Bennett (2), J.A. Bostock (2), C.M. Cymbalista (2), D.J. Rolton (2), T.J. Scarrott (2), P.J. Grant (2), A.M. Carter (2)

(1) Leeds Institute of Molecular Medicine and (2) Academic Unit of Molecular Vascular Medicine, University of Leeds, Leeds, UK

Highly heritable anthropometric, metabolic and fibrinolytic factors are associated with an increased risk of cardiovascular disease. The aim of this study was to test for association between 15 of these factors and single nucleotide polymorphisms (SNPs) in three genes: CCAAT/enhancer binding proteins alpha, beta and delta (CEBPA, CEBPB and CEBPD). The study is based on 537 intensively phenotyped subjects from 89 multigenerational pedigrees. Each trait was tested for association with a comprehensive set of 24 SNPs in these genes (11 in CEBPA, 11 in CEBPB and 2 in CEBPD). Single SNP analysis was carried out in STATA using random effect models allowing for the correlation among family members. Two-locus haplotype analysis was carried out using the family-based association test implemented in PBAT. At significance level 0.001 association was found between a SNP in CEBPA and fasting glucose, a SNP in CEBPD and tissue plasminogen activator and several SNPs in CEBPB with waist-to-hip ratio (WHR). The association between CEBPB SNPs and WHR was further confirmed but not strengthened by haplotype analysis, which also revealed weak evidence of association between Factor XIII protein and SNPs in CEBPA. This study demonstrates the utility of studying complex diseases through their intermediate phenotypes.

## 122
## kerfdr: kernel based estimation of the local False Discovery Rate

G. Nuel(1), M. Guedj(1,2), A. Celisse(3) and S. Robin(3)
(1)Statistics and genome (UMR Evry Univ, CNRS 8071, INRA 1152), Evry, France, (2)Serono, Geneva, Switzerland, (3)Statistics and genome (UMR AgroParisTech, INRA 518), Paris, France

The use of current high-density microarray genomic association studies leads to the simultaneous evaluation of a huge number of statistical hypotheses and at the same time, to the multiple-testing problem. As an alternative to the often too conservative FWER (Family-Wise Error Rate), the FDR (False Discovery Rate) criterion as been introduced by Benjamini-Hochberg. One drawback is that the FDR is associated to a given rejection region and hence refers to all the test statistics within the region without distinguishing those that are close to the boundary and those that are not. As a result, Efron introduced recently the local FDR which quantifies marker-specific evidence for being associated.
In this context, we first propose a didactive presentation of the notion of local FDR through simple but illustrative Gaussian mixtures then we introduce with kerfdr (R package) an efficient implementation of a new non-parametric approach based on kernel estimations. This method allows both to consider complex heterogeneities in the alternative hypothesis and to take into account a priori knowledge (from expert judgment or previous study) by offering a semi-supervised mode. Finally, to demonstrate the versatility of our new approach, we apply kerfdr to a wide range of datasets (DNA patterns, gene expression, association studies).

## 123
## A Simulated Genetic Structure for Bipolar Illness

John I. Nurnberger, Jr. MD PhD
Indiana University School of Medicine

Bipolar illness is conceptualized as a polygenic condition. Based on candidate gene findings from the literature to date, up to 23% of the genetic risk may be explained by 6 gene variants with an average allele frequency of 0.58 in cases and 0.54 in controls. The mean allele specific relative risk (ASRR) for these variants is 1.07. Initial results from genomewide association studies tend to confirm this estimate of effect size. A 30 allele model is presented in which the average affected person would carry 22 susceptibility variants, and the average unaffected person would carry 15. In a comparable model with 100 alleles, the average affected person would carry 62 susceptibility variants compared with 50 in unaffecteds. Thus common gene variants associated with bipolar disorder may be expected to be widely distributed in the general population. The neurobiology of the replicated candidate genes is considered, allowing an initial delineation of relevant biological pathways.

## 124
## Training in Genetic Epidemiology - Implementation of a technology based training course

F. Pahlke(1), I.R. König(1), M. Bischoff(2), A. Ziegler(1)
(1)Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Germany, (2)oncampus Fachhochschule, Lübeck, Germany

Even though the importance of genetic epidemiology as a scientific field has been widely recognized over the past decades, only very few technology assisted training opportunities have been offered in the last years. Specifically, no comprehensive technology assisted training course with sound didactical justification is available for our field. The goal of our project is to construct a self-learning online training course with a content covering about a five days course, based on the textbook "A Statistical Approach to Genetic Epidemiology" from A. Ziegler and I. R. König. Until now, the content of about a quarter has been implemented as a highly interactive e-learning module. In this presentation, we describe the process of building the raw concept and the storyboard. Also, the implementation of multimedia elements is illustrated, e.g. interactive Flash-based pedigree diagrams and interactive problems with algorithm-based free-text correction. Acknowledging that the e-learning projects at universities are produced under different conditions than in industry, we used a specific procedure tailored for academic use, which allows a high degree of flexibility and emphasizes the didactical concept – instead of the technical implementation. With our course, students and scientists of very different fields of research will get a flexi-time and flexi-location training opportunity in genetic epidemiological methodology and design.

## 125
## An adjusted instrumental-variable model for Mendelian randomization

T.M. Palmer, P.R. Burton, J.R. Thompson, M.D. Tobin
Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, Leicester, UK

The principle of Mendelian randomization uses the associations between genotype and disease and between genotype and phenotype to make inferences about the association between phenotype and disease [1]. In the case where all variables are continuous measures traditional two-stage least squares instrumental variable methods are appropriate, however, the majority of genetic epidemiological studies are of case-control design. Therefore a modelling approach allowing for a dichotomous disease variable is required.
A standard approach would be to use a logistic regression model at the second stage of the instrumental variable procedure, with perhaps some adjustment of the standard errors [2]. However with a dichotomous disease variable this standard approach is affected by shrinkage bias and unmeasured confounding [3]. An adjusted instrumental variable model for the analysis of a genetic case-control study is proposed and investigated through a simulation study. The adjusted model is shown to have superior properties in terms of reducing the bias in the parameter estimates and accounting for unmeasured confounding factors compared a standard instrumental-variable approach to a dichotomous disease variable. The adjusted two-stage model is also shown to produce equivalent but more precise parameter estimates for a linear disease outcome measure.

References

1. Davey Smith, G. and Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. International Journal of Epidemiology, 2003, 32, 1–22.
2. Hardin, J.W. and Carroll, R.J. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. The Stata Journal, 2003, 3, 342–350.
3. Zeger, S.L.; Liang, K and Albert, P.S. Models for longitudinal data: a generalized estimating equation approach. Biometrics, 1988, 44, 1049–1060.

## 126
## The Ordered Transmission Disequilibrium Test: a method for modifier gene detection

Hervé Perdry (1,2), Marie-Claude Babron (1,2), Françoise Clerget-Darpoux (1,2)
(1) INSERM UMR-S 535, Villejuif, F-94817
(2) Univ. Paris-Sud, Villejuif, F-94817

A modifier gene in a disease induces a clinical heterogeneity that may be measured by a quantitative variable, such as the age of onset or the severity of the disease.
We designed the Ordered Transmission Disequilibrium Test (OTDT) to test for a relation between such a quantitative variable, and marker genotypes of a candidate gene. The method applies to trio families with one affected child and his parents. Each family member is genotyped at a bi-allelic marker M of a candidate gene. As the TDT (Spielman et al, 1993), the OTDT is based on the observation of the transmission rate T of a given allele at M; as the Ordered Subset Analysis (Hauser et al, 2004), it uses the values of the variable to order the sample of families.
The OTDT aims to detect whether T depends on the value of the variable. The principle of the test is to seek for a threshold value of the variable, which separates the diseased in two groups, in which transmission rates are different. The threshold is chosen such that the significance of the difference between the two transmission rates is maximal.
We compare the powers of the OTDT and of the Quantitative TDT (QTDT) (Abecasis et al, 2000). This comparison shows that, as expected, the QTDT is more powerful under models in which the variable follows normal distributions, the means of which depend on the genotype of the diseased. However, the OTDT is robust to non-normality, allowing even to detect a locus which modulates the mode of the distribution of the variable, but not its mean and variance.

