ABSTRACTS FROM THE

FIFTEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

St. Petersburg, Florida
November 16–17, 2006

1

Application of three approaches to select relevant SNPs in genetic association studies

A.G. Heidema(1,2,3), E.J.M. Feskens(2), J.C. van Houwelingen(1,4), P.A.F.M. Doevendans(5), E.C.M. Mariman(3) and J.M.A. Boer(1)
(1) Nat Inst for Public Health and Env (RIVM), NL, (2) Div of Human Nutr, Wageningen Univ, NL, (3) Funct Genomics, Maastricht Univ, NL, (4) Med Stat, Leiden Univ Med Center, NL, (5) Heart-Lung Centre, Univ Med Centre Utrecht, NL

Approaches have been developed to analyze large numbers of SNPs in relation to disease status. We compared the set association approach (SAA), multifactor dimensionality reduction (MDR) and random forest (RF) to select from 93 SNPs a subset of SNPs that are important in determining HDL-cholesterol levels. The study population consisted of a random sample from a Dutch monitoring project for cardiovascular disease risk factors and was dichotomized into cases (low HDL-cholesterol, n=533) and controls (high HDL-cholesterol, n=545) based on gender-specific median values. All approaches prioritized three SNPs as important (CETP Taq1B, CETP -629 C/A and LPL ser447ter). SNPs with weaker main effects or only relevant in interaction are not prioritized by all methods. MTHFR 677 C/T was selected in combination with CETP Taq1B as best model by MDR. APOC3 3175 G/C, CCR2 val62ile and NOS2A asp346asp were additionally prioritized by RF. P-values were significant for both SAA ($p=0.0007$) and RF ($p<0.03$), the best model obtained with MDR was not significant ($p<0.14$). The application of different approaches provides information whether SNPs contribute by their main and/or interaction effects. SNPs with clear main effects are selected by all methods, strengthening our confidence that these SNPs are truly involved in determining HDL-cholesterol levels.

2

Testing Association between Disease and Multiple SNPs in a Candidate Gene

W.J. Gauderman, C. Murcray, F. Gilliland, D.V. Conti
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20247

Current technology allows investigators to obtain genotypes at multiple SNPs within a candidate locus. Many approaches have been developed for using such data in a test of association with disease, ranging from simple genotype-based tests to complex haplotype-based tests. We develop a new approach that involves two basic steps. In the first step, we use principal components (PC) analysis to compute combinations of SNPs that capture the underlying correlation structure within the locus. The second step uses the principal components directly in a test of disease association. The PC approach captures linkage-disequilibrium information within a candidate region, but does not require the difficult computing implicit in a haplotype analysis. We demonstrate by simulation that the PC approach is typically as or more powerful than several other methods, including both genotype- and haplotype-based approaches. We also analyze association between respiratory symptoms in children and four SNPs in the GSTP1 locus, based on data from the Children's Health Study. We observe stronger evidence of an association using the PC approach ($p=0.044$) than using either a genotype-based ($p=0.13$) or haplotype-based ($p=0.052$) approach.

3

A powerful test of association of multiple genes with disease

I Mukhopadhyay*(1), A Thalamuthu(2), E Feingold(3), DE Weeks(3)
(1) Department of Human Genetics, Univ of Pittsburgh, USA (2) Department of Statistics, Univ of Madras, Chennai, India (3) Department of Human Genetics and Biostatistics, Univ of Pittsburgh, USA

When multiple genes might influence disease risk, it can be useful to globally test for the simultaneous effect of the multiple genes on disease risk. We propose a powerful test based on kernels for testing association of multiple markers acting simultaneously on disease, using case-control data. We used the idea of analysis of variance (ANOVA) with the scores of symmetric kerel functions on genotypes of each marker (Schaid et al., 2005) as observations. Thus the variation between cases and controls and the variation within each class led us to propose a testing procedure for the detection of association. We study each marker separately and combine them to get a global statistic that is finally used to test for disease-marker association. We carried out a simulation study to calculate the Type I error and power of our new statistic, varying

liability loci from one to five out of a total of ten markers. For a variety of relative risks and allele frequencies, our proposed statistic has much higher power than some other statistics in the literature. We studied several different kernels and it appears that no particular kernel turns out to be the best in all models; however there is very little difference in power among them. Schaid DJ, McDonnell SK, Hebbering SJ, Cunningham JM, Thibodeau SN. 2005. Nonparametrtic tests of association of multiple genes with human disease. Am J Hum Genet 76:780–793.

4

The general transcription factor FOXO1A on 13q14 is associated with obesity

I.B. Borecki(1), M.F. Feitosa(1), D. Ma(1), K.E. North(2), S. Williamson(3), J. Laramie(3), R.H. Myers(3)
(1) Div Stat Genomics, Washington Univ, St. Louis, MO; (2) Dept Epidemiol, Univ. N.Carolina, Chapel Hill, NC; (3) Dept Neurology, Boston Univ, Boston, MA

We sought to identify loci for body mass index (BMI) in the linked region on chromosome 13q14 (42 cM; LOD=3.2) in Caucasian pedigrees from the NHLBI Family Heart Study; Black subjects were not genotyped in this project phase. We tested 29 positional candidate genes located between 26 and 76 cM, selecting tag SNPs within each gene, for a total of 227 SNPs. SNPs and haplotypes were screened using a TDT implemented in TRANSMIT with obesity (BMI>30 kg/m^2). There was strong evidence of association for FOXO1A (forkhead box O1A; p=0.008), which is a key transcription factor in insulin signaling in liver and adipose tissue, and influences the expression of pancreatic beta-cell genes. DGKH (diacylglycerol kinase), with a role in glycerolipid metabolism, was associated with obesity (p=0.004); and suggestive signals were found for AKAP1 (a-kinase anchor protein 11; p=0.02) and HTR2A (5-hydroxy tryptamine receptor 2A; p=0.03). AKAP1 may influence the cell cycle with its general function of binding to the regulatory subunit of protein kinase A, while HTR2A is a receptor for serotonin, a key mediator in the control of satiety. A novel, iterative routine for exploring haplotype associations, HAPLOBUILD, revealed extended haplotypes associated with BMI in both FOX01A (P=0.025) and HTR2A (P=0.0019) under an additive model. Common variants in several genes located under the 13q14 linkage peak may have a role in obesity and adiposity.

5

Single worldwide origin for a common low-penetrance RET mutation in Hirschsprung disease (HSCR)

F Lantieri(1,2), J Amiel(3), G Antinolo(4), S Borrego(4), G Burzynski(5), I Ceccherini(1), E Emison(6), R Fernandez(4), M Garcia-Barcelo(7), P Griseri(1), R Hofstra(5), C Kashuk(6), S Lyonnet(3), P Tam(7), A Tullio-Pelet(3), K West(6), A Chakravarti(6)
(1) Inst. Gaslini, Genoa, (2) Univ.Genoa, Italy; (3) INSERM U-393, Paris, France; (4) HH UU Virgen del Rocio, Seville, Spain; (5) Univ.Groningen, The Netherlands; (6) Johns Hopkins Univ. Baltimore, MD; (7) Univ.Hong Kong, China

HSCR is a complex genetic disease characterized by colonic aganglionosis with multiple genes mutations. RET proto-oncogene coding mutations explain < 50% of cases. Analysis of 14 RET SNPs in 876 HSCR families (2,672 inds.) from six countries, showed the highest transmission from parents to affected sibs for the T allele of the intron1 I1.357 SNP (p=5.1x10-66), previously demonstrated to be an enhancer mutation. Its frequency is 59% on transmitted (T) and 24% on untransmitted (U) alleles, with differences among populations consistent with HSCR prevalence. Haplotype reconstruction, LD and recombination analyses were carried out to prove identity by descent of the variant. The two most frequently transmitted haplotypes in Caucasians, long and short, include the I1.357 T allele and share a common allelic combination extending from 5'UTR to exon 5. In contrast, in Chinese the long haplotype is almost exclusively present. A recombination hotspot between introns 5 and 8 was demostrated by LD analysis. We conclude that the enhancer mutation arose on the long haplotype, which rearranged after the Asian-European split to originate the short haplotype. We thus provide a first view on the genetic history that has led to the current worldwide RET allele distribution.

6

Copy number variants of drug metabolizing enzyme genes typed in the HapMap samples

F.C.L. Hyland(1), K. Li(1), C. Barbacioru(1), K. Haque(2), F.M. De La Vega(1), R.A. Welch(2), and K. Lazaruk(1)
(1) Applied Biosystems, Foster City, CA, USA, and (2) Core Genotyping Facility, Division of Cancer Epidemiology and Genetics, SAIC Frederick, National Cancer Institute, Gaithersburg, MD USA

Gene copy number variants (CNVs) have been associated with complex phenotypes, including cancer, immunological and neurological disorders, and with variations in drug response. Recent reports show that thousands of CNVs across the human genome are polymorphic, suggesting they may be more relevant to common disease than previously anticipated. Furthermore, falsely assuming diploidy can lead to inaccuracy in SNP genotyping. Finally, CNVs can interact with SNPs to influence their phenotypic impact.We have developed real-time quantitative PCR TaqMan® assays to quantify copy number in genes, including 5 important drug metabolism genes: CYP2D6, CYP2E1, CYP2A6, GSTM1 and GSTT1, in which CNVs (0 to 4 copies) are known to be associated with drug response. We developed copy number calling algorithms. We genotyped these CNVs in the HapMap samples, comprising 270 individuals from African, Asian, and European populations, including 60 trios. The CNVs analyzed were in Hardy-Weinberg equilibrium in all populations. There were statistically significant differences in CNV allele frequency between one population and one or all others for CYP2A6, CYP2D6, GSTT1, and GSTM1, but no differences for CYP2E1. These results suggest that some of the CNV alleles may have been subject to recent natural selection. In most but not all cases, the number

of copies in a child was consistent with Mendelian inheritance.

7

High density association model building and gene discovery

X. Gu(1,2) and C.I. Amos(1)
(1) Department of Epidemiology, M.D. Anderson Cancer Center and (2) Department of Biometry, U.T. Health Science Center, Houston, TX

With the advent of platforms for inexpensive genotyping using genomic high-density arrays new approaches are required to identify variants associated with disease susceptibility. At issue in the analysis are the large number of false positives that may be generated by individual SNP analysis of the data. Therefore, we have been studying the performance of various alternative approaches that can perform model building to incorporate all of the data from a genome-wide association study. Approaches that we evaluated include i) single SNP analysis using either Bonferonni correction or false-discovery to control false positive results, ii) stepwise forward selection. We have applied these methods to several simulated data sets including both a genome-wide simulation as well as a candidate gene simulation derived using the linkage disequilbrium structure of specific genes. Results show poor power for single SNP analyses. For example analysis of the complex 'Rivalside' trait from the pharmacogenetics network simulation study showed 64% power for single SNP analysis with FDR correction and 90% power using stepwise approach. At the nominal 5% level, the FDR approach showed unacceptably low type I error (0.2%) because of the LD among the markers, while the stepwise approach was only moderately conservative (2.87%). When applying these approaches to data from genetic analysis of data derived from genotyping using Affymetrix arrays, we had to develop procedures to only include those markers that show significance to allow for missing genotyping.

8

Using combined datasets to find true associations in genome wide association studies

J.C. Barrett, K.T. Zondervan, L.R. Cardon
Wellcome Trust Centre for Human Genetics, Oxford University, UK

Many genome wide association (GWA) studies involving hundreds of thousands of SNPs genotyped in thousands of samples are currently underway or being planned. Because only a few commercial marker sets are available on this scale, many such studies will generate genotype data on an identical set of markers. In addition, a number of studies have turned to the use of universal or 'common' controls, where one control group is compared with multiple disease samples. The Wellcome Trust Case Control Consortium (WTCCC) offers a uniquely powerful resource to illuminate the potential benefits of this design. The WTCCC is a UK-based GWA study of 2000 cases from each of seven common diseases and two sets of 1500 control individuals. To date, we have completed genotyping of the Affymetrix 500k chip on more than 15,000 WTCCC samples. Comparisons of allele frequencies in unrelated diseases (e.g. Type 2 diabetes and bipolar disorder) can aid in separating true associations from spurious results due to genotyping error or population structure. We show, for example, that one previously published association observed in our data is confirmed by similar allele frequencies between other case groups and controls, whereas other apparent significant associations are more questionable because case to external case frequencies are a closer match than external case to control. We extend this idea to a more generalized method for extracting maximum information from additional cases and consider its potential future application to combining data from many large, independent GWA studies.