**127**

**A Simple Correction for Population Stratification in Genome-Wide Haplotype Sharing Analysis**

D. Qian
Department of Biostatistics, City of Hope, Duarte, California, USA

Population stratification is an important issue in case-control genome-wide association studies. Haplotype sharing analysis is an attractive approach because of its robust signals at true disease loci and low false positives at non-disease loci. To correct population stratification as well as some other genetic heterogeneities, an inverse function of shared intervals is defined to quantify the between individual distance, a density-based clustering is used to assign subjects to clusters, and the association between disease status and haplotype sharing is evaluated and summed over the clusters. A simulation is conducted to generate case-control data from a population consisting of multiple latent subpopulations on multiple chromosomes, and the haplotype sharing association results are compared with and without the correction of population stratification.

**128**

**FFIGdb – A Fast, Flexible, Integrated Genotype Database capable of storing very large genotype and phenotype data sets**

N.W. Rayner, E. Zeggini, M.I. McCarthy
and the international type 2 diabetes 1q consortium.
Oxford, UK

Recently published genome wide association studies have shown the speed with which data sets are growing and with genotype imputation data set sizes are set to continue to grow rapidly. Efficient exploitation of these data requires dedicated data management systems capable of storing and manipulating these massive data sets.

FFIGdb is a high-throughput data management system initially developed for the International Type 2 Diabetes 1q Consortium. The system is capable of storing both phenotype and genotype data as well as analysis results and is currently based on an Oracle database utilising a Perl API and client interface.

Unlike most previous systems the system has been designed from the outset to utilise Enterprise level software and is capable of processing multibillion point data sets on lower end hardware. Exports can be in a wide variety of formats, including Stata and Plink. The database utilises an Entity-Attribute-Value (EAV) data model for storing phenotype, sample and marker data, which provides great flexibility allowing for user-defined data sets to be easily loaded thereby permitting it to be extended to any project.

The system also incorporates many features for data management such as genotype and phenotype Quality Control, duplicate sample- and marker-checking and audit trails.

To date, the system is handling multibillion point genome wide association data sets with ease. Development continues to ensure the speed, storage capability and flexibility all keep pace with ever-increasing data sizes being generated.

**129**

**SNP-pair Tetrahedron: Geometric Presentation of Haplotype Space of Pairwise SNPs**

R. Yamada(1,2)
(1)HGC, IMS, University of Tokyo, Japan, (2)CGM, Grad. School of Medicine, Kyoto University, Japan

We recently published a new method to express frequency of haplotypes for a locus with multiple SNPs, which gave a generalized definition of linkage disequilibrium (LD) for multiple SNPs. Assume a locus with n SNPs. This new method gave a set of haplotype frequencies a linear expression with 2^n-1 variables, that were mutually independent. In case of SNP pairs, a frequency vector of four haplotypes is expressed with 2^2−1=3 variables. Two of the three variables correspond to frequency of two SNPs and the last one represents LD. The three variables are defined so that they are mutually independent; therefore the corresponding three vectors are orthogonal. We do not place the three orthogonal vectors as base vectors for three dimensional Euclidean space but propose to let them construct a regular tetrahedron so that components of the tetrahedron represent conditions of pairwise SNPs: Tetrahedron is consisted of four vertices, six edges and four faces. Four vertices stand for clonal conditions. Four out of six edges correspond to conditions where only one of two sites is polymorphic. The two other edges are conditions where a LD index r^2=1. Four faces are conditions where another LD index, D′=1. The space inside of the tetrahedron represents conditions where four haplotypes exist. The tetrahedron contains a curved surface of linkage equilibrium. This presentation characterizes the SNP-pair tetrahedron and introduces its applications to evaluations of genetic heterogeneity and linkage disequilibrium.

**130**

**Evaluation of different type of bivariate analyses in a genome-wide search for pleitropic loci on two Bone Mass Density quantitative traits**

A Saint-Pierre (1), M Cohen-Solal (2), A Ostertag (2), MC de Vernejoul (2), JM Kaufman (3), M Martinez (1)
(1) INSERM U563, France, (2) INSERM U606, France, (3) Gent University, Belgium

Multivariate linkage can efficiently analyze multiple traits while controlling for type-I error rate, but it may be computationally intensive. Alternative "simplified multivariate tests" have been proposed: the dimension of the data is reduced by constructing independent Quantitative Traits from, for instance, principal components [Elston et al., Genet Epidemiol, 2000] analysis. Joint linkage can be assessed either from univariate tests or the Combined Test [Mangin et al., Biometrics, 1998] of the new QTs. In simulated nuclear family with one marker data, CT was found liberal and less powerful than the multivariate test [Gorlova et al., Ann Hum Genet, 2002] or the univariate test(PC1)on the 1st principal component [Kaabi & Elston, Genet Epidemiol, 2003]. Here, we applied CT and PC1 tests to a genome-wide search for Bone Mass Density at the Femoral Neck and at the Lumbar Spine in the NEMO (103 extended pedigrees) data. Multipoint linkage analyses are conducted using variance

component approach with Merlin. Type I error levels of CT, PC1 and LOD of LS/FN-BMDs tests are derived through simulations in the NEMO data. Overall, we found good agreement between the two "simplified bivariate tests" and, PC1 seemed to perform better in these data. We are performing efficient bivariate linkage analysis using SOLAR. We will discuss the significance of the detected joint linkages under the different strategies.

## 131
## Integration of SIMLA and SIMLAPLOT: A Graphical User Interface for Complex Disease Simulation and Analysis

Silke Schmidt, Ren-Hua Chung, Xuejun Qin, Xuemei Lou, Elizabeth R. Hauser
Center for Human Genetics, Duke University Medical Center, Durham, NC

We have integrated two software packages that facilitate simulation studies for complex human diseases: an enhanced version of our previously distributed SIMLA package (Schmidt et al. 2005) that is now available with a Graphical User Interface (GUI) for generating the control file, and our recently developed visualization tool SIMLAPLOT (Qin et al. 2007). Specifically, we have implemented the following new features: simulation of unrelated case-control datasets; simulation of up to three modifier loci, each of which can generate a categorical phenotypic feature, such as disease severity; simulation of a biallelic quantitative trait locus (QTL) that may influence the distribution of a continuous disease risk factor and/or generate a disease-unrelated trait for QTL analysis; simulation of X-linked disease loci and sex-specific relative risks; simulation of up to four blocks of markers in linkage disequilibrium (LD); LD calculation tool for generating founder haplotype frequencies given user-specified values of $D'$ or $r^2$; and a sibling recurrence risk ($\lambda_s$) calculation tool. SIMLAPLOT graphically illustrates a variety of models by which continuous environmental or clinical covariates may influence the risk of complex human diseases, in concert with genetic susceptibility. Examples for such models include gene-environment interaction, the QTL mechanisms described above, and genetic main effects with covariate-based heterogeneity. SIMLAPLOT may be used to better understand the role of various model parameters in a SIMLA control file by graphically displaying the relationship between disease locus (or QTL) genotypes and covariate values. When applied to real datasets, plots produced by SIMLAPLOT may assist in the interpretation of statistical analysis results and the exploration of plausible disease models.

## 132
## Improvement of haplotype sharing analysis in localization of disease loci using entropy based marker selection

A Schulz(1), C Fischer(2), J Chang-Claude(1), L Beckmann(1)
(1)German Cancer Research Center DKFZ, Heidelberg, Germany
(2)Institute of Human Genetics, University of Heidelberg, Germany

The previously introduced Mantel Statistics using Haplotype Sharing method correlates genetic and phenotypic similarity to map complex disease-genes, and proved powerful in narrowing candidate regions. Haplotype estimation for consecutive markers not sharing a common recent evolutionary history may lead to reduced estimated haplotype heterogeneity. This may increase the correlation of subsequent pointwise haplotype sharing analysis, and cause neighboring markers to be significant at similar magnitude in regions of high LD. We present a new method to more precisely localize risk genes by combining Mantel Statistics with iterative entropy-based marker selection. For each marker to be tested, a subset of surrounding markers is chosen by maximizing multilocus linkage disequilibrium, measured by the Normalized Entropy Difference. We evaluate the approach w.r.t. type I error, power, the length of sharing covered, and localization of the disease variant, comparing it in simulation scenarios to the algorithm i) without marker selection and ii) considering haplotype block structure. Despite its considerable conservativeness in some simulations, the new algorithm yields better power than no selection at the causal locus. There the new method is clearly more often significant than at neighboring markers, thus improving the precision of the localization of risk genes. The approach may be favorable in settings where recent mutations interrupt shared chromosomal regions around a disease.