9

Statistical methods for the identification of trait-associated microdeletions

Y. Li(1), C.I. Amos(2), M.S. McPeek(1)
(1) Dept. of Statistics, University of Chicago, USA, (2) Dept. of Epidemiology, University of Texas MD Anderson Cancer Ctr, USA

Microdeletion (the removal of a chromosomal segment of size up to several megabases) can play an important role in some genetic diseases. Identifying regions of microdeletion related to a genetic disease could lead one to the causal gene(s) located within the micro deletion. With high density genotype data from cases and controls, microdeletion associated with a trait would be observed as excess homozygosity associated with the trait. We develop statistical methods to detect such trait-associated homozygosity. We propose a hidden Markov model which allows for long homozygous segments that are present at background levels in the general population and for heterogeneity (i.e. some proportion of the case haplotypes do not carry deletion). This framework leads to efficient algorithms for likelihood calculation and maximization, which allow us to perform likelihood-based inference (hypothesis testing and confidence interval construction) to detect and localize microdeletions associated with a trait. We also develop fast permutation tests that allow us to detect microdeletions by combining information from all available markers. We demonstrate the power of our methods using simulation based on a dense map of markers on chromosome 18 (2300 SNPs in a 10 cM region). Supported by HG001645 and AR44422.

10

Genome-wide association (GWA) scan of type 2 diabetes (T2D) in the Old Order Amish

E. Rampersaud, C.M.Damcott, B.D. Mitchell, J. Shelton, J. Ying, M. Fu, J. Shi,Y. Zhao, S. Ott, J. O'Connell, A.R. Shuldiner
Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, USA

We conducted a GWA scan of T2D in the Amish, a genetically well-defined closed Caucasian founder population. We genotyped 116,204 single nucleotide polymorphisms (SNPs) using the Affymetrix GeneChip Mapping 100K Array set in 124 T2D cases and 295 normal glucose tolerant (NGT) controls. The average call rate was 95% based on the BRLMM genotype-calling algorithm (range: 90–99%). Following quality-control checks, 102,079 informative SNPs were included in our analyses. We selected an initial set of T2D-associated SNPs ($p < 0.0002$; n=212) using age-adjusted logistic regression analysis. To properly account for family structure, we then re-analyzed these SNPs using a variance component approach. Our most significant results (lowest p-value=1.07e-5) were identified in a cluster of SNPs spanning 46 kb on chromosome 7p12.2. These SNPs are in moderate LD (r2: 0.30–0.78). Significant results (p-value=1.7e-5 to 0.0003) were also found on chromosomes 4q21.21, 10q21.3-22.1, 12q23.1 and 17p11.2. Efforts are currently underway to replicate these associations in other populations and to identify the genes and their functional polymorphisms within replicated regions.

11

Accounting for Epistasis in Linkage Analysis of General Pedigrees

Y.J. Sung(1), E.A. Thompson(2), E.M. Wijsman(1,3)
(1) Div. of Medical Genetics, Dept. of Medicine, (2) Dept. of Statistics, (3) Dept. of Biostatistics, Univ. of Washington, USA

Complex traits are influenced by multiple loci, and identification of such loci is more difficult for epistatic than for additive loci. We describe a new program lm_twoqtl in the package MORGAN. lm_twoqtl performs parametric linkage analysis with a model that can include two epistatic quantitative trait loci (QTLs) and a polygenic component to model additional familial correlation from other sources. lm_twoqtl provides computationally tractable analysis of general pedigrees with many markers for these complex trait models. To illustrate the advantages of using complex trait models, in particular a model that properly accounts for epistasis, we simulated data sets on extended and nuclear pedigrees using an epistatic two-QTL model, derived from the GAW10 simulated epistatic model, and analyzed them with 5 different analysis models. Analyses with more complex models provided higher lod scores than did analyses with simpler models: the lod scores were highest with an epistatic two-QTL model (max lod=32.2) and lowest with a single-QTL model (max lod=11.6). In particular, accounting for epistasis provided higher lod scores and better localization than did analysis without epistasis (max lod=20.9). The difference between maximum lod scores for epistatic and non-epistatic models was greater in extended (diff=11.3) than in nuclear pedigrees (diff=3.2), even though the nuclear pedigree data contained more individuals (3600 vs 1600). When epistasis was properly accounted for, localization of both QTLs was considerably improved.

12

Bayesian combination of the Case-Only and Case-Control analysis: A powerful and robust test for gene-environment interaction

D. Li(1), D.V. Conti(2)
Dept. of Preventive Medicine, University of Southern California, Los Angeles, CA. USA

Gene-environment interaction may play an important role in the genetic basis of complex diseases. The traditional ways to detect the interaction, the likelihood-ratio test on the interaction term or the heterogeneity test in the case-control design, suffer from low power. The case-only analysis, which is much more powerful than the case-control analysis, has been proposed. However, the case-only analysis is based on the assumption that the two interacting factors are absolutely independent in the parent population, and it can not estimate the main effects of the interacting factors, both of which have limited its application. Here we propose a Bayesian combination of the Case-only and Case-control analysis (BCCC) approach, in which the result of the case-only analysis is set to be the prior for the case-control analysis and the cutoff for final inference is set through permutation. Power of our approach to detect the interaction is compared with several other methods across a variety of scenarios. We demonstrate that the BCCC approach is more powerful than traditional approaches and almost as powerful as the case-only analysis when the risk factors are independent in the parent population. Furthermore, when the risk factors are dependent, the false-positive rate of the BCCC test is similar with the case-control analysis. Additionally, the main effects and the interaction effect can be estimated when the BCCC test is integrated in the traditional case-control analysis. The implication of our approach for the multi-locus model is discussed.

13

Inference for candidate gene and environmental effects from combined family and case control data

R.P. Pfeiffer
Biostatistics Branch, DCEG, National Cancer Institute Bethesda, MD, 20892-7244, USA

Genetic association studies may take one of two approaches, either family-based, often using families with multiple affected members, or population based, in the form of case-control studies. Sometimes data on both types of studies are available, and combining the studies seems a sensible approach to improve power to detect genetic associations for rare genetic variants. I propose two different approaches to combine data from a case-control study and a family study that collected families with multiple cases. In the first approach, the family is viewed as the sampling unit. The joint likelihood for the members of a family is specified using a two-level mixed effects model to account for varying genetic correlations among family members. The ascertainment of the families is accommodated by conditioning on the number of cases in a family. The individuals in the case-control study are treated as families of size one, and their unconditional

likelihood is combined with the conditional likelihood for the families. This approach yields subject specific maximum likelihood estimates of genetic and environmental effects in the setting of a rare disease. In the second approach, I view an individual as the sampling unit and, using two-phase sampling techniques, estimate marginal covariate effects in the presence of a "family history" parameter that captures residual effects in the data. Data from a case-control and a family study from North-Eastern Italy on cutaneous melanoma and a low-risk melanoma-susceptibility gene, MC1R, are used to illustrate both approaches.

14

Detection of genes responsible for disease sub-types: taking advantage of symptom patterns in families

Bureau(1,2), A. Labbe(1,3), C. Mérette(1,4)
(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 phenotypes related to the diagnosis, such as detailed symptoms or endophenotypes. One hypothesis is that more homogeneous disease sub-types influenced by a smaller number of genes may be detectable from differences in patterns of these phenotypic measurements. We are developing a comprehensive statistical approach to detect disease susceptibility loci in this context using linkage analysis. Its major components are 1) modeling the multivariate symptoms of subjects in families as a function of latent homogeneous disease classes; 2) validation of genetic homogeneity of latent classes based on the heritability of these classes and 3) integration of heritable latent classes in linkage analysis. We first extended latent class analysis to allow dependence between the latent disease class status of relatives within nuclear families. An EM algorithm maximizes the likelihood and a cross-validation approach selects the optimal model. Latent classes are then used to define the phenotype in linkage analysis. We have simulated datasets where two genes can cause the same clinical diagnosis but each produces distinct patterns of symptoms. We present a comparison of our method to linkage analysis under heterogeneity and identity-by-descent sharing methods applied to the clinical diagnosis on the basis of their ability to detect the disease genes.

15

MCMC provides practical approaches for genome scans on general pedigrees with many multiallelic or dense diallelic markers

E.M. Wijsman(1,2), J.H. Rothstein(2), EA Thompson(3)
(1) Medical Genetics, (2) Biostatistics, (3) Statistics, Univ. of Washington

Linkage analysis needs to adapt to dense SNPs and multipoint analysis with either SNPs or multiallelic STRs. While exact computation is available for small pedigrees, exact computation for large pedigrees remains intractable. Markov chain Monte Carlo (MCMC) provides the only computationally practical option, but no systematic comparison of performance of programs is available. We used simulation to evaluate performance of two MCMC-based programs: lm_markers (LMM) from the ORGAN package and SimWalk2 (SW). 100 replicates of each condition were simulated on pedigrees of 14, 52, or 98 individuals with up to 4 generations of missing data. Exact computation or a surrogate was available for comparison. Both programs were fast and accurate for multipoint analysis with STRs. The 52-member pedigree (PED52) required 2-2.5 CPU min. for 3-marker analysis with LMM or SW; the median discrepancy in MCMC-based LOD scores was $<10\%$ of the exact LOD score, and for 10 markers, computation times increased to only 5.7-14 CPU min. For large numbers of dense SNPs only LMM was able to provide accurate results in computationally practical time. For PED52 and analysis of 67 dense SNPs, LMM required $\sim$11 CPU min/pedigree while SW required $\sim$11 CPU hrs/pedigree; median discrepancy compared to exact results were 19% and 62% for LMM and SW, respectively. The smallest pedigree gave similar results. Thus the MORGAN package is the first computationally practical option for linkage analyses with both large numbers of SNPs and large pedigrees.

16

A new algorithm to split complex pedigrees for linkage analysis

C.M. van Duijn, F. Liu, Y.S. Aulchenko, A. Arias Vásquez, B.A. Oostra
Genetic Epidemiologic Unit, Department of Epidemiology & Biostatistics and Clinical Genetics, Erasmus U

Genetically isolated populations have proven to be a powerful setting for the localization and isolation of genes. Not only association studies but also genome screens using linkage have been conducted successfully in these populations for rare disorders. Also for complex traits, large pedigrees can be constructed often spanning 10-15 generations. However, computational problems occur when conducting linkage studies of such complex pedigrees, since patients are often related through multiple lines and phenotypic information is lacking for most ancestors. We have developed a new method to cut large pedigrees into sub-pedigrees. For cutting large genealogies, we developed a recursive algorithm based on 3 criteria: (1) each sub-pedigree includes 2-6 patients, (2) a patient can be assigned to only one sub-pedigree, (3) the kinship of a sub-pedigree is maximized. We applied our algorithm to a pedigree including 103 Alzheimer patients from the Genetic Research in Isolated Populations (GRIP) program, which is based in the Southwest of the Netherlands. The pedigree spanning 17 generations was cut into 35 sub-pedigrees with on average 3 patients. We conducted 100 genome-wide simulations under no linkage, which showed that a LOD score of 3.63 in these sub-pedigrees corresponded to a genome wide type 1 error of 0.05. Two regions passed this threshold in the AD genome screen using 420 micro-satellite markers, one known region on chromosome 1 and one new region on

chromosome 3. These findings show that we have developed a powerful algorithm for splitting complex extended pedigrees.

17

Design and Analysis of Genome Wide Association Studies: Application to Type 2 Diabetes

M. Boehnke, A. Skol, L.J. Scott, and G.R. Abecasis for the CIDR and FUSION study investigators
Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA

The catalog of common human variants provided by the International HapMap project and the precipitous drop in genotyping costs together have made genome wide association studies a practical approach to studying the genetic basis of complex human diseases. Such studies are now in progress, often requiring genotyping hundreds of thousands of genetic markers on hundreds or thousands of subjects. In this talk, I discuss optimal experimental design and analysis approaches for two-stage genome wide association studies in which a subset of a sample is genotyped on all genetic markers in stage 1, and the remaining samples are genotyped on the most interesting markers in stage 2. Consistent with prior work of Satagopan and Elston, we find that such two-stage designs can maintain nearly the same power to detect association as the corresponding one-stage design in which all samples are genotyped for all markers. We argue that joint analysis of stage 1 and 2 samples is more powerful than replication-based analysis, despite the much larger number of tests implied. We address the impact of design parameters (proportion of sample in stage 1 and proportion of markers followed up in stage 2) and study setting (per genotype cost differences and etiologic heterogeneity between in stages 1 and 2) on optimal design and analysis approach. I illustrate these ideas with results from stage 1 of a genome wide association study of type 2 diabetes.

18

Whole-genome multilocus association mapping using localized haplotype clusters

S.R. Browning and B.L. Browning
Dept. of Statistics, The University of Auckland, New Zealand

We demonstrate the power and computational efficiency of a new method for multilocus association analysis that uses variable-length Markov chains to define localized haplotype clusters. The method automatically adapts to the levels of linkage disequilibrium in the data, so that the choice of haplotype clusters balances degrees of freedom with multilocus information extraction. Our implementation of this method can be applied to whole genome association scans with hundreds of thousands of markers and thousands of individuals, and uses permutation testing to correct for multiple tests. We show that the haplotype clusters provide better power than single marker tests when the causal variant has low frequency, while single marker tests have better power for common causal variants. Thus we suggest that both single marker and haplotype cluster tests be performed, with permutation to account for multiple testing, in order to obtain good power whatever the frequency of the causal variant. Single marker allelic and haplotype cluster analysis of 750,000 markers on 6000 case haplotypes and 6000 control haplotypes with 1000 permutations takes approximately 2 days on a desktop Windows single-processor computer with 2GB RAM, and less than 24 hours on a quad-processor Linux server with 5GB RAM.

19

Genome-wide association mapping under the Malecot model and composite likelihood

W. Zhang(1), W. Jia(2), C. Hu(2), N. Maniatis(1), A. Collins(1), N.E. Morton(1)
(1) Human Genetics Div., Univ. of Southampton, Southampton, UK, (2) Shanghai Diabetes Institute, Shanghai Jiaotong Univ. Affiliated Sixth People's Hospital, Shanghai, China

The evolutionary Malecot model and composite likelihood (CL) have been successfully applied to candidate gene association analysis. To extend the method to genome-wide scans, chromosomes were divided into regions based on linkage disequilibrium (LD) maps in LD units (LDUs). A minimum of 10 LDUs and 30 SNPs were assumed for a region individually analyzed. The expected association was predicted by the Malecot model, which is a function of the distance between the disease causing variant and the marker. CL combining associations at multiple loci was maximized to estimate the location of the disease variant. Statistical tests for association were through contrasting hierarchical models. Significance levels were assigned empirically through simulation under the null hypothesis with no association. Starting with a case-control sample, we simulated the case/control status based on genotypes of a randomly chosen SNP taken as causal for a region. Results showed that on average, the estimated point locations for the causal SNPs were only 23 to 30 kb apart from the true locations in these data with SNP density of one per 7.5 kb. Both power and location accuracy depended on the LD between the causal SNP and the nearest surrounding markers, as well as the similarity of minor allele frequencies between them. Also, LD maps provided higher power and location accuracy than kb maps. Our method is both practical and efficient for genome-wide LD scans.

20

Efficient p-value estimation in massively multiple testing problems

X.F. Shi(1,2), R. Kustra(1), D.J. Murdoch(3), C. Greenwood(1,2) and J. Rangrej(2)
(1) Dept. Of Publich Health Science, University of Toronto, Canada, (2) Hospital for Sick Children, Toronto, Canada, (3) Dept. Of Statistics & Actuarial Sciences, University of Western Ontario, Canada

We focus our attention on developing a computationally feasible algorithm for obtaining estimates of statistical

significance for millions of test statistics. This algorithm includes a method for quickly obtaining approximate estimates of the p-values associated with the test statistics, using a Random Forest model, a Bayesian scheme for deciding which interactions are of interest, and where permutations could be effectively used to improve the p-value estimates. This algorithm is illustrated and evaluated in an analysis of interactions between haplotypes of SNP markers in candidate genes, in a case-control study of colorectal cancer (the ARCTIC study), which includes more than 400,000 haplotype pairs, and approximately 1200 colorectal cancer cases and 1200 controls from Ontario, Canada. This algorithm is applicable to different test statistics where massive numbers of tests are being performed yet asymptotic significance levels are not appropriate and permutation is needed to estimate significance.

21

Potential of a Sequential Replication Filter to Detect Disease Associated SNPs

M.D. Ritchie, A.A. Motsinger, X. Liang, S.M. Dudek, J.L. Haines
Vanderbilt University, Center for Human Genetics Research, Nashville, TN 37232

Whole genome association (WGA) using dense SNP sets has rapidly become the norm rather than the exception for studies of complex disease. The current challenge is not in the generation of these data, but in the analysis and interpretation of the results. One major concern of an approach, such as a chi-square test of association, is the large number of false positive results. We propose a solution that identifies associated SNPs while minimizing the number of false positive results: the SEquential Replication Filter (SERF). SERF is based on the premise that a result that replicates in one or more independent datasets is more interesting than a result that is generated in only one dataset. SERF merges the goals of cross-validation and sequential multiple decision procedures by allowing the number and size of the replication datasets to be determined dynamically from the overall dataset. Intuitively, SNPs with strong associations will replicate more often than those with weaker associations. To test this approach, we simulated 1000 datasets with 500 cases and 500 controls and ten different single gene models with an odds ratio between 1.6 and 2.5. We applied SERF to chi-square tests of association, although it can be applied to any number of analytical methods. SERF replicates the functional SNP three or more times for all models simulated ($p < 0.05$ based on permutation testing). More importantly, SNPs with nominal significance in the overall dataset (false positives) replicate less than two times for all simulations. Thus using a replication criterion can reduce the overall false positive rate below the nominal false positive rate of 5% (for an alpha of 0.05). Thus, SERF reduces the number of SNPs that must be examined in additional association, biological, or functional experiments.

22

Harmonizing Human Genome Epidemiology Initiatives Worldwide

I. Fortier(1,2), P. Burton(2,3), A. Hansell(4), J. Little(2,5), M.J. Khoury(2,6)
(1) Soc. & Prev. Med., Univ. Montreal, Canada, (2) P3G Consortium, Canada, (3) Inst. Genetics, Univ. Leicester, UK, (4) Epi. & Pub. Hlth., Imp. Coll., UK, (5) Epi. & Comm. Med., Univ. Ottawa, Canada, (6) OGDP, CDC, USA

The contribution of genetic epidemiology to aetiological research will depend critically on the valid information that can be extracted from the numerous association-based human genome epidemiology projects being set up worldwide. Based on the design of UK Biobank, we will briefly present power calculations taking realistic account of the size of the small aetiological effects to be detected (most risk ratios < 1.5) and the true complexity of observational data in human genome epidemiology. At least 10K-20K cases of a complex disease of interest are needed to study rigorously its genetic and environmental determinants. Simulations of the real-time generation of disease events in UK Biobank will be presented to show which complex diseases may generate this many cases and over what time-frame. On a stand-alone basis, only the largest cohort and case-control initiatives currently under design will have adequate power for the commonest complex diseases. This highlights the need for harmonization across studies: both prospectively to optimize future data pooling in studies yet to begin, and retrospectively, to synthesize potentially harmonizable data that have already been collected. We will outline the current status of biobanking internationally, and describe on-going approaches to the harmonization of population-based cohort studies and case-control studies, including the closely linked initiatives P3G and HuGENet.

23

Replication Strategies for Whole Genome Association (WGA) Studies

G.M. Clarke(1), K. Carter(2), L.J. Palmer(2), A.P. Morris(1), L.R. Cardon(1)
(1) Wellcome Trust Centre for Human Genetics, Univ. of Oxford, (2) Univ. Western Australia

Whole genome association (WGA) studies have the potential to identify 1000's of possible disease loci each requiring further scrutiny for validation. The ability to replicate results in an independent study is therefore increasingly important. So far, such confirmation studies have a dismal record and, even when successful, can cite different markers, alleles and phenotypes from that found in the initial study. We examined two commonly used replication strategies analytically and via simulation. We concluded that the best strategy for replication is one in which only the variants initially identified are subsequently tested. Whilst is is theoretically possible to benefit from a replication strategy in which additional markers are identified for follow up, in practice, it requires an impossibly judicial selection of markers in regions of high inter-marker linkage disequilibrium (LD) and an extremely judicial selection of markers in regions of low

inter-marker LD. We observed the intuitively obvious fact that the power of the initial WGA is paramount in determining the probability of replication suggesting that strategies aiming to reduce genotyping costs by minimizing the genotyping in the initial WGA study are actually counter-productive. We also found that allele reversal, where one study finds an allele to be protective and another finds it to be associated with increased disease risk, is extremely unlikely to occur by chance but can arise due to population stratification, genotyping error, unmeasured interactions and other factors.

24

Estimation of haplotype frequencies from data on unrelated people

Moumita Sinha, Robert C. Elston
Department of Epidemiology and Biostatistics, Case Western Reserve University

Several methods/algorithms have been proposed in the literature for estimating haplotype frequencies. Some of the most popular methods have been the expectation-maximization(EM)maximum likelihood(ML) algorithm, and the Bayesian approach as incorporated in the software PHASE. Here we propose a novel method/algorithm, the limited linkage disequilibrium (LLD) algorithm, to estimate haplotype frequencies from case-control data. We estimate the allele frequencies of the markers and linkage disequilibrium (LD) coefficients (up to 4-locus LD, i.e. up to fourth order) from which the haplotype frequencies are calculated. Because we only consider up to fourth order LD coefficients, assuming all higher order LD coefficients are 0, our method estimates the haplotype frequencies as functions of fewer parameters and hence can handle a larger number of loci, even when the sample size is large. The LLD algorithm uses repeated quasi-maximum likelihood estimation and, when estimating the haplotype frequencies for 20 markers, the number of parameters to be estimated is only 6195 as compared to 1048575 parameters when EM-ML is used. We have tested the effect of deviation from Hardy Weinberg proportions for inbreeding coefficients of 0.03, 0.05 and 0.1 and the effect on the algorithm is minimal-whereas the effect on the EM algorithm is relatively large. Also, in the absence of such deviation, the efficiency of the LLD method is comparable to that of the EM-ML algorithm, and in fact in some cases the variance of the estimates from the EM-ML algorithm is higher than that of the estimates obtained by the LLD algorithm.

25

Testing and estimation of genotype and haplotype effects in family-based analysis of quantitative traits with missing genotype data

H.J. Cordell(1), J.R. O' Connell(2), E. Wheeler(3)
(1) Institute of Human Genetics, Newcastle University, UK, (2) School of Medicine, University of Maryland, USA, (3) The Wellcome Trust Sanger Institute, Cambridge, UK

A convenient method for testing genetic association with a quantitative trait is to perform a linear regression of trait on variables coding for the genotype or haplotypes present in an individual. Such analysis is not generally valid in the presence of population stratification, but a valid analysis can be achieved with family data by incorporation of additional stratification parameters based on the mating type of an individual's parents. Correlations between individuals from the same family (due to additional unmeasured genetic or shared environmental effects) can be accounted for by use of an empirical variance estimate clustered by family. In practice, many individuals may have missing genotype data and/or unobserved phase (haplotype) resolutions, although the possible configurations may be inferred probabilistically given genotype data from other family members and/or by taking account of linkage disequilibrium with neighboring loci. A valid analysis can be achieved by use of a conditioning argument that essentially discards individuals or families in which the required information is not inferable, but at the cost of some loss in power and efficiency. Instead, we propose incorporation of the posterior probabilities of the possible configurations through a weighted regression procedure. In simulations, this method performs well in comparison to competing approaches such as multiple imputation.