## 133
## Genome-wide association studies get in a flow

D.F. Schwarz(1), I.R. König(1), S. Szymczak(1), A. Götz(1,2), F. Pahlke(1), T. Strom(4), W. Lieb(2), B. Mayer(2), T. Meitinger(3,6), H.E. Wichmann(4,5), J. Erdmann(2), H. Schunkert(2), A. Ziegler(1)
(1) Institut für Medizinische Biometrie und Statistik and (2) Medizinische Klinik II, Universitätsklinikum Schleswig-Holstein, Campus Lübeck; (3) Institut für Epidemiologie and (4) Institut für Humangenetik, GSF-Forschungszentrum, Neuherberg; (5) Institut für Humangenetik, Klinikum rechts der Isar der TU München; (6) Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie, LMU München

A challenge of GWAs is the amount of generated data, with accompanying storage problems and extensive time for analyses. Even optimized work flows generate ~130MB per person of raw cel-files using the 500K chip from Affymetrix. Also, a useful interface to exchange the results with co-operation partners is required.
Based on our experience from several GWAs, we established a nearly automatic work flow encompassing (1) generation of genotypes, (2) intensive quality checks, (3) statistical analysis, and (4) result management. In (1), we perform BRLMM genotype calling. Among other checks in (2), we generate and automatically evaluate signal intensity cluster plots. Analyses in step (3) are performed utilizing the toolset Plink, which allows to add own implementations, e.g., random forest variable importance. In (4), results are merged with marker information and fed into a MySQL database. A secure www based user interface to access data is built and made available to co-operation partners. This presentation details our work flow, which enables us to accomplish the first analysis of a GWA in a few days.

**134**

**DNA methylation patterns tell a tale of cancer stem cells**

K.D. Siegmund (1), P. Marjoram (1), D. Shibata (2)
(1) Department of Preventive Medicine and (2) Department of Pathology, USC Keck School of Medicine, University of Southern California, Los Angeles, CA 90089 USA

Cancer is believed to start from a single transformed cell and progress through a series of clonal expansions. However, virtually nothing is known about the phylogenetic structure of individual human cancers. Cancer cells are thought to be immortal, in the sense that given the right conditions, unlimited progeny are possible. However, recent studies have suggested a hierarchy of cancer cells consisting of relatively rare cancer "stem" cells that produce much larger numbers of "differentiated" cancer cells which have more limited proliferative potential. Cancer growth requires genome replication that, just like phylogeny, can be summarized using ancestral trees. Genome replication involves both the duplication of base order and the duplication of epigenetic patterns such as DNA methylation at CpG sites. At certain genomic sequences, DNA methylation errors accumulate at much faster rates than mutations. We explore a population genetic approach to reconstruct tumor histories from non-functional replication errors, using DNA methylation as a somatic cell "molecular clock". We apply rejection methods to analyze data from three human cancers, finding evidence supporting the existence of cancer stem cells. We also estimate age of the tumor, a variable of considerable interest to epidemiologists who are trying to identify different disease pathways by identifying "young" tumors that may be extremely aggressive from "old" tumors that may be relatively benign.

**135**

**Genetic Substructure in New Hampshire**

C. Sloan, A. Andrew, E. Duell, M. Karagas, J.H. Moore
Dept. of Genetics, Dartmouth Medical School, USA

The impact of geography and ecology on the genetic architecture of common diseases is largely unknown. Understanding the geographic distribution of genetic background is likely to improve our ability to identify both genetic and environmental risk factors. The goal of the present study was to characterize genetic structure in the population of New Hampshire as a first step toward ecogeographic genetic epidemiology of spatially distributed common diseases. We sampled 865 control subjects for an epidemiologic study of cancer from across the state of New Hampshire. We measured 1474 SNPs from approximately 500 cancer susceptibility genes and used Bayesian clustering implemented in the Structure program to identify genetic substructure in this spatially extended sample of subjects. Clusters were evaluated using fixation index (Fst) and admixture statistics along with a novel Hamming distance metric. The Bayesian clustering results suggest four distinct genetic subgroups within New Hampshire (Fst=0.0699, 0.0798, 0.0466, 0.0204). The observed genetic structure may arise from the state's unique ethnic composition that includes a large proportion of Caucasians of French Canadian descent as well as Caucasian individuals of English-Irish descent. The identification of genetic substructure among a largely Caucasian population of European descent in the state of New Hampshire is consistent with recent genetic structure results from studies in Europe including countries such as Iceland. These results will play an important role in helping to explain regional differences in incidence of common human diseases.

**136**

**The Ordered Penetrance Test for Detecting Single-locus Association and Gene-gene Interaction**

M. Song(1), D.L. Nicolae(2)
(1) Dept. of Stat, The Univ.of Chicago, USA., (2) Dept. of Medicine and Stat, The Univ. of Chicago, USA.

In genome-wide studies that search for loci affecting complex traits, two stage strategies, where the analyses in the second stage are done only on markers associated in the first stage have become a common choice for researchers. In this talk, we propose methods for detecting association and interaction in a 2-stage strategy which is shown to be more powerful than the classic approaches. Our method makes use of the fact that many traits are monotone in penetrance and mean and incorporating this knowledge can dramatically increase power. For qualitative traits, we develop likelihood ratio tests for both association and interaction where the asymptotic distributions for both cases are shown to be Chi-bar-squared i.e. weighted sums of chi-squared distributions. Our simulation studies for various models show that the ordered penetrance tests are more powerful compared to other popular tests especially at genome-wide scale. For quantitative traits, analogous tests based on ordered means are proposed and asymptotic results are obtained. We will also show an important extension of our method to testing untyped variation.

**137**

**Robust Statistics for Sib-Pair Linkage Analysis of Quantitative Trait Loci (QTLs)**

M. Sow(1), L. Briollais(2), G. Durrieu(1)
(1)GEMA, University of Bordeaux 1 and CNRS, FRANCE, (2) Samuel Lunenfeld Research Institute, Toronto, CANADA

Study of quantitative traits or intermediate quantitative traits associated with a complex disease may help to dissect its underlying genetic mechanisms. Model-free Haseman and Elston sib-pair methods are appropriate to detect linkage of chromosomal regions to quantitative traits in candidate gene or genome-wide search. The method is based on the least-squares regression of the squared sib-pair trait differences upon the estimated proportion of alleles shared identical-by-descent by the sib-pair at each marker locus. In practice, the standard t-test based on least-squares regression is sensitive to outliers associated to extreme observations and distribution assumption of the trait studied. In this work, we propose to evaluate robust alternatives to the Haseman and Elston approach for linkage analysis of QTLs based on maximum likelihood estimator (M-estimator), linear combinations of order statistics (L-estimator), Wald statistic in non-parametric model using kernel density estimator. Accordingly, we develop and evaluate statistically robust procedures and bivariate regression models for the HE approach. Simulation studies

with 500 sib-pairs show that in presence of outliers generated under a skew-t or skew-normal bivariate distribution, the M- and L-estimators give a type-I error close to the nominal value of 5% and robust procedures have greater power to detect the QTL (approximately 80%) than the standard t-test under the alternative hypothesis. This work could provide a general framework for robust linkage analysis of QTLs.

**138**

**Models for Integrating Genotyping Error into Pedigree Analysis: Application to Linkage Analysis**

W.C.L. Stewart, V. Gateva, G. Abecasis
University of Michigan

To mitigate the negative and potentially severe effect of genotyping error, we describe a method that incorporates genotyping error within a pedigree analysis. Our method treats true genotypes as latent information that influences observed genotypes through an error model. We illustrate the utility of our method in the context of affected sibpair linkage analysis, and compare its performance to standard approaches. In contrast to our method, the standard approaches can only analyze Mendelian-consistent subsets of the observed data, and cannot integrate error within the linkage analysis.