26

Associating Haplotypes with Trajectories from Longitudinal Studies

Q Li, J Ning, K Bandeen-Roche, MD Fallin
Johns Hopkins University

The use of haplotypes in association studies can be more powerful compared to a single marker, however, they require statistical estimation of haplotype phase. Methods for associating haplotypes with outcomes have been developed for case-control, cross-sectional and time-to-event studies, but little is available for analysis of repeated measures. We investigate four methods to evaluate haplotype associations with repeated measures. This is currently possible by either (1) performing a 2-step mixed effects model where phase is estimated in step 1 and time∗ haplotype effects are estimated in step 2, or by (2) performing a GLM using fitted slopes as the trait of interest, which can be implemented for haplotypes using the iterative updating of phase and regression parameters described by Lake et al., 2003. What has been missing to our knowledge is the integration of these ideas so that a mixed effects models can be used while simultaneously updating phase information in a one-step process. We have developed SAS and R code to perform this one-step mixed modeling approach for longitudinal haplotype data analysis (method 3). In addition we have also developed a U-statistic method based on fitted slopes (method 4) that does not require phase estimation. We performed simulations assuming a linear relationship in outcome with time to compare type I error, bias, power and coverage across each method. These methods are also applied to data from the Women's Health and Aging study to investigate haplotype associations in inflammatory genes with change in cognitive test performance over time.

27

On the use of phylogeny-based tests to detect association between quantitative traits and haplotypes

C. Bardel(1), V. Danjean(2), P. Darlu(3) and E. Gnin(3)
(1) UMR 5145, CNRS-MNHN-Univ Paris VII, France (2) ID-IMAG, Univ Grenoble 1, France (3) INSERM U535, Villejuif, France

With the development of molecular techniques, numerous Single Nucleotide Polymorphisms (SNPs) are now available to establish their possible role in regulation of quantitative traits. However, the number of haplotypes drastically increasing with the number of investigated SNPs, the power of usual association tests between traits and candidate genes decreases unless some strategies are applied to group haplotypes. In this work, we propose a new test of association between quantitative traits and clusters of haplotypes based on their evolutionary relationships depicted by a phylogenetic tree. This test consists in a one-way analysis of variance (anova) comparing the level of the quantitative trait between the groups of haplotypes defined at each level of the tree. The type one error rate and the power of the test are assessed through simulations for various sample size and for different genetic models underlying the quantitative trait. The efficiency of our test was compared with a standard one-way anova directly performed on all ungrouped haplotypes and with another phylogeny-based test, TreeScan (Templeton et al., 2005). We found that the phylogeny-based tests often performed better than the standard haplotype test. Their respective power highly depends on the recurrence rate of mutations at the quantitative trait loci along the tree and on the genetic model, our test being more powerful than the others when the recurrence rate is high and for additive and dominant models.

28

Effects of genotyping error on estimated individual ancestry proportions and consequences for case-control studies in admixed populations

A. Ahn(1), N. Lim(2), S.J. Finch(2), D. Gordon(3), G.A. Chase(1), M.D. Fallin(4)
(1) Health Evaluation Sciences, Penn State College of Medicine, Hershey, PA 17033, (2) Applied Math & Stat., Stony Brook University, Stony Brook, NY 11790, (3) Genetics, The State University of New Jersey, Piscataway, NJ 08854, (4) Epi., Johns Hopkins School of Public Health, Baltimore, MD 21205

It is well known that genotyping error adversely affects parameter estimation for case/control association studies. To date, there is little research on genotyping error effects when estimating individual ancestry proportion(IA) in an admixed population used for association tests. Therefore, we have examined (1) if the genotyping error can increase the difference between the true IA and estimated IA using single nucleotide polymorphisms(SNPs) from case/control studies of unrelated admixed individuals, and (2) How the genotyping error, and resulting IA estimation affects case/control association tests when IA is adjusted for as a confounder. We simulated case/control data of admixed individuals with various genotyping error models, and genetic models then estimated IA using STRUCTURE and MLE methods, and tested association using logistic regression, adjusting for IA estimates. We also varied the number of informative markers for estimating IA and investigated how the genotyping error impacts on the differences as the number of informative markers increases. We found that genotyping error increased the difference between the true and estimated IAs, and we will show how this can lead to conflicting results of association tests when adjusting for admixture.

29

Use of covariates in case-control association analyses when the covariates are measured in cases only

P. Holmans
Biostatistics & Bioinformatics Unit, Cardiff University, UK

Use of covariates can increase the power of both linkage and association studies of complex traits (e.g. by reducing heterogeneity). When the covariate is only measured in cases (e.g. disease severity), standard methods, such as logistic regression, cannot be used. Macgregor et al. (2006, Eur J Hum Genet 14:529-534) proposed a strategy in which cases are added to the analysis sequentially in order of their covariate value. A standard chi-squared association statistic is calculated after each case is added, and the maximum of these is taken as the overall test statistic. Significance is assessed by randomly permuting genotypes between cases and controls. The method was applied to a study of bipolar disorder by Macgregor et al. In this presentation, the power of the sequential addition method is systematically investigated under a variety of models, including pleiotropy and gene-environment interaction. Power is compared to that of a standard case-control test (no covariates), andalso to a combined test where association is declared if either a standardcase-control test or a regression of covariate on genotype in the cases is significant.

30

Use of structural equation modelling to estimate pooled DNA frequencies

J Knight (1), D Campbell (1), F Rijsdijk (1), & P Sham (1,2)
(1) Institute of Psychiatry, Kings College London, UK. (2) University of Hong Kong, HK

The recent development of high throughput technologies allows hundreds of thousands of markers to be simultaneously genotyped. Although the price of this technology is falling it can still be financially prohibitive to type large samples. One cost reducing strategy involves performing preliminary investigations using pooled DNA and following up markers of interest with individual genotyping. This methodology has increased error hence reduced power but still provides a useful screening tool. In this paper we present a novel method for increasing the accuracy of pooled allele frequency estimates. We use structural equation modelling and fit a model where the genotypes of each marker are estimated not only on the basis of the pooled measure of their own frequency but

also using the pooled measure of allele frequency of other highly correlated markers. To demonstrate the utility of this approach work with 12 markers which have been both individually genotyped and typed in pooled samples. (The pools constructed of around 200 individuals each were genotyped using the Affymetrix 500K Gene Chip). We will investigate the methodology further using simulated data.

31

Association tests allowing for missing data in combined samples of nuclear families and unrelated subjects

F. Dudbridge
MRC Biostatistics Unit, Cambridge, UK

Missing data problems occur in genetic association studies when, for example, parents are missing from nuclear families, haplotype phase is uncertain, or sporadic genotyping failures occur. Although these problems have received much attention, none of the current solutions are entirely satisfactory. The popular TRANSMIT program has a known bias when testing multiple offspring in the presence of linkage; FBAT only gives basic tests and does not estimate effect sizes; and various programs for unrelated subjects do not integrate readily with nuclear family data. I propose a modification to the nuclear family likelihood that reduces to the TDT when data are complete, is strongly correlated to TRANSMIT when data are incomplete, but is unbiased in the presence of linkage. This is achieved by decoupling the relative risk parameter in the offspring from that in the parents, and introducing a novel step of conditioning on the inheritance vectors compatible with the family data. By regarding unrelated subjects as nuclear families with two missing parents, samples from both designs can be combined in a single analysis. The methods are implemented in a computer application, UNPHASED, available from the author.

32

Early Age-at-Diagnosis is a Characteristic of Familial Lobular Breast Carcinoma

K. Allen-Brady & L. Cannon-Albright
Department of Biomedical Informatics, University of Utah, USA

We and others have previously shown that lobular breast cancer (LOB) exhibits increased familial clustering; however, whether differences in breast cancer characteristics exist between familial and ''sporadic'' LOB cases has not been reported. Using the Utah Population Database, we identified 837 female LOB cases, with genealogy back at least 2 generations. Of 837 total cases, 383 belonged to familial clusters with significant excess ($p < 0.05$) of LOB, and the remaining 454 cases were termed sporadic. Among familial LOB cases a significantly higher percent of cases (39.7% vs. 28.6%) were diagnosed at a younger (<55 years) age ($\div 2=11.4$. $p=0.001$). Although familial cases may be screened earlier because of family history resulting in an earlier age-at-diagnosis, we observed no significant difference by group for either more localized disease (i.e., in-situ or localized carcinoma) or more well-differentiated cells, stratified by age-at-diagnosis (Stage: MH OR=0.88 95% CI=0.67-1.17; Grade: MH OR=1.01 95% CI=0.75-1.38). BRCA1/2 mutation screening had been performed previously on 27 (7%) familial cases and 7 (1.5%) sporadic cases. Five familial LOB cases (18.5% of those screened) were identified with deleterious BRCA1/2 mutations; no mutations were identified among sporadic cases. As earlier age-at-diagnosis is suggestive of genetic etiology, and as BRCA1/2 mutations are present in a minority of familial LOB cases, these findings add further evidence that familial LOB may include a heritable component and pedigrees with excess LOB may be informative for isolating additional breast cancer genes.

33

Detectable odds ratio in association studies

C.I. Amos and Y. Han
U.T. M.D. Anderson Cancer Center

With the advent of high-density SNP association platforms, procedures for the design of studies are still lacking. Here we present the availability of software (with versions in SAS and C) for detectable odds ratio and sample size computation for a variety of scenarios. In particular, we provide power estimation for case samples that have been selected through singletons or through a case with an affected sibling, or for mixtures of singletons and affected siblings. We also allow the user to specify various metrics including D′ or R-squared measures for linkage disequilibrium with the marker locus, and the user can specify either the allele frequency or genotype frequency. The software restricts the range of acceptable R-squared values if the marker and allele frequencies are not identical. Reductions in detectable odds ratios are achieved for virtually all combinations of allele frequency and prevalence when sampling cases with affected siblings, and these reductions can be dramatic for recessive diseases. Power is also dramatically influenced by the degree of similarity of the marker and disease allele frequencies. Power computations that use R-squared as a metric often implicitly assume a close correspondence of the marker and allele frequencies and so may be overly optimistic.

34

Power estimation in association studies

C.I. Amos and Y. Han
U.T. M.D. Anderson Cancer Center

With the advent of high-density SNP association platforms, procedures for the design of studies are still lacking. Here we present the availability of software (with versions in SAS and C) for detectable odds ratio and sample size computation for a variety of scenarios. In particular, we provide power estimation for case samples that have been selected through singletons or through a case with an affected sibling, or for mixtures of singletons and affected siblings. We also allow the user to specify various metrics including D′ or R-squared measures for linkage disequilibrium with the marker locus, and the user can specify either the allele frequency or genotype frequency. The software restricts the range of acceptable R-squared values if the marker and allele frequencies

are not identical. Reductions in detectable odds ratios are achieved for virtually all combinations of allele frequency and prevalence when sampling cases with affected siblings, and these reductions can be dramatic for recessive diseases. Power is also dramatically influenced by the degree of similarity of the marker and disease allele frequencies. Power computations that use R-squared as a metric often implicitly assume a close correspondence of the marker and allele frequencies and so may be overly optimistic.