We simulated multipoint linkage data at single nucleotide polymorphisms (SNPs) in the presence of genotyping error. In our simulations, we considered a variety of different error models, error rates, marker maps, missing data patterns and family structures. The results of our analyses show that (1) when one or both parents are not genotyped, our method is considerably more powerful than the standard approaches; (2) relative to the information contained in the true genotypes and provided that high-resolution SNP based linkage maps are used, our method retains most of the linkage signal, even for error rates as high as 2%; and (3) in almost all cases, the negative effect of genotyping error decreased as the resolution of the map increased.

We have implemented our approach within the MERLIN package, where it can be used for many common pedigree analyses. In addition, our method estimates error rates from available genotype data, and can be used to evaluate the performance of new genotyping technologies.

**139**

**A comparison of the distribution of extreme p-values under alternative genetic models**

H. Sung, A.J.M. Sorant, A.F. Wilson
Genometrics section, NHGRI, NIH, Baltimore, MD, USA

As the number of SNPs that can be genotyped has increased, the probability of finding ¡extreme¡ ± p-values has increased proportionally. When results are presented from GWAS with 500K SNPs (or more), there is concern that the results reported may mostly be type I errors or that normal distribution theory breaks down in these situations. In this study we compared the distributions of extremely significant p-values under the null and alternative hypotheses. Samples were generated using the G.A.S.P. (v3.3) under a variety of conditions. In each case, two traits were generated: one caused in part by a genetic locus and one unrelated to the locus. Samples varied in structure, size (100, 500), allele frequency (1%, 10%, 20%), trait heritability (1%, 5%, 10%), dominance, adherence to Hardy-Weinberg equilibrium, and an additional genetic component. Each trait was analyzed with a simple linear regression on the genotype of the trait locus, assuming an additive model. Occurrences of very small p-values were compared between the associated trait and the unassociated one (the null distribution). For fewer than 2% of the models, the distribution under the null hypothesis had the most significant (most extreme) p-value. Those instances occurred when the heritability was low and (mostly) small sample size. Violation of the additivity assumption in the analysis also appeared to affect the distribution of the p-value. In general, very significant p-values were more frequent for the associated trait (under the alternate hypothesis), as predicted by normal theory.

**140**

**Comparing Variable Selection Methods for Genetic Association Studies**

M.D. Swartz, (1,2), R.K. Yu, (1) and S. Shete, (1)
(1) Dept. of Epi, U. T. M. D. Anderson Cancer Center, USA
(2) Dept. of Stat, Texas A&M, USA

Genetic epidemiologists are primarily interested in discovering the modulating factors for a disease, both genetic and non-genetic. One of the more common models used for such investigations is logistic regression. However variable selection methods, especially Bayesian, are under-utilized in the search for risk factors. We performed a simulation study to compare five methods of variable selection: (1) using a confidence interval approach for significant coefficients (CI), (2) backward selection (BS) (3) forward selection (FS) (4) stepwise selection (SS) and (5) stochastic search variable selection (SSVS). We defined our simulated model mimicking odd ratios for cancer risk found in the literature for environmental factors such as smoking and alcohol; diet risk factors such as fiber and folate; and genetic risk factors such as MTHFR and XPD. We also modeled the distribution of our covariates after the reported empirical distributions of these factors, including correlation. We simulated multiple interaction models. Preliminary results from one scenario shows that for average false positives, SSVS had the lowest rate (0.5%), then the CI approach (8.7%) followed by FS and SS (tied at 15%) and BS (17%). Using the average false negative rate, BS, the CI approach and SSVS performed comparably (0.5%, 0.2% and 1.5%, respectively) while FS and SS tied for last (15%). The covariate specific false positive and false negative rates, and other models will also be presented.

**141**

**Mutual information in genetic association studies: a simulation-based comparison with parametric and nonparametric ANOVA**

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

In genetic association studies using gene expression data, conditional distributions of quantitative phenotypes given genotype are highly skewed, heavy tailed or contaminated with outliers. This is in contrast to usual normality assumptions of common statistical methods and thus, robust nonparametric procedures are sensible alternatives. Recently, Tsalenko et al. (J Bioinform Comput Biol, 4, 2006, 259–274) proposed an entropy based "quantitative mutual information score" (QMIS) to analyze the relationship between single-nucleotide polymorphisms (SNPs) and transcript levels. In our work, we compare QMIS with parametric analysis of variance (ANOVA) and its nonparametric counterpart, the Kruskal-Wallis test (KW) using simulated expression and genotype data. More precisely, we determine the number of significant p-values of these methods considering various types of distributions. We focus on different location, variance and skewness parameters of low and also heavy tailed distributions. In addition, we vary the overall sample size and genotype frequencies. We observe substantial differences between the statistical methods of interest depending on specific model assumptions. In general, if data distributions are highly skewed nonparametric methods should be preferred to determine differences in location. In particular, the QMIS approach is robust against extreme cases of different shapes and consequently, leads to sensible results.

## 142
## Quantifying disease genotype enrichment under multiple causative loci

F. Takeuchi, R. McGinnis
Wellcome Trust Sanger Institute, Cambridge, UK

Enrichment of disease-causing genotypes in affecteds versus unaffecteds plays a key role in detectability of disease loci by linkage or association. Designate a polymorphism being tested for linkage or association as "foreground" and other disease loci as "background". In unrelated affecteds, mean foreground genotype frequency is unchanged by background loci but in related affecteds the enrichment of foreground susceptibility genotypes can be attenuated by background loci. However, understanding of this phenomenon rests on simulated results [Howson et al. Genet Epidemiol 2005. 29:51–67].
We quantify enrichment of foreground genotypes by assigning a "risk" to each background and foreground genotype such that the penetrance of each multilocus genotype is determined by combining these risks either multiplicatively or additively. For comparison purposes, we also consider a single locus model with no background loci.
As previously shown [ibid], foreground genotype frequencies for multiplicative models are identical to those of a corresponding single locus model. By contrast, the frequency of disease-predisposing foreground genotypes in additive models varies between the highly enriched frequency of the single locus model and unenriched frequency in the general population. For related affecteds, the degree of enrichment in this range is the ratio of recurrence risks for the single locus and corresponding multilocus models.
Our quantification of this phenomenon and of an analogous one impinging on allele sharing in affected sib pairs enables better evaluation of association and linkage studies of diseases caused by multiple loci.

## 143
## The effect of MBL2 gene variants on the age of first infection in Cystic Fibrosis (CF) patients is modified by TGFβ1 gene variants

C. Taylor (1), B. Huang (2), R. Dorfman (2), J. Zielenski (2), A. Sandford (3), P. Paré (3), P. Durie (1), K. Schmansky (3), M. Corey (1).
(1) Child Health Evaluative Sciences, Toronto Hospital for Sick Children, Canada, (2) Molecular Medicine, Toronto Hospital for Sick Children, Canada, (3) University of British Columbia, Canada

The poor correlation between CFTR gene variants and lung disease severity in CF has stimulated the search for genetic modifiers that may be related to disease events or progress. Many of these studies have been inconclusive or contradictory.
We studied two potential modifiers – Mannose Binding Lectin (MBL) and Transforming Growth Factor **β**1 (TGF**β**1). In the Canadian CF Modifier Study, we studied the effect of MBL and TGF**β**1 on age at first *Pseudomonas aeruginosa* (PA) infection in 1000 CF patients. PA is an organism that most CF patients eventually acquire and infection is correlated with lung function and survival.
Using time to event analysis, we found that MBL gene variants are related to age at first infection: those producing less MBL protein developed PA earlier. Although TGF**β**1 was not related to age at first infection, the effect of MBL on age at first infection was modified by TGF**β**1. High TGF**β**1 expressing variants worsened the negative effect of low level MBL.
In conclusion, genetic modifiers of CF lung disease can themselves be modified by other genetic factors indicating that CF is a complex genetic disease in which we can expect to observe interactions between genetic and environmental factors.