35

Comparison of nontraditional designs for gene-environment interaction (GEI) detection

N. Andrieu(1), AM. Goldstein(2), W. Schill(3), P. Wild(4)
(1) INSERM U794, Inst. Curie, France, (2) NCI/NIH/DHHS,USA, (3) Univ. Bremen, Germany, (4) INRS, France

As the interest in GEI advances, it is clear that traditional designs do not have enough power for detecting many interactions. Alternative designs that allow for simultaneous estimation of the main and interactive effects, have been proposed to increase efficiency for GEI detection: two-phase, countermatched, flexible matching, flexible two-phase and case-related control designs. These designs all attempt to increase the frequency of the rare factor(s) through oversampling. Two phase and countermatched designs sample subjects from phase1. After classification according to an environmental or genetic exposure or an exposure surrogate, subsamples of subjects are selected in phase2 for GEI assessment. The variables used for subsampling must be available at very low cost for a large number of subjects at phase1 and the surrogates must have high sensitivity and specificity. We have compared the efficiency of GEI detection for the 5 designs. We calculated the relative efficiency defined as the ratio of the GEI variance for each design to the variance for a classical case-control study with equal numbers of cases and controls. We examined a rare and common dominant gene with an at-risk gene frequency of 1% and 20% respectively for different scenarios. Details of comparisons will be presented. In summary, the most appropriate design(s) will depend on the environmental and genetic factor characteristics and their effects. Further, the ability to accrue cases and controls and the costs per subject will influence design suitability.

36

Evidence for sex-specific quantitative trait loci underlying asthma-related phenotypes in the French EGEA study

H. Aschard (1), E. Bouzigon (1), M.H. Dizier (2), A. Ulgen (1), M.P. Oryszczyn (3), J. Maccario (3), M. Lathrop (4), F. Kauffmann (3), F. Demenais (1)
(1) INSERM U794, France (2) INSERM U535, France (3) INSERM U780, France (4) Centre National de GÃ ©notypage, France

Sex differences in asthma-associated phenotypes are well known but the genetic factors which may account for these differences have received little attention. To localize sex-specific quantitative trait loci (QTL) underlying asthma-related phenotypes, we extended our genome screen conducted in the whole sample of 295 EGEA families to male-only and female-only datasets. Five quantitative phenotypes were examined: immunoglobulin E (IgE) levels, sum of positive skin prick tests to allergens (SPTQ), eosinophil counts (EOS), percent predicted forced expiratory volume in 1 s (%FEV1) and bronchial responsiveness (BR) to methacholine test. Linkage was investigated using the Variance Components method (QTDT programme). Genome-wide suggestive significance for sex differences in linkage signals was assessed by permutation testing. Evidence for sex-specific linkage was found for two phenotypes in six regions: 3 regions for %FEV1 (9q33, 12q24, 20q13) and 2 regions for SPTQ (3q26, 7q34). The highest peak LOD score (LOD=2.66 , p=0.00025, 7q36) was reached for SPTQ in the male-only dataset while the other four sex-specific signals had p-values between 0.00035 and 0.0009. Most signals weren't revealed at the 0.5% level in the whole data set. These results show the importance of taking into account sex-specific effects to increase our power to identify genes underlying complex traits.

37

***XRCC3* variants, tobacco exposure, and incident CHD: The Atherosclerosis Risk in Communities (ARIC) Study**

CL Avery(1), DA Canos(1), D Couper(2), AF Olshan(1), C Poole(1), MS Bray(3), KE North(1)
Depts. of (1) Epid and (2) Biostat, Univ. of North Carolina, US (3) Baylor College of Medicine, US

DNA repair, including double strand break (DSB) repair pathways, may mediate atherogenic properties of tobacco smoke and thus influence CHD risk. As *XRCC3* is an integral DSB gene, we conducted a series of case-cohort analyses to examine how seven *XRCC3* variants modify the relationship between tobacco exposure and CHD in the ARIC cohort. All incident CHD cases 1987–1998 (n=1086) and a stratified random sample (n=1085) at baseline were selected from 15,792 participants. Analyses were stratified by race and adjusted for sampling strategy and study center. Incidence rate ratios (IRR) were estimated with controls weighted proportional to person-time at risk. Departures from additivity were assessed by interaction contrast ratios (ICR) and SNPs were measured with a dominant model. When combined with an ever-smoking history, the intronic SNP rs1799796 increased the effect of smoking on CHD in Caucasians *ICR=0.6(0.1, 1.0)*, with IRR estimates for both exposures, smoking alone and the SNP alone of 1.6(1.1, 2.2), 1.2(0.9, 1.8) and 0.7(0.5, 1.0), respectively. The nonsynonymous SNP rs861531 was associated with a reduction in the effect of smoking on CHD in Caucasians *ICR=-0.8(-1.7, 0.1)*, as IRR estimates for both exposures, smoking alone and the SNP alone were 1.7(1.2, 2.5), 2.2(1.5, 3.2) and 1.4(0.9, 2.0), respectively. Even as smoking prevention and cessation remain priorities, smoking-related CHD susceptibility may be modulated by *XRCC3* variants.

38

A Comparison of Operating Characteristics of Haplotype Logistic Analysis Using HelixTree and HaploStats

K. R. Bailey, E. Boerwinkle, A. B. Chapman, P. Thapa, B. L. Fridley, S. T. Turner
Mayo Clinic

One of the tools for Genetic Association Analysis is haplotype regression modeling, including logistic models that are linear in the counts of each of the haplotypes. Two software packages, HelixTree, and HaploStats, use different approaches and statistic s for testing genetic association with a binary phenotype. Specifically, HelixTree uses a likelihood ratio chi-squared statistic, while HaploStats uses a score chi-squared Statistic. The same basic algorithm (an EM algorithm) is used by both programs to generate haplotype frequency estimates under the null hypothesis. Nevertheless, differences in the two methods' approaches to hypothesis testing appear to result in different operating characteristics under the null hypothesis (Type I error rates). In a specific example arising from a pharmacogenomic study (GERA) of hypertensive therapy response in which the AFFY 100K SNP chip array was acquired on 193 African American Hypertensive subjects, a simulation study (n=4) using random case-control assignments applied to all of the SNP's on Chromosome 12 (N=4673) was conducted. A moving window of 3 SNP's was used, and the associated haplotype analyses conducted using both software systems. The results suggested that, at the 0.05 and 0.005 nominal levels of significance, HelixTree software operated at an approximately 0.065 and 0.0067 Type I error rates, while HaploStats operated at Type I error rates of 0.040 and 0.0032, respectively. This suggests that HaploStats, but not HelixTree, provides P-values that can be interpreted as no smaller than the probability of obtaining as extreme a result by chance. Moreover, Helixtree failed to return a P-value approximately 11% of the time, while HaploStats invariably returned a P-value. The reasons for these differences, their dependence on sample size and haplotype frequencies, and implications for interpretation of genome-wide scans will be discussed.

39

Multiple QTL influence the serum Lp(a) concentration: A genome-wide linkage screen in the PROCARDIS study

Barlera(1), C. Specchia(1), M. Farrall(2), B.D. Chiodini(1), M.G. Franzosi(1), S. Rust(3), E. B. Nicolis(1), PROCARDIS Consortium(4)
(1) Istituto Mario Negri, Milano, Italy (2) Wellcome Trust Centre for Human Genetics, Oxford, UK (3) Leibniz-Institut, Münster, Germany (4) www.procardis.org

Objective and Methods: The serum concentration of lipoprotein Lp(a) is known to be highly heritable and associated with cardiovascular risk. A genome-wide variance component linkage analysis was performed to localise QTLs influencing Lp(a) levels in a large cohort collected in the PROCARDIS coronary heart disease study. Results: Highly significant linkage was detected at the previously described LPA locus on chromosome 6q27 (LOD 108). Taking into account the effect of the locus detected on chromosome 6, a highly significant LOD score was detected on chromosome 13q22-31 (LOD 7.0). Other regions suggesting linkage were observed on chromosomes 11p14-15 (LOD 3.5), 15q23-25 (LOD 2.9) and 19q13 (LOD 2.7). The significant peak at 13q22-31 shows an essential overlap with a locus modulating cholesterol in familial hypercholesterolemia. If the gene underlying these loci is the same, it will be a promising candidate target for manipulating LDL-cholesterol and Lp(a). We also confirmed the presence of a previously identified locus influencing Lp(a) on chromosome 1q23 (LOD 1.5). Conclusions: Our findings provide new and confirmatory information about genomic regions involved in the quantitative variation of Lp(a) and serve as a basis for further studies of candidate genes in these regions.

40

Genetics Informatics: The Universal Genetics Database and the Genetics Information Commons

M. M. Barmada
Dept of Human Genetics, Univ of Pittsburgh, Graduate School of Public Health

Increased data flow from high-throughput genomic technologies such as whole-genome microarrays and "SNP-chips", large consortium studies, proteomics, and computerized medical records initiatives are creating a deluge of genetic data that threatens to completely overwhelm the ability of most analytic personnel (or resources) to manipulate in a meaningful and efficient manner. A requirement for properly dealing with this amount of data is a resource for integrating disparate information sources and for presenting appropriate realizations of the data for each member of a multidisciplinary research team. To this end we have created a Universal Genetics Database (UGD) encompasing a peer-to-peer data store with a unique, extensible, schemaless information architecture. As a result of this architecture, the UGD encourages the free dissemination and reuse of information. Public data sets in the UGD form a Genetics Information Commons, and are free to mix with other public data in the larger Information Commons http://www.maya.com/infocommons/). Other key benefits of the system include the use of universal identifying keys to enable standoff annotation and the reuse of results in virtual datasets and meta-projects. The UGD promises to provide a resource that promotes multidisciplinary and multi-center studies, decreases the time required for preparation of results, and enhances the ability of researchers to investigate data in a cross-domain fashion, decreasing the time-to-publication of studies and increasing the ability of research teams to make use of the multitude of data already in existence.

41

A Latent Class Model for Heterogeneity in a Regression-Based Test of Linkage.

Bastone(1), M. Putt(1), T. TenHave(1), R.S. Spielman(2)

(1) Division of Biostatistics, (2) Department of Genetics, University of Pennsylvania, USA

Heterogeneity due to the presence of non-segregating families poses a challenge to QTL-mapping. Non-segregating families, i.e. those with neither parent heterozygous at the QTL, are likely to be present in an unselected sample. Modeling heterogeneity is important since the presence of non-segregating families reduces the ability of statistical methods to detect linkage. When linkage is detected, it is desirable to know which families contribute to the linkage evidence. We extend Haseman-Elston regression to account for heterogeneity and apply our method to map regulators of gene expression. Expression phenotypes with significant ''cis linkage'' in 14 CEPH pedigrees from Morley et al. *Nature 430, 743–747 (2004)* are analyzed using a latent class model to account for heterogeneity and test for linkage to SNP markers. We provide the theoretical motivation for our method and derive expectations of the parameters in the latent class regression model, in terms of an assumed additive genetic model. We compare the use of various phenotype functions (e.g. mean-corrected product) as the outcome in the latent class model. A permutation test is used to test the joint hypothesis of heterogeneity and linkage. Families are classified with confidence as segregating or non-segregating, based on their posterior probability of latent class membership. Simulations are used to assess the statistical properties of the hypothesis test and classification procedure. This method can be extended to model the genetic (locus) heterogeneity of complex traits.

42

Application of Mantel statistics using haplotype sharing to APM1 and adiponectin plasma levels

L. Beckmann(1), I. Heid(2), J. Chang-Claude(1), S. Wagner(2), B. Paulweber(3) F. Kronenberg(4)
(1) German Cancer Research Center, Heidelberg, Germany (2) Natl. Research Center for Environment and Health, Neuherberg, Germany (3) Paracelsus Private Medical Univ., Salzburg, Austria (4) Innsbruck Medical Univ., Innsbruck, Austria

Recently, we proposed a new approach based on Mantel statistics that correlate phenotypic similarity and genetic similarity based on haplotype sharing (Beckmann et al. Hum Hered. 2005 59(2):67–78). So far, the proposed method was applied only to data from case-control studies with binary outcomes. The aim of the study presented here was to elucidate the applicability of the Mantel statistic to quantitative traits. We have thus applied the method on the example the *APM1* gene and its association with plasma adiponectin levels. We analyzed 18 SNPs in two haplotype blocks in *APM1* on chromosome 3q27 with an average intermarker distance of 1.3 kB. The sample consisted of 1,727 unrelated healthy Caucasians from the SAPHIR study (Salzburg Atherosclerosis Prevention Program in subjects at High Individual Risk), Austria. We were able to confirm the significant associations of the SNPs in *APM1* with adiponectin plasma levels that were found by Heid et al. (Diabetes 2006 55:375–384). After correction for the family-wise error rate using a step-down minP algorithm, the magnitude of statistical significance was $p < 0.0001$ for all SNPs in block 1, and ranged between $p=0.0001$ and $p < 0.0005$ in block 2. Our results suggest that Mantel statistics using haplotype sharing is an appropriate approach for the analysis of quantitative traits.

43

A composite likelihood approach for inference on directly associated polymorphisms

J.M. Biernacka, H.J. Cordell
Inst. of Human Genetics, Newcastle Univ., UK

Direct association of disease with a sole causal variant in a region should explain all the linkage in that region; whereas indirect association due to incomplete linkage disequilibrium (LD) between a candidate polymorphism and a causal variant may not be able to explain the observed linkage. Recently, several methods have been proposed for identification of disease associated polymorphisms that can explain an observed linkage signal (e.g. Li et al. 2005, Biernacka and Cordell, 2005). These methods assume a single causal SNP in a region, and attempt to identify it from among a set of candidate SNPs, by testing whether association with any of the candidate SNPs can explain the observed linkage result. Biernacka and Cordell (2005) used a likelihood approach and estimated relative risk and LD parameters, allowing for a test of the null hypothesis that the test SNP is the only causal variant in the region, or is in complete LD with the sole causal variant in the region. Furthermore, they extended this method to test for complete LD between a disease locus and a haplotype composed of two candidate SNPs. Here we present a composite likelihood approach for estimation of relative risk and LD parameters via joint modelling of information from several candidate SNPs. A likelihood ratio test of whether association with any one of the candidate SNPs can fully explain the observed linkage can then be carried out. We compare and contrast the approaches described by Biernacka and Cordell (2005) with the composite likelihood approach, present simulation results, and illustrate the methods using type 1 diabetes data.

44

Genome-wide search for multiple loci in type 2 diabetes

J.T. Bell(1), K. Elliot(1,2), A. Morris(1), C.J. Groves(1,2), S.E. Fiddy(1), T.M. Frayling(3), A.T. Hattersley(3), M. Walker(4), G.A. Hitman(5),M.I. McCarthy(1,2), S. Wiltshire(1)
(1) WTCHG and (2) OCDEM, Univ Oxford, (3) Peninsula Medical School, Exeter, (4) Dept Medicine, Univ Newcastle, (5) Barts and The London School of Medicine and Dentistry, UK

Multiple loci contribute to type 2 diabetes susceptibility, potentially involving complex epistatic interactions. We performed a systematic two-dimensional (2D) genome scan for multiple susceptibility loci in 573 pedigrees from the Warren 2 consortium. Our results identified one genome-wide significant gene-gene interaction (1q24-

10q23, two-locus MLS=5.83) and 15 other pairs of regions that surpassed genome-wide suggestive evidence for two-locus linkage. For each interaction, we identified the most likely additive, multiplicative, or epistatic underlying genetic model. We searched across two-locus penetrance models to select a set of models consistent with the two-locus effects observed at the 16 2D peaks. We undertook an automated search of the published molecular interaction databases using the regions involved in our peak two-locus findings. The resulting pair-wise molecular interactions underlying the 2D linkage peaks identified several candidates, including components of the FAS-regulated apoptotic pathway. In preparation for analyzing the 2D interactions at the nucleotide level, we performed two-locus simulations of case-control data under the two-locus models obtained from the 2D linkage results to explore future analyses for two-locus association with epistasis. Our findings support the evidence for joint action of multiple loci in type 2 diabetes.

45

Multivariate extension of the Maximum-Likelihood-Binomial method shows pleiotropic effect of 5q13 on asthma expression and age of onset in the EGEA study

E. Bouzigon(1), V. Siroux(2), MH. Dizier(3), A. Ulgen(1), C. Pison(4), F. Kauffmann(5), I. Pin(2), F. Demenais(1)
(1) INSERM U794, (2) INSERM U578, (3) INSERM U535, (4) INSERM 221, (5) INSERM U780, France

Asthma is a complex disease displaying variable expression and age of onset. Although many genome-wide screens have been conducted for asthma as a binary trait, there is limited information regarding the genetic factors underlying the variation of asthma expression. We conducted univariate and bivariate linkage analyses of the two following phenotypes: 1) a categorical asthma score representing the spectrum of disease expression and 2) age of onset of asthma in 110 EGEA families. We used the Maximum Likelihood Binomial (MLB) method, suited to categorical and continuous phenotypes. This method was extended to bivariate analysis which consisted of first conducting a principal components (PC) analysis of the two phenotypes which were then subjected to independent MLB analyses. The test statistic was constructed by summing the MLB likelihood-ratio statistics, following an approach suggested by Mangin et al (Biometrics, 1998). This test was assumed to follow a mixture of chi-square distributions. Univariate linkage analyses detected linkage to the asthma score on 18p11 (LOD=2.4, p=0.0004) and to asthma age of onset on 1p31 (LOD=1.6, p=0.003) and 5q13 (LOD=1.7, p=0.003). Bivariate linkage analysis led to a substantial improvement of the linkage signal on 5q13 (p=0.00007). This result indicates the presence of a pleiotropic genetic factor on 5q13 influencing the variation of asthma expression and age of onset.

46

A Polygenic Model to Identify SNP Profiles Associated with Breast Cancer Risk

V. Onay , S. Savas, H. Ozcelik
L. Briollais (1), V. Onay (2), S. Savas (2), H. Ozcelik (2)
(1) Prosserman Centre for Health Research (2) Fred A. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Toronto, Canada

Mutations of only few genes (e.g. BRCA1,BRCA2, TP53,ATM) were shown to contribute to breast cancer predisposition. Such high penetrant and rare mutations account for less than 5% of breast cancer cases. The genetic background of the remaining breast cancer cases is not known. Low to moderate-penetrant commonly occurring SNPs are also shown to be associated with increased breast cancer. However, compared to BRCA mutations, the magnitude of risk detected by these SNPs is considerably small. Polygenic models have been proposed to explain the joint multiplicative effect of many susceptibility alleles on breast cancer risk in the population (Pharoah et al., 2002). However, the risk prediction made by the authors is theoretical and has not been validated in real data sets. In this study, we have evaluated the polygenic susceptibility to breast cancer explained by 20 SNPs in the DNA-repair pathway. These SNPs have been genotyped in breast cancer cases (n=400) and population controls (n=400) from Ontario, Canada, and also have been characterized on the basis of functionality using bio-informatics computational modeling and a protein-protein interaction map of DNA repair pathway. The effects of the SNPs on breast cancer risk has been modeled by log-linear and graphical model approaches. The polygenic model suggests that a risk profile based on combination of SNPs and other risk factors is more likely to provide risk discrimination that has practical value for health care than single risk factors.

47

Validation of Clusters Derived from Genetic Marker Data

G.N. Brock(1), R.N. Baumgartner(2), M.L. Slattery(3), T. Byers(3), A. Guiliano(4), K.B. Baumgartner(2)
(1) Dept. of Bioinformatics & Biostatistics, Univ. of Louisville, USA, (2) Dept. of Epidemiology & Population Health, Univ. of Louisville, USA, (3) Health Research Center, University of Utah, USA, (3) University of Colorado School of Medicine, USA, (4) Moffitt Cancer Center, USA

A variety of methods currently exist for determining population substructure using genetic marker data, including Bayesian (Structure 2.1), hybrid Bayesian/classical (ADMIXMAP), classical maximum-likelihood (PSMIX), and neighbor-joining methods (Neighbor in Phylip). The number of markers required to accurately cluster individuals depends on a variety of factors, including marker informativity, the number of source populations, the percentage of admixture within each individual, and the sample size of individuals from each source population. Several papers have addressed the theoretical aspects of which markers are the most informative for determining population substructure. While each of these studies gives some empirical recommendations on the number of markers needed for accurate clustering, currently there is no overall consensus on this issue. In this work, we examine several measures to inspect the consistency

of clustering results and help determine whether a sufficient number of markers has been used for clustering. We evaluate our methods using simulated data and data from a case-control study of breast cancer among Hispanic, American Indian and non-Hispanic white women in the 4-Corners area of the United States.

48

Linkage Analysis of the Age of Onset of Thrombosis

A. Buil(1), L. Almasy(2), J.C. Souto(1), J. Fontcuberta(1), J. Blangero(2), J.M. Soria(1)
(1) Unitat de Hemostasia i Trombosis, Hospital de la Santa Creu i Sant Pau, Barcelona, SPAIN. (2) Dept. Genetics, Southwest Foundation for Biomedical Research, San Antonio TX, USA

The analysis of the "age of onset" of a disease is often more informative than the analysis of the simple yes/no affection status. As our sample is composed by related individuals, grouped in families, we use an extension of the Cox model that enables the study of non-independent observations including correlated random effects in the model. In fact, this strategy is a hybrid between the traditional Cox model and the variance components methods widely used in the field of statistical genetics. We applied this method to the age of onset of thrombosis in the Genetic Analysis of Idiopathic Thrombophilia (GAIT) project. The GAIT sample consists in 399 individuals organized in 21 extended pedigrees. Our response variable is the age at the first event of thrombosis for the patients and the age at the beginning of the study for the healthy individuals (censored individuals). We obtained five genomic positions, in chromosomes 2, 3, 4 and 12, with a LOD score greater than 1. Due to the novelty of the method, it is difficult to evaluate the significance of these results. Previous simulations suggest that a LOD score cut-off of 1.3 can provide a moderate power with a small type I error. In our sample, the linkage signal on chromosome 12 gets a LOD score of 1.3. We applied a general variance components method for time-to-event data to the age of onset of thrombosis, and we obtained evidence of suggestive linkage in five genomic positions.

49

Two-stage Genomewide Association: Power and Sample Size for Replication

S.B. Bull(1,2), L. Sun(2,3), X. Xie(1), L.Y. Wu(1), A.D. Paterson (3)
(1) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Univ of Toronto, Toronto, Canada. (2) Department of Public Health Sciences, Univ of Toronto, Toronto, Canada, (3) Hospital for Sick Children, Toronto, Canada

Because of improved efficiency with respect to genotyping costs and/or sample accrual, high density genomewide association (GWA) studies are typically designed to have multiple genotyping and analysis stages. Whether the second stage involves an independent sample of individuals or the use of a family-based design, one can expect effect estimates at the second stage to be less optimistic that those obtained at the first stage, due to the selection bias that arises from genomewide screening for significance. The degree of bias depends on the sample size, the true genetic effect size, and the stringency of the significance criteria. Application of computationally-intensive bootstrap estimation in the first stage can yield less biased effect estimates, and hence more realistic design specifications for replication in a second stage. Motivated by the development of a genomewide association study of the genetics of complications of type I diabetes, we evaluate the implications of selection bias for power and sample size in alternative designs and analytic strategies for detection and mapping of candidate gene regions using high-density SNP arrays.

50

Multifactor Dimensionality Reduction using Normalized Mutual Information

WS Bush, TL Edwards, SM Dudek, MD Ritchie.
Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37215

Multifactor Dimensionality Reduction (MDR) has been previously introduced as a non-parametric statistical method for detecting gene-gene and gene-environment interactions. MDR performs a dimensional reduction by assigning multi-locus genotypes to either high or low risk groups and measuring the percentage of cases and controls incorrectly labeled by this classification – the classification error. The combination of variables that produces the lowest classification error is selected as the best or most fit model. The correctly and incorrectly labeled cases and controls can be expressed as a two-way contingency table. We sought to improve the power of MDR to detect interaction effects by replacing classification error with a different metric to assign model fitness. In this study, we compare the power of MDR using a variety of fitness metrics for two-way contingency table analysis. We simulated 40 genetic models, varying the number of loci in the model (2 - 5), allele frequencies (.2/.8 or .4/.6) and the epistatic heritability of the model (.3–.05). Normalized Mutual Information is a measure of information transfer that improves the power of MDR to detect gene-gene interactions in simulated data and has useful theoretical properties for measuring classification performance.