## 144
## Gene based association tests using inferred haplotypes for case-control samples

A. Thalamuthu (1), D.E. Weeks (2)
(1) Division of Human Genetics, Genome Institute of Singapore, Singapore, (2)Department of Human Genetics and Biostatistics, University of Pittsburgh, Pittsburgh, USA

We propose a haplotype-based approach for testing joint association of markers within a candidate gene with a disease. First we statistically infer the haplotypes of individuals in a sample of cases and controls. We then model the probability of the disease given the two haplotypes of an individual using product of two logistic transformations. A weighted Bernoulli likelihood accounting for the phase ambiguity of the haplotypes is constructed for case-control samples. A Likelihood Ratio (LR) test is derived for the global or joint association of all the haplotypes. The novelty of the global test proposed here is that the number of degrees of freedom for the test depends only on the number of SNP markers used in the construction of haplotypes and not on the number of observed haplotypes. The method proposed here can be used testing gene-based associations involving several markers

within a candidate gene. Further, to identify specific disease causing variants within the gene, an indirect test of association of individual markers using Wald statistic is proposed. Our simulation study shows that the method proposed here has increased power for joint effect of markers within a gene compared to two popular methods (HAPLO and WHAP). Also the tests of individual marker specific associations proposed here is able to identify the disease causing variant within a candidate gene.

**145**

**Shared genomic segment analysis. A novel approach to mapping disease redisposition genes in extended pedigrees using dense single nucleotide polymorphism assays**

A. Thomas, N.J. Camp, J.M. Farnham, K. Allen-Brady, L.A. Cannon-Albright.
Department of Biomedical Informatics, University of Utah, USA.

We examine the utility of high density genotype assays for predisposition gene localization using extended pedigrees. Results for the distribution of the number and length of genomic segments shared identical by descent among relatives previously derived in the context of genomic mismatch scanning are reviewed in the context of dense single nucleotide polymorphism maps.
We use long runs of loci at which cases share a common allele identically by state to localize hypothesized predisposition genes. The distribution of such runs under the hypothesis of no genetic effect is evaluated by simulation.
Methods are illustrated by analysis of an extended prostate cancer pedigree previously reported to show significant linkage to chromosome 1, p23.
Our analysis establishes that runs of simple single locus statistics can be powerful, tractable and robust for finding DNA shared between relatives, and that extended pedigrees offer powerful designs for gene detection based on these statistics.

**146**

**Confronting Complexity in Late-Onset Alzheimer Disease: Application of Two-Stage Analysis Approach Addressing Heterogeneity and Epistasis**

T.A. Thornton-Wells(1), J.H. Moore(2), E.R. Martin(3), M.A. Pericak-Vance(3), J.L. Haines(4)
(1) Vanderbilt Kennedy Center for Research on Human Development, Vanderbilt University Institute for Imaging Science
(2) Departments of Genetics and Community and Family Medicine, Dartmouth Medical School
(3) Miami Institute for Human Genomics, Miller School of Medicine, University of Miami
(4) Center for Human Genetics Research, Vanderbilt University Medical Center

Common diseases with a genetic basis are likely to have a very complex etiology. A new comprehensive statistical and computational strategy for identifying the missing link between genotype and phenotype has been proposed, which emphasizes the need to address heterogeneity in the first stage of the analysis and gene-gene interactions in the second stage. We applied this two-stage analysis strategy to late-onset Alzheimer disease (LOAD) from 654 families and an independent set of 451 cases and 699 unrelated controls. Bayesian Classification found significant clusterings ($p < 0.002$) for both datasets, which used the same five SNPs in LRRTM3 as the most influential in determining cluster assignment. In subsequent analyses to detect main effects and gene-gene interactions, SNPs in three genes—PLAU, ACE and CDC2—were found to be associated with LOAD in particular subsets of the data based on their LRRTM3 multilocus genotype ($p < 0.05$). All of these genes are viable candidates for LOAD based on their known biological function. Further studies are needed to replicate these statistical findings and to elucidate possible biological interaction mechanisms between these genes and LRRTM3.

**147**

**Gender specific effects of a common genetic variant in the NOS1 regulator NOS1AP on cardiac repolarization**

M.D. Tobin (1), M. Kähönen (2), P. Braund (3), T. Nieminen (5), C. Hajat (1), M. Tomaszewski (3), J. Viik (6), R. Lehtinen (2), G. Andre Ng (3), P.W. MacFarlane (7), P.R. Burton (1), T. Lehtimäki (4), N.J. Samani (3).
(1) Dept. Health Sciences, Univ. of Leicester, LE1 7RH, UK, (2) Dept. Clinical Physiology, Univ. of Tampere, Finland, (3) Dept. Cardiovascular Sciences, Univ. of Leicester LE3 9QP, UK, (4) Dept. Clinical Chemistry, Univ. of Tampere, Finland, (5) Dept. Pharmacological Sciences, Medical School, Univ. of Tampere, (6) Ragnar Granit Institute, Tampere Univ. of Technology, Finland 8 Div. Cardiovascular Medical Sciences, Univ. of Glasgow G31 2ER

A longer heart-rate corrected QT interval QTc on the surface ECG is associated with ventricular arrhythmias and sudden death. Women have a longer resting QTc interval and are 2–3 times more likely than men to develop drug-induced QT prolongation. Molecular regulation of QTc is incompletely understood. We studied association of a common NOS1AP variant recently shown to affect resting QTc with QT interval at rest in 919 women and 918 men from 504 families in the GRAPHIC study, representative of the UK population. NOS1AP genotype was significantly P=1.3*10−5 associated with QTc interval in GRAPHIC participants: the minor allele G of rs10494366 prolonged QTc by 4.59 milliseconds 95% CI 2.77 to 6.40; P=7.63*10−7 in women, but only by 1.62 ms 95% CI −0.15 to 3.38; P=0.073 in men gender-SNP interaction term P=0.025. We will also report emerging findings on the impact of the SNP on QTc at rest and following exercise in a population of Finnish men and women referred for exercise stress testing.

**148**

**Strong evidence for association of UNC13b gene with diabetic nephropathy: the EURAGEDIC study**

the EUropean Rational Approach for GEnetics of DIabetic Complications (EURAGEDIC) Study group

In the first part of this study, 437 SNPs located in 144 candidates genes were sought for association with diabetic nephropathy (DN) in a large Type 1 diabetes case/control (1176/1323) study fro 3 European populations. A multi-stage strategy involving Random Forest, DICE and haplotypic methods was used to detect marginal and epistatic effects. Among the genetic effects identified by this strategy, one single locus effect was found consistently in the 3 populations. The T allele of the UNC13b_134327 SNP was less frequent in cases than in controls (0.281 vs 0.324) leading to a common odds ratio of 0.746 [0.624 – 0.892] (p=0.0013). In a second step, 20 additional tagging SNPs spanning the UNC13b gene were further genotyped to clarify the contribution of the gene to the susceptibility to DN. Single- and multi-locus analyses were performed and identified another SNP as a candidate risk factor for DN. Homozygous carriers of the T allele at this SNP were more frequent in cases than in controls (0.20 vs 0.14), leading to an increased risk of DN of 1.70 [1.31–2.21] ($p < 10\text{-}4$) under a recessive model. This association was consistent in Denmark (OR=1.57 [0.98–2.52]), Finland (OR=1.95 [1.26–3.01]) and in France (OR=1.51 [0.94–2.44]). The effect of UNC13b_134327 observed in our initial multi-stage strategy was in fact the consequence of its LD with this SNP.
Together with previous studies suggesting UNC13b to be involved in hyperglycemia control, our analysis identifies UNC13b as a strong candidate gene for DN.

**149**

**Investigation of Type I Error in Linkage Analysis of Complex Qualitative Traits with Common Disease Alleles and Quantitative Covariates**

T.N. Turley-Stoulig (1), A.J.M. Sorant (2), J.E. Bailey-Wilson (2), D.M. Mandal (3)
(1) Southeastern Louisiana University, Hammond, LA, USA; (2) NHGRI/NIH, Baltimore, MD, USA; (3) LSUHSC, New Orleans, LA, USA.