51

When a Case is Not a Case: Power Loss for the TDT with Misdiagnosed Cases

S. Buyske(1,2) D. Gordon(2)
(1) Dept of Statistics, Rutgers University, USA, (2) Dept of Genetics, Rutgers University, USA

Of all the assumptions that enter a power analysis for a genetic study of a binary trait, the one that is the least explicit is that "cases" are actually cases. We consider the power loss in a case-parent design analyzed with the transmission/disequilibrium test when a portion of the claimed cases are improperly diagnosed. We provide both exact calculations (with one simplifying assumption) and

simulation results. The power loss due to including false cases is largely insensitive to the factors, such as marker and disease allele frequency, disequilibrium, and genetic model, that determine the probability of transmission of a marker allele to a true case. The relative power loss does depend on the significance level of the test as well as the nominal power of the design The table below shows, for 2 power levels and various proportions of true cases, the ratio of the sample size needed for the actual power compared to the sample size for the nominal power when all cases are erroneously assumed to be true cases. A significance level of alpha=.001 was used in the table; higher ratios are required for smaller alpha.

Proportion of True Cases:					
Power	.99	.95	.90	.85	.80
–.8	1.05	1.28	1.67	2.21	3.01
.9	1.07	1.44	2.17	3.47	6.13

Can this power loss be alleviated by including a subset of "gold standard" diagnoses? We show how tests based on summary statistics from the gold standard and non standard subsets can improve the power.

52

Localization of a Prostate Cancer Predisposition Gene to an 880 kilobase Region on Chromosome 22q12.3 in Utah High-Risk Pedigrees

N.J. Camp, J.M. Farnham, L.A. Cannon-Albright
Division of Genetic Epidemiology, Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA

Chromosome 22q is a region of interest for prostate cancer (PRCA). We previously identified a LOD=2.42 at chromosome 22q12.3. This region has also been noted by eight other studies (LODs 1.50-3.57). Here, we perform fine-mapping and localization using a pedigree-specific recombinant estimation approach in 14 informative, high-risk Utah pedigrees. The 14 pedigrees used in the localization were considered to be linked or haplotype sharing, or both. Linked pedigrees were those with pedigree-specific LOD>0.588 ($p < 0.05$) at the 22q12.3 region, regardless of the number of PRCA cases sharing the segregating haplotype. Haplotype sharing pedigrees were those with at least 5 PRCA cases sharing a segregating haplotype in the 22q12.3 region, regardless of the linkage evidence. In each pedigree the most likely haplotype configuration (in conjuntion with the multipoint LOD graph) was used to infer the position of recombinant events and delimit the segregating chromosomal segment. These pedigree-specific segments were then overlaid to form a consensus recombinant map across all 14 pedigrees. Using this method, we identified a 881,538 bp interval, between D22S1265 and D22S277, that is the most likely region containing the PRCA predisposition gene. The Utah extended high risk pedigrees remain powerful for this type of localization approach. We are mutation screening candidate genes in this region to identify specific genetic variants segregating in these pedigrees.

53

Efficient computation of individual latent variable scores from data with multiple missingness patterns

DD Campbell(1), FV Rijsdijk(1) and PC Sham(1,2)
(1) MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park (Box P080), London SE5 8AF, UK(2), Department of Psychiatry and Genome Research Centre, Faculty of Medicine, University of Hong Kong, Hong Kong

Latent variable models are used in biological and social sciences to investigate characteristics that are not directly measurable. The generation of individual scores of latent variables can simplify subsequent analyses. However, missing measurements in real data complicate the calculation of scores. Missing observations also result in different latent variable scores having different degrees of accuracy which should be taken into account in subsequent analyses. Web present a publicly available software tool that addresses both these problems, using as an example a dataset consisting of multiple ratings for ADHD symptomatology in children. The program computes latent variable scores with accompanying accuracy indices, under a 'user-specified' structural equation model, in data with missing data patterns. Since structural equation models encompass factor models, it can also be used for calculating factor scores. The program, documentation and a tutorial, containing worked examples and specimen input and output files, is available at http://statgen.iop.kcl.ac.uk/lsc.

54

Estimating Haplotype Relative Risks in Complex Disease From Unphased SNPs Data in Families Using a Likelihood Adjusted for Ascertainment

J. Carayol, A. Philippi, F. Tores
Integragen, Evry, France

We propose an ascertainment adjusted likelihood-based method to estimate haplotype relative risks using pooled family data coming from association and/or linkage studies that were used to identify specific haplotypes. Haplotype-based analysis tends to require a large amount of parameters to capture all the information that leads to efficiency problems. An adaptation of the Stochastic Expectation Maximization algorithm is used for haplotypes inference from genotypic data and to reduce the number of nuisance parameters for risk estimation. Using different simulations, we show that this method provides unbiased relative risk estimates even in case of departure from Hardy-Weinberg equilibrium.

55

Hierarchical Modeling in Genome Wide Association Studies

G.K. Chen, E. Jorgenson, J.S. Witte
Dept. of Epi & Biostat, Univ. of CA, San Francisco, USA

The completion of the HapMap project and advances in high-throughput genotyping methods has made feasible

genome-wide association (GWA) studies. Such studies generally evaluate the relationship between hundreds of thousands of single nucleotide polymorphisms (SNPs) and one or more phenotypes. A critical unresolved issue in GWAs is how to analyze the enormous amount of information generated in a manner that is most likely to detect causal variants. The conventional analysis approach entails estimating the association between each SNP (or multiple SNPs) and a phenotype, and then using the corresponding p-values to prioritize the results. This approach, however, ignores existing information about the SNPs, can lead to spurious results as well as suffer from low power. To address these issues, we propose here a hierarchical model, adding a prior model that incorporates information about the SNPs into a conventional analysis. In particular, we develop a hierarchical model that uses the following information on the SNPs: 1) LD with neighboring SNPs; 2) potential functionality; 3) previous evidence of linkage and/or association. Using SNP genotype data from HapMap phase II and publicly available gene-expression phenotypes, we show empirically how integrating this information in a hierarchical model may improve the ability of GWAs to determine the location of causal variants.

56

Multi-marker association test while controlling population stratification

H.S. Chen, Q. Sha, S. Zhang
Department of Mathematical Sciences, Michigan Technological University, USA

Statistical tests for association studies using unrelated individuals tend to have inflated type I error when there is population stratification among the individuals. Several methods such as genome control (GC), structured association (SA), and semi-parametric test (SPT) have been proposed to control for the population stratification. However, the existing association tests that can control population stratification are all single marker methods, that is, testing one marker at a time. For complex disease, to develop methods that can jointly considering multiple markers has become more important. Current multi-marker association methods are valid for homogeneous population and will lead to false-positive results in the presence of population stratification. In this paper, we develop computational efficient multi-marker methods to control for false-positives due to population stratification.

57

The role of BDNF in obesity etiology: results from a population-based cohort study

L. Chen(1), I. Day(1), N. Timpson(1), J. Heron(1), J. Golding(1), G. Davey-Smith(1), Y. Yao(2)
University of Bristol, UK

Several lines of evidence indicate that brain-derived neurotrophic factor (BDNF) plays an important role in eating behaviour and early-onset mood disorders. More recently, it has been shown that intraventricular administration of BDNF in rats controls food starvation and body weight loss, while BDNF or its specific receptor NTRK2 knockout mice develop obesity. In this study, we conducted a nested case-control study to test for association between BMI and two functional SNPs in the BDNF gene in 7,529 women aged between 41 and 45 years old. Study subjects are known as ALSPAC mothers who were collected from the old Avon County situated on the south west coast of England along the Severn Estuary, 120 miles west of London. We examined two functional SNPs: BDNF Val66Met and −270C/T in BDNF. Using the statistical methods implemented in Haplo.stats, we detected a significant association between BMI (treated as a quantitative trait) and a haplotype of these two SNPs. Analyses conducted by Haplo.stats gave a global p-value of 0.0017 and a haplotype-specific p-value of 0.0006 with a protective effect. When covariates such as age, marriage status were adjusted in the regression model, the association remained to be statistically significant. To summarize, our results suggest an inverse association between BMI in ALSPAC mothers. We will replicate this findings in ALSPAC children with detailed life-style information.

58

Identifying SNPs Explaining Partially a Linkage Signal: Combining Homozygote Sharing and Transmission/Disequilibrium Tests Using Affected Sib Pairs

M.-H. Chen(1), P. Van Eerdewegh(2), J. Dupuis(3)
(1) Dept. of Math & Stat, Boston Univ., USA, (2) Genizon BioSciences Inc., Canada & Dept. of Psychiatry, Harvard Univ., USA, (3) Dept. of Biostat, Boston Univ., USA

Once a linkage region is identified, the next step is often to test single nucleotide polymorphisms (SNPs) for association with the disease within the region to identify the causal gene(s). If a SNP shows significant association with the phenotype, a question of interest is to determine whether the linkage evidence could be explained partially or fully by the SNP. Methods that have been developed for this purpose either condition on parental genotypes, such as the Homozygote Sharing Test (HST), or use offspring genotypes, such as implemented in the software LAMP proposed by Li, Boehnke and Abecasis (Am J Hum Genet (2005) 76:934–49). To identify SNPs explaining partially the linkage evidence, we propose a method that combines HST and the transmission disequilibrium test (TDT) for affected sib pairs because these two tests are complementary: HST examines IBD sharing from homozygous parents to determine if a SNP explains the linkage evidence, while TDT uses information from heterozygous parents only. We name the combined statistic HSTDT. We performed a simulation study to determine the power and type-I error of four methods for identifying SNPs explaining partially the linkage evidence: HST, TDT, HSTDT and LAMP. LAMP gave slightly inflated type-I error for most models examined. All other methods gave appropriate type-I error. For most models examined, TDT and LAMP had the highest power, followed closely by HSTDT.

59

Association between prostate cancer risk and rs1447295 at chromosome 8q24: a replication study

I. Cheng(1), X. Liu(1), S.J. Plummer(2), G. Casey(2), J.S. Witte(1)
(1) Department of Epidemiology and Biostatistics and Center of Human Genetics, University of California, San Francisco, San Francisco, CA, USA, (2) Department of Cancer Biology, Lerner Research Institute, The Cleveland Clinic, Cleveland, OH, USA

Deciphering the genetic basis of prostate cancer has proved quite difficult, though an extremely compelling marker was recently detected on chromosome 8q24. We provide independent confirmation of this in an association study of 1,012 men. We found that men carrying the common SNP rs1447295 at 8q24 have a statistically significant 40% increased risk of being diagnosed with advanced disease (P=0.03). Taken together, these results strongly implicate this common variant at 8q24 in this complex disease.

60

Flanking makers in linkage equilibrium reduce bias due to linkage disequilibrium among dense SNPs in multipoint linkage analysis of both qualitative and quantitative traits

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

Most current multipoint genetic linkage analysis methods assume linkage equilibrium (LE) between genetic markers. This assumption is violated among dense SNPs, where linkage disequilibrium (LD) may exist. In qualitative trait linkage analysis, studies have shown LE assumption among dense markers induces false-positive evidence for linkage with missing parental genotypes. This bias may be influenced by SNPs in LD, which can cause apparent oversharing of multipoint IBD between relative pairs. In our study, we examined the effects on bias of adding flanking markers in LE around the two previously studied markers in LD in both affected sib pair (ASP) and quantitative trait locus (QTL) linkage analyses. Our simulation under no linkage considered various levels of LD, with none, one or both parental genotypes missing. Pedigrees consisted of two to four offspring, each with at least two, one or none selected from the upper 20% of the phenotype distribution. Six analysis methods are compared: four regression approaches and two variance component statistics (likelihood ratio and score statistic). We observed a notable decline to complete elimination of bias with flanking markers in LE in both ASP and QTL linkage analyses with missing parental genotypes. In addition, our study showed a decrease in bias with increasing number of offspring in a family and with decreasing LD between markers. We found that bias can be greatly reduced with complete parental genotype information and by having flanking markers in LE around the markers in LD.