Previous work has shown that when a good estimate of the heredity model was available, model-dependent lod-score methods (in LODLINK) provided the greatest power when analyzing a qualitative trait where disease etiology involved a quantitative covariate. With an uncertain heredity model, sib-pair analysis with the squared sib-pair trait difference as the dependent variable was a good alternative. However, NPL scores and Kong and Cox LOD analyses with MERLIN and ALLEGRO do not allow for covariate effects and lose power in the presence of strong environmental effects. In the current simulation study, both model based (LODLINK with/without covariates) and model free (SIBPAL with /without covariates, MERLIN, ALLEGRO) methods of linkage analysis were compared for the effect of covariate inclusion on the Type I error rates of linkage analysis of a qualitative trait with common disease alleles. As is the case with rare disease alleles, Type I error rates were not inflated and virtually unaffected with covariate inclusion when using linkage analysis methods where covariate inclusion was possible. For methods that do not allow covariate inclusion, Type I error was marginally greater than expected only when using the exponential Kong and Cox LOD methods provided in ALLEGRO (in a few instances approaching 0.003 at the 0.001 nominal p-value and 0.11 at the 0.05 nominal p-value).

**150**

**Parent of origin effect in multiple sclerosis in a genetic isolate in the Netherlands**

Cornelia M. van Duijn, Ilse A. Hoppenbrouwers, Fan Liu, Yurii.S. Aulchenko, George C Ebers, Ben A Oostra, Rogier Q Hintzen
Departement of Epidemiology & Biostatistics, Clinical Genetics and Neurology, ErasmusMC, Rotterdam, the Netherlands; Wellcome Trust Centre for Human Genetics and Department of Clinical Neurology, University of Oxford, Oxford, UK

Multiple sclerosis (MS) is a complex disease, resulting from genetic as well as environmental factors. Parent of origin effects may influence the risk for MS.
This study is part of a larger research program named Genetic esearch In Isolated Populations(GRIP), in the South West of the Netherlands. Twenty-four MS patients from this population could be linked to the most recent common ancestor in 14 generations. We computed and compared the average kinship of the parents of the 24 MS patients. We further explored genealogic links between MS patients via all of their common ancestors. Most often, multiple connections exist between two patients, only the shortest ones were used. We compared the resultant distribution with the expected distribution under no parent of origin effect (25% maternal-maternal, 25% paternal-paternal, and 50% maternal-paternal). Among a total of 814 shortest connections, 333 (41%) were maternal-maternal, 98 (12%) paternal-paternal, and 383 (47%) were maternal-paternal (HWE test, $p < 0.001$). Mean kinship among mothers ($3.05 \times 10^{-3}$) was 3.8 times higher than the one among fathers ($0.81 \times 10^{-3}$), (t-test, $p < 0.001$). For the parents of patients with Alzheimer's disease and ADHD, no such differences were seen. The significant excess in maternal relationship suggest a maternal effect in MS occurrence.

**151**

**A simple approach for assessing the strength of evidence for association at the level of the whole gene**

Anna E. Vine (1), David Curtis (1), Jo Knight (2)
(1) Centre for Psychiatry, Queen Mary's School of Medicine and Dentistry, London E1 1BB, UK
(2) Social Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK

**Introduction** It is expected that different markers may show different patterns of association with different pathogenic variants within a given gene. It would be helpful to combine the evidence implicating association at the level of the whole gene rather than just for individual markers or haplotypes. Doing this is complicated by the fact that different markers do not represent independent sources of information. **Method** We propose combining the *p* values from single locus and/or multilocus analyses of different markers

according to the formula of Fisher, $X=\Sigma(-2\ln(p_i))$, and then assessing the empirical significance of this statistic using permutation testing. We present an example application to 19 markers around the HTRA2 gene in a case-control study of Parkinson's disease. **Results** Applying our approach shows that, although some individual markers produce low *p* values, overall association at the level of the gene is not supported. **Discussion** Approaches such as this could be useful in assimilating the overall evidence supporting involvement of a gene in a particular disease. Information can be combined from biallelic and multiallelic markers and from single markers along with multimarker analyses. Single genes can be tested or results from groups of genes involved in the same pathway can be combined in order to test biologically relevant hypotheses.

**152**

**Two-Stage Strategies for Detecting Gene-Environment Interaction in a Genome Wide Association Study**

HE Volk(1), JP Lewinger(1), DC Thomas(1)
(1)Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, USA

We evaluate the performance of combinations of case-only and case-control study designs in the context of a two-stage genome-wide association study (GWAS) with the goal of identifying gene-environment interaction (GxE). The case-only design has demonstrated increased power to detect GxE in the absence of gene-environment association, but if used as a screening test in the first stage of a GWAS, could lead to many more false positives. Standard case-control tests are unbiased in this case, but less powerful in the absence of gene-environment association. We evaluate the effect of gene-environment association and the proportion of markers carried from stage I to stage II on combinations of case-only and case-control designs in a two-stage GWAS. The power of a GWAS to identify a true interacting marker at a fixed type I error rate (controlled by estimating the marginal null distribution of the interaction test across the empirical distribution of G-E associations) is simulated. We show that the most efficient design combination is strongly related to the underlying amount of gene-environment association, that the correct type I error can be preserved, and that power equivalent to a single stage design can be achieved at a reduced genotyping cost.

**153**

**Unbiased and Efficient Estimation of the Effect of Candidate Genes on Quantitative Traits in the Presence of Population Admixture**

Y. Wang(1), Q. Yang(2), D. Rabinowitz(3)
(1)Dept. of Biostat, Columbia Univ., USA, (2)Dept. of Biostat, Boston Univ., USA, (3)Dept. of Stat, Columbia Univ., USA

Population admixture and stratification can be a confounding factor in gene-association design. Family-based methods have been proposed in both testing and estimation settings to adjust for this confounding, especially in case-only association studies. Family-based methods rely on conditioning on the observed parental genotypes or on the minimal sufficient statistic for the genetic models. In some cases these methods do not capture all the available information; the conditioning strategy is too stringent. General efficient methods to adjust for population admixture that use all the available information have been proposed. However these general approaches are not easy to implement in practical situations. A previously developed easy-to-compute approach adjusts for admixture by adding additional covariates to linear models. Here is shown that this method can be combined with general efficient methods to provide computationally friendly and efficient adjustment. After deriving efficient covariates, the adjusted analysis can be carried out by standard statistical packages. The approach is illustrated through an analysis of the influence of apolipoprotein E (APOE) genotype on plasma low density lipoprotein (LDL) concentration. The results provide evidence that non-trivial gains in efficiency may be obtained by using information not accessible to methods that rely on conditioning on the minimal sufficient statistics.

**154**

**Haplotype-based Case-Control Association Studies with Related Individuals: A Quasi-Likelihood Score Test Approach**

Z. Wang and M. S. McPeek
Dept. of Stat, Univ. of Chicago, USA

Haplotype-based association analysis has been widely used in case-control studies, with haplotypes potentially providing more information on untyped variants as well as on interactions among tightly-linked typed variants. Most previous methods focus on specific family-based designs or on unrelated case-control designs. We propose a method for very general study designs which is an extension of the single-marker quasi-likelihood score test approaches of Bourgain et al (Am. J. Hum. Genet. 73:612–626, 2003) and Thornton and McPeek (Am. J. Hum. Genet., 2007, to appear). Our method allows one to test for association of a binary trait with haplotypes based on multiple tightly-linked markers when only genotype data are available, and it can be used in samples containing arbitrary combinations of related and unrelated individuals with relationships specified by known pedigrees. Such samples commonly arise when families sampled for a linkage study are included in an association study. Furthermore, power to detect association to complex traits can be increased when affected individuals with affected relatives are sampled, because they are more likely to carry disease-associated haplotypes than are randomly sampled affecteds. Our proposed method obtains high power by explicitly taking advantage of this phenomenon. Parental genotype data is incorporated, when available. Our method provides global tests for association, as well as haplotype-specific tests. We perform simulation studies to compare the power of our method to that of Browning et al (Genet. Epidemiol. 28:110–122, 2005).

**155**

**The role of potential obesity loci in a population at high risk of cardiovascular disease**

J. Wheeler (1), C. Cluett (2), D. King (3), S. John (1), D McHale (1), G. Johnston (1).
(1)Pfizer Inc, UK (2)Pfizer Inc, US

Obesity is a common trait which predisposes subjects to several diseases including cardiovascular disease. Multiple genetic loci (including 127 candidate genes) have been associated with obesity in family and population-based association studies. The aim of this study was to assess the genetic contribution to obesity in a population at high risk of cardiovascular disease.
The study population was drawn from the ACCESS phase III cardiovascular trial. 857 randomly selected subjects were genotyped for 611 SNPs on 34 genes previously associated with obesity. Data were analysed as linear regression models of log bmi adjusting for demographic characteristics, smoking and lipid levels.
The analysis was restricted to the 89% of subjects with Caucasian ancestry, 68% were male, mean age 68 years and 28% clinically obese. Quality control of genotyping was used to exclude poorly performing SNPs from the analysis using call rate and deviation from Hardy Weinberg Equilibrium. Of the 34 genes selected, 18 were found to be significantly associated with BMI (uncorrected p-value $p < 0.05$). Multiple SNPs were associated in the glucocorticoid receptor (p=0.0001 to 0.033), 11-b-hydroxysteroid dehydrogenase (p=0.0005 to p=0.041), b2-adrenergic receptor (p=0.016 to p=0.043), Vitamin D3 receptor (p=0.0003 to p=0.026) and Neuropeptide Y receptor type 2 (p=0.016 to 0.025). These data provide additional support for obesity genes and further analysis will assess the contribution of these genes in this population at high risk of cardiovascular disease compared to published data.

**156**

**Analysis of multiple SNPs in a candidate gene or region**

J.C. Whittaker, J. Chapman
London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

We consider the analysis of multiple SNPs within a gene or region. The simplest analysis of such data is based on a series of single SNP hypothesis tests, followed by correction for multiple testing, but it is intuitively plausible that a joint analysis of the SNPs will have higher power. However, standard tests, such as a likelihood ratio test based on an restricted alternative hypothesis, tend to have large numbers of degrees of freedom and hence low power. This has motivated a number of alternative test statistics.
Here we compare several of the competing methods, including the multivariate score test (Hotelling's test) of Chapman et al (2003), Fisher's method for combining p-values, the minimum p-value approach, a Fourier transform based approach recently suggested by Wang and Elston (2007) and a Bayesian score statistic due to Goeman et al. (2005). Some relationships between these methods are pointed out, and simulation results given to show that the minimum p-value and the Goeman et al (2005) approaches work well over a range of scenarios. The Wang and Elston approach often performs particularly poorly, and we explain why.

**157**

**A block-based SNP selection approach for population association studies**

S Won (1), Robert C. Elston(1)
(1) Dept. of Epidemiology and Biostatistics, Case Western Reserve Univ., USA

Approaches for detecting disease-SNP associations can in general be direct studies of candidate, potentially causal, polymorphisms, or indirect studies of neutral markers, with the aim of detecting causal variants via linkage disequilibrium. Thus, in direct approaches SNPs in a gene that affects the trait of interest are genotyped, whereas in indirect approaches SNPs in the gene that affect the trait are not genotyped The analytical power of indirect approaches depends on both the strength of linkage disequilibrium (LD) between markers and the causal gene and the disease model, i.e. disease mode of inheritance. In addition, it is known that alleles at SNPs near a disease locus have a tendency to be present in different frequencies in cases and controls, which provides a theoretical basis for association methods. However, in reality we find that, when a causal SNP is simulated on to a Hapmap dataset, there are non-informative SNPs even very near the causal SNP. As a result, two things can be concluded. First, it is preferable to consider association between a window of SNPs and a disease instead of between each individual SNP and the disease. Second, given a window of SNPs, power is improved if only informative SNPs are selected for analysis. Thus, we suggest a block-based SNP selection approach, with adjustment of standard errors to correct for the diminution of the standard error that arises as a result of model selection. Several disease models are simulated on to Hapmap marker data to quantify this improvement in power.

**158**

**Comparing affected-relative allele-sharing models in a small number of moderate-sized pedigrees**

Chao Xing
Department of Clinical Sciences and McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390‑8591

For some complex but less common diseases, doctors can usually only collect through probands a small number of moderate-sized pedigrees with multiple affected members. Linkage screen for allele sharing identical by descent (IBD) among affected relatives is routinely the first step toward identifying the disease-disposing genes. In this study we compared the power of three affected-relative allele-sharing models including the non-parametric linkage (NPL) method, the so-called Kong and Cox linear model, and the exponential model in a small number of moderate-sized pedigrees. The

NPL method gives liberal p-values whereas the Kong and Cox linear model provides conservative p-values, which is different from their usual behavior that the former is more conservative while the later is more appropriate in case of incomplete inheritance information. The Kong and Cox exponential model gives proper significance levels. In summary, the exponential model should be advanced in case of a small number of moderate-sized pedigrees with multiple affected members. Moreover, the "odd" phenomenon of liberal NPL scores but conservative Kong and Cox linear scores at the same region may indicate excess allele sharing IBD among the affected.

**159**

**Linkage analysis may not be effective for detecting a modifier locus: example using melanoma, CDKN2A and MC1R**

XR Yang, MA Tucker, AM Goldstein
Genet Epidemiol Br, DCEG/NCI/NIH/DHHS, Bethesda, MD, USA

For many complex diseases, identifying a major susceptibility gene is only the first step to explaining the disease phenotype. We often want to detect other genes that may influence risk; these genes may be classified as low-risk or modifier genes (if they modify the penetrance associated with the major gene). We examined whether a linkage strategy in families with mutations in a major gene would detect a modifier gene. We used data from a genomewide scan of cutaneous melanoma (CM) to determine whether we could observe evidence for linkage to a modifier gene (MC1R) among families with mutations in the major high-risk CM-susceptibility gene (CDKN2A). MC1R also influences CM risk; it is classified as a low risk/modifier gene since MC1R variants modify penetrance in CDKN2A-mutated families. We used data from a genomewide scan of 1000 microsatellite markers (avg density 4cM) in 10 large CM families with CDKN2A mutations (69 CM patients; 117 mutation carriers). Maximum 2-pt lod scores near CDKN2A (9p21) ranged from 6.5–7.1. In contrast, using a dominant model, there was no overall evidence for linkage to MC1R (16q24). One family (11 CM cases), however, had $Z>2.2$ for the markers closest to MC1R. We plan to conduct additional analyses using the MC1R variants directly to determine the maximum Z that could be expected from this locus. Since MC1R variants may increase CM risk additively, we will also evaluate other inheritance models. Although MC1R is a known modifier gene for CDKN2A, a linkage analysis strategy may not be effective for detecting evidence for this type of locus.

**160**

**Estimating heritabilities and genetic correlations of insulin resistance and related metabolic traits in Indian families using a multivariate maximum likelihood approach**

D. Zabaneh (1), J.C. Chambers (1), P. Elliott(1), R. Baliga (2), J. Scott(1), D.J. Balding(1) and J.S. Kooner(1)
(1) Imperial College London, UK,(2)Davis Heart & Lung Institute, Ohio, USA

Insulin resistance (IR) and related metabolic conditions have a higher prevalence among Indian Asians (IA) living in the UK compared to Northern Europeans (NE). The heritabilities of IR related traits have been studied extensively in NE populations, but none have been published recently for IA. Most studies utilised univariate and bivariate methods in their analyses, in this study we use a multivariate restricted maximum likelihood procedure which estimates variance components of many traits simultaneously. This allows for the estimation of heritabilities in addition to genetic correlations arising mainly from pleitropy. Models with an additive genetic and a shared common environmental effect were used, comparison between models was carried out using LRT. Two statistical packages were employed: SOLAR for the univariate analysis and VCE for the multivariate one. Phenotypic data was available for 1518 individuals from 181 IA families living in London and ascertained on Coronary Heart Disease (CHD) status. Our results show that genes contribute to a significant proportion of the total variance in IR related traits in IA families with a high risk of CHD. Heritability estimates were similar from both packages and are within the published range from studies on other populations. However, we believe that estimates from the multivariate approach are more accurate as correlations between all traits are taken into account.

**161**

**Genome-wide Association and Replication Analyses in UK Subjects Reveal Multiple Novel Type 2 Diabetes Susceptibility Loci**

E. Zeggini(1), M. Weedon(2), N. Timpson(1), T. Frayling(2), K. Elliott(1), W. Rayner(1), C. Lindgren(1), UK T2D Genetics Consortium, Wellcome Trust Case Control Consortium, A. Hattersley(2), M. McCarthy(1).
(1)Oxford, UK, (2)Exeter, UK

The capacity to undertake genome-wide association (GWA) scans and to follow up emerging signals in large-scale replication datasets can deliver insights into the etiology of type 2 diabetes (T2D). Starting from GWA data (on 393453 SNPs passing quality control) for 1924 T2D cases and 2938 controls generated by the Wellcome Trust Case Control Consortium, we set out to detect replicated T2D signals in 3757 additional cases and 5346 controls, and by integration of our findings with data from other international consortia (DGI and FUSION, total sample n>32000). This work has identified novel T2D loci in and around *CDKAL1* (rs10946398, all UK data OR 1.16[1.10-1.22], additive $P=1.3\times10^{-8}$; combined UK, DGI and FUSION data OR 1.13[1.09-1.17], $P=1.4\times10^{-12}$), *CDKN2A/CDKN2B* (rs10811661, all UK OR 1.19[1.11-1.28], $P=4.9\times10^{-7}$; combined OR 1.20[1.15-1.25], $P=2.2\times10^{-15}$) and *IGF2BP2* (rs4402960, all UK OR 1.11[1.05-1.16], $P=1.6\times10^{-4}$; combined OR 1.14[1.11-1.18], $P=8.6\times10^{-16}$) and confirmed the associations at *HHEX/IDE* (rs5015480, all UK OR 1.13[1.07-1.19], $P=4.6\times10^{-6}$; combined OR 1.13[1.08-1.17], $P=5.7\times10^{-10}$) and *SLC30A8* (rs13266634, all UK OR 1.12[1.04-1.19], $P=1.2\times10^{-3}$; combined OR 1.12[1.07-1.17], $P=3.5\times10^{-7}$). Our findings indicate the contribution of multiple modest-effect variants to the genetic architecture of T2D. We are

currently undertaking further replication rounds based on genome-wide imputation data meta-analyses.

**162**

**Continuous and Discrete Association Analyses of Body Mass Index and Obesity**

J.H. Zhao(1), S. Li (1), J.A. Luan(1), Q. Tan(2), E. Wheeler(3), S. Debenham(4), M. Inouye(3), P. Deloukas(3), M. Sandhu(4), I. Barroso(3), R. McGinnis(3), R. Loos(1), N.J. Wareham(1)

(1)MRC Epidemiology Unit, Cambridge, UK, (2)Odense University Hospital, Denmark, (3)The Wellcome Trust Sanger Institute, Hinxton, UK; (4)Department of Public Health & Primary Care, Institute of Public Health, University of Cambridge, UK

Body mass index (BMI, weight (kg)/height (m) 2) provides a reliable indicator of body fatness for most people and adult obesity (BMI>=30). Apart from the general concern over the loss of statistical efficiency associated with the dichotomous definition, the impact of the dichotomisation relative to BMI as a continuous trait in the study of association with genetic polymorphisms is interesting but not well characterised. We set to explore this using computer simulation, where the recently-established FTO gene enables us to use a more realistic model than has previously been reported. In addition, we have examined results from our genomewide association study of obesity involving ~3500 unrelated individuals genotyped on Affymetrix 500k genechips. This is a case-cohort design with a randomly selected sub-cohort as controls from the European Prospective Investigation in Cancer (EPIC) Norfolk study. We discuss synthesis of results from cohort data, case-only data and case-control data by regression analyses. Our findings will give insights into the problem aforementioned. We discuss related issues and give some recommendations for analysis.

**163**

**Optimal DNA Pooling-based Two-Stage Designs in Case-Control Association Studies**

Yihong Zhao and Shuang Wang
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032

Study cost remains to be the major limiting factor for genome-wide association studies due to the need of genotyping a large number of SNPs on a large number of subjects. DNA pooling strategy and two-stage designs have been proposed to reduce genotyping costs. In this study, we propose a cost-effective two-stage approach with DNA pooling strategy. In stage I, all markers are evaluated on a subset of individuals with DNA pooling. The most promising set of markers are then evaluated with individual genotyping on all individuals in stage II. The goal is to determine the optimal parameters ($\pi^p_{sample}, \pi^p_{marker}$) that minimize the cost of a two-stage DNA pooling design while maintaining a desired overall significance level and achieving a similar power of the one-stage individual genotyping design. We consider the effects of three factors on the optimal two-stage DNA pooling designs: the DNA pooling-related measurement errors, the ratio of the per-genotyping cost in stage II to that in stage I, and the DNA pool size. Our results suggest that the optimal two-stage DNA pooling designs may be much more cost-effective than the optimal two-stage individual genotyping designs (which use individual genotyping at both stages) under most scenarios considered when study powers are fixed.

**164**

**Mathematical modeling for left truncated HIV data to predict the time between primary infection and enrolment in the study**

J.J. Zhuang, R.C. Griffiths
Department of Statistics, University of Oxford, UK

Most HIV data has the characteristic of not only being right censored, as the end date of study is set and not all the patients enrolled in the study died on or before such date, but also left truncated, as patients enroll in the study at different stages of their infections. A series of mathematical models is specifically developed in order to predict the time between a patient's primary infection with HIV-1 and his/her enrolment in the study using the patient's CD4 count at enrolment. The model for predicting the time between a patient's enrolment and either his/her death or the end of the study given his/her CD4 count at enrolment is also developed in order to verify the accuracy and consistency of the models after comparisons with real data.

**165**

**Gene-environment interaction: prospects and pitfalls**

Peter Kraft
Department of Epidemiology and Biostatistics, Harvard School of Public Health

Complex diseases are by definition multifactoral, and for many there are established environmental risk factors (e.g. smoking and lung cancer, asbestos and mesothelioma). It is therefore plausible that allowing for potential genetic effect modification by environmental exposures can improve our ability to discover causal genetic variants or better understand known genetic risk factors. However, standard tests for statistical gene-environment interaction are notoriously underpowered and difficult to interpret. Choosing appropriate exposures, accounting for multiple testing, distinguishing between "external" and "internal" exposures, and adjusting for exposure measurement error can also prove quite difficult. I review how these issues affect (i) standard tests for gene-environment interaction and (ii) a joint test for gene main effect and gene-environment interaction that can be useful in the context of genome-wide association scans. I present situations where accounting for gene-environment interaction can boost power to detect a risk locus–and situations where it will not. I close with a discussion of the relevance of gene-environment interaction to biology, public health, and individualized treatment.

**166**

**Turning a flood of data into a deluge: "in silico" genotyping for genome-wide association scans**

Gonçalo Abecasis
Center for Statistical Genetics, Department of Biostatics, University of Michigan, Ann Arbor, MI 48103

With millions of single nucleotide polymorphisms (SNPs) identified and characterized, genome-wide association studies have begun to identify susceptibility genes for complex traits and diseases. These studies involve the characterization and analysis of very high-resolution SNP genotype data for hundreds or thousands of individuals. Nevertheless, and despite continuing improvements in SNP genotyping technologies, most genome-wide association studies only directly genotype a subset of all existing SNPs. I will review computationally efficient approaches for estimating unmeasured genotypes and evaluating the association between these unmeasured genotypes and relevant traits. These approaches all rely on the intuition that even apparently unrelated individuals will share stretches of chromosome that include many SNPs. Once one of these stretches has been characterized in detail in a few individuals, the alleles it contains can be imputed in other carriers, with different degrees of accuracy. I illustrate the performance of the method and its potential utility using data from ongoing genome-wide association scans, including both scans that examine samples of related individuals and scans that examine samples of apparently unrelated individuals. I also examine the performance of these approaches in different populations, using data from the Human Genome Diversity Panel.

**167**

**Can Forests Have Lotus Effects? Or: Data Mining in Genome-Wide Association Studies**

A. Ziegler, Institute for Medical Biometrics and Statistics, University of Lübeck

Genome-wide association studies (GWAs) were one "area to watch" for 2007 according to a comment in Science in fall 2006. Indeed, several GWAs with surprising results have been published in the first half of 2007. But what are the aims of GWAs? To answer this question, I will describe gene identification and risk prediction studies in the first part of the presentation.

Various data mining approaches have been proposed to perform these tasks. After a brief overview of the data mining techniques, I will put a focus on tree-based methods for yes/no predictions in the second part of the talk. I will specifically discuss feature selection approaches, computational issues and the detection of interactions. Some of the data mining approaches will be illustrated by re-analyzing data from a GWA on myocardial infarction and simulated data provided for the Genetic Analysis Workshop 15.