61

Bias and Efficiency of Family-based Designs for Estimating the Lifetime Risk Associated with Rare Mutations of a Disease-Susceptibility Gene

Y.-H. Choi and L. Briollais
Dept. of Epi. & Biostat., Samuel Lunenfeld Research Institute, Toronto, Canada

Many clinical decisions require accurate estimate of disease risks (penetrance) associated with inherited mutations of disease susceptibility genes. The penetrance is often estimated from family-based designs (population- or clinic-based with random or carrier proband) that require appropriate correction for the sampling bias. There is no consensus on which of these different designs provides more accurate and efficient estimates of penetrance. Our objective is to evaluate their relative performance under several genetic models (including one or two disease genes). The bias and efficiency are studied via simulations using prospective, retrospective and joint likelihoods. The proposed methods were also applied to a real data set of HNPCC pedigrees from Newfoundland. The simulation results show that the clinic-based design with random probands yields the most efficient estimates of the relative risk for the major gene effect, whereas the population-based design with carrier probands provides the most efficient penetrance estimates. Overall, the proposed likelihood methods provide nearly unbiased estimations both in the relative and absolute risks when there is no or small genetic heterogeneity. However, the presence of a strong second gene effect increases the bias in risk estimation. We also found that the joint likelihood serves as the most efficient likelihood method. In conclusion, this work could help to develop optimal study designs for penetrance estimation of rare mutations involved in complex human disease.

62

Genetic susceptibility of prostate cancer: genome-wide screen of men with non-aggressive disease

G.B. Christensen, N.J. Camp, J.M. Farnham, L.A. Cannon-Albright
Department of Biomedical Informatics, University of Utah, USA

The genetic basis of Prostate Cancer (PCa) is complex and poorly understood. It has been proposed that studying alternative phenotypes, such as tumor ggressiveness, may be a solution for overcoming the apparent heterogeneity that has hindered the identification of PCa susceptibility genes. Familiality analysis within the Utah Population Database (UPDB) has shown that localized, non-metastatic PCa demonstrates especially strong familiality, making this phenotype a good candidate for linkage analysis. The familial relative risks for second and third degree relatives of men with this phenotype are significantly greater than the corresponding relative risks observed for PCa in general. We present the results of a genome-wide linkage analysis for localized PCa susceptibility loci. We identified 115 subjects with localized PCa in 24 extended and nuclear families ascertained from the UPDB. Families included

between 2 and 9 affected subjects with genotype information. Parametric multipoint linkage statistics were calculated under dominant and recessive inheritance models for a genome-wide set of 401 microsatellite markers using the MCLINK software package. Preliminary results showed no significant linkage findings at the genome-wide level, although LOD scores greater than 1 were observed on chromosomes 2p, 9p, and 15q. These results are preliminary and do not represent all available data. Analysis of the complete data resource and selected strata is currently in progress.

63

Detection of the true disease susceptibility site in the presence of missing data

P. Croiseau(1), H.J. Cordell(2), E. Génin(1)
(1) INSERM U535 and Université Paris XI, Villejuif, FRANCE, (2) Institute of Human Genetics, Newcastle University, UK

To test for association between a disease and a set of linked markers, or to estimate disease risks, several different methods have been developed. Most methods require that individuals be genotyped at the full set of markers and that phase be reconstructed. Individuals with missing data are excluded from the analysis. This can result in an important decrease in sample size and a loss of information. It is also possible than another locus in linkage disequilibrium (LD) can become more significant than the true disease susceptibility site (DS). A possible solution to this problem is to use multiple imputation method to infer missing data. Briefly, the method consists in estimating from the available data all possible phased genotypes and their respective probabilities derived from the estimated haplotype frequencies. These posterior probabilities are then used to generate imputed data sets via one of a number of possible data augmentation algorithms. We performed simulations to examine, for different patterns of missing data, how often the true DS site gives the highest association score among different loci in LD. We found that multiple imputation usually allows a better detection of the true DS site even if the number of missing data is high. Multiple imputation presents the advantage of being much more easy to use and flexible than the other methods that deal with missing data in association tests. It is therefore a promising tool in the search for DS sites involved in complex diseases.

64

PedGenie 2.0: Meta Genetic Association Testing in Mixed Family and Case-control Designs

K. Curtin, J. Wong, K. Allen-Brady, N.J. Camp
Div. of Epidemiology, Dept. of Biomedical Informatics, University of Utah School of Medicine, USA

In the study of common diseases and genes with modest effects, large consortium and multi-center efforts hold the promise of increased power to detect associations, but also present analysis challenges. Studies differ geographically and ethnically, and considerable differences in case-control ascertainment and pedigree structures between resources are likely. Currently, no software package exists that allows association testing in mixtures of family-based and independent resources, between or within studies. PedGenie 2.0 (beta-version) extends the functionality currently available in PedGenie 1.2 (Allen-Brady et al. 2006, BMC Bioinformatics, 7:209) by incorporating meta statistics for combined analysis of multi-study resources, along with Monte Carlo significance testing which allows for a mixture of pedigree members and independent individuals. Briefly, study-specific allele or haplotype frequencies for markers of interest are used within studies. A Mendelian gene-drop simulation is performed independent of trait; each possible null genotype configuration is used to create an empirical null distribution for the significance testing of meta statistics. The currently incorporated meta statistics for genotype, composite genotype or haplotype analysis across studies are based on Cochran-Mantel-Haenszel (CMH) techniques to calculate odds ratios, chi-squared test of independence, and chi-squared test of trend. Future efforts will include meta-extensions for other quantitative and transmission statistics, such as the transmission-disequilibrium test (TDT). PedGenie is a flexible, easily implemented analysis tool that is enhanced significantly in beta version 2.0 by the incorporation of meta-statistics to allow valid combined analysis of multiple studies in the detection of genetic association with common disease.

65

Identifying secondary loci in existing genome scan data

E.W. Daw, S. Shete
Department of Epidemiology, MD Anderson Cancer Center, USA

In many genetic disorders in which a primary disease-causing locus has been identified, evidence for additional trait variation due to genetic factors has been found. These findings have led to several studies in which secondary modifier loci are being sought. Identification of such modifier loci provides additional insight into disease mechanisms and may provide additional treatment targets. We believe that some secondary loci can be identified by re-analysis of genome screen data while controlling for the effects of the primary locus. To test this hypothesis, we simulated multiple replicates of typical genome screening data on to two real family structures used to identify a genetic mutation causing hypertrophic cardiomyopathy. We simulated a trait with characteristics similar to one measure of hypertrophic cardiomyopathy. This trait was influenced by a primary gene, a secondary modifier gene, and a third very small effect gene. We examined the power and false positive rates to map the secondary locus while controlling for the effect of the primary locus with two types of analyses. First, we examine Monte Carlo Markov chain (MCMC) combined segregation and linkage analysis as implemented in Loki. For this method, we calculated two scoring statistics: LOP and l-score. Second, we calculate LOD scores using an individual-specific liability class based on the quantitative trait value. We find that both methods produce scores that are significant in some replicates, with the MCMC methods generally performing

better. We conclude that mapping of modifier loci in existing samples is possible with these methods.

66

Assessment of replication results from quantitative trait linkage analyses using Rochester Family Heart Study Phases I and II

M. de Andrade(1), Z. Ye(2), E.J. Atkinson(1), S.T. Turner(3)
(1) Biostatistics, Mayo Clinic, USA, (2) Mathematical Sciences, Michigan Technological University, USA, and (3) Nephrology and Hypertension, Mayo Clinic, USA

The lack of replication in linkage analysis may be due to different study design, ascertainment, population stratification, genotype calling. Between 1984 and 1991, 3978 members from 601 households underwent standardized medical interviews, physical examinations, and blood sampling at Mayo Clinic. This study is known as the Rochester Family Heart Study (RFHS), a community-based cross-sectional study of the genetic epidemiology of atherosclerotic coronary artery disease and essential hypertension. The RFHS Phase I consisted of 2135 participants from 279 randomly ascertained multigenerational pedigrees, and the RFHS Phase II consisted of 1809 participants from 252 randomly ascertained multigenerational pedigrees. Recently we published the first trivariate genome scan for quantitative linkage analysis using blood pressure measures and body mass index data from the Phase I RFHS. Only one region on chromosome 10 showed strong evidence of linkage (LOD=5.10, $p < 0.000044$) (Turner et al.; 2004). When the same analysis was performed in the Phase II RFHS, the linkage results were not replicated. In this paper, we will present linkages results from both phases and our assessment for the lack of replication.

67

Linkage genome scan and subsequent association studies show involvement of the vitamin D receptor gene in idiopathic short stature

A. Dempfle(1), S.A. Wudy(2), K. Saar(3), S. Hagemann(2), S. Friedel(4), A. Scherag(1), L.D. Berthold(2), G. Alzen(2), L. Gortner(5), W.F. Blum(2,6), A. Hinney(4), P. Nürnberg(7), H. Schäfer(1), J. Hebebrand(4)
(1) Univ. Marburg, (2) Univ. Gießen, (3) Max Delbrück Center, Berlin, (4) Univ. Essen, (5) Univ. Saarland, (6) Eli Lilly, (7) Univ. Cologne, all Germany

Stature is a highly heritable trait under both polygenic and major gene control. To identify genetic regions linked to idiopathic short stature (ISS) in childhood, we performed a whole genome scan in 92 families each with two affected children with ISS, including constitutional delay of growth and puberty and familial short stature. Chromosome 12q11 showed significant evidence of linkage to ISS and height (maximum non-parametric multipoint LOD scores 3.18 and 2.31 at 55-58 cM), especially in sister-sister pairs (LOD score of 1.9 for ISS in 22 pairs). The region on chromosome 12q11 had previously shown significant linkage to adult stature in several genome scans and harbors the vitamin D receptor gene, which has been associated with variation in height. A SNP (rs10735810, FokI), which leads to a functionally relevant alteration at the protein level, showed preferential transmission of the transcriptionally more active C-allele to affected children (p=0.04) and seems to be responsible for the observed linkage (p=0.05, GIST test). The C allele has an allele frequency of 0.63 and leads to estimated genotype relative risks of 1.3 and 1.9 for heterozygous and homozygous carriers.

68

Semiparametric Models for Linkage and Association Analyses of Quantitative Traits in Longitudinal Pedigree Studies

G. Diao(1) and D.Y. Lin(2)
(1) Dept. of Applied & Engineering Statistics, George Mason University, USA, (2) Dept. of Biostatistics, University of North Carolina at Chapel Hill, USA

Longitudinal pedigree studies provide a valuable resource for evaluating genetic effects on complex human traits over time. Most commonly used approaches for the genetic analyses of longitudinal quantitative trait data in human pedigrees pertain to the variance-component (VC) models. Virtually all the existing VC models for longitudinal data analysis assume that the trait values are normally distributed and are limited to linkage analysis. Violation of the normality assumption would inflate the type I error and reduce the power. We present powerful and robust VC methods for the linkage and association analyses of longitudinal quantitative traits. Our methods are based on semiparametric transformation linear models, which allow arbitrarily distributed quantitative traits. The proposed models consist of two parts, of which the mean part models the effects of environmental variables and the association between marker genotype and trait value, and the variance-component part models the correlations among trait values within a family as well as intrasubject correlation among multiple measurements per subject. We present efficient likelihood-based procedures to estimate unknown parameters and test for linkage and association. Extensive simulation studies showed that the new methods outperformed the existing ones in practical situations. We provide an application to the Framingham Heart Study data provided to Genetic Analysis Workshop 13. A computer program is freely available.

69

A Genomic Scan for Age at Onset of Alzheimer's Disease from the NIMH Genetic Initiative

MR Dickson(1), R CP Go(1), H Wiener(1) D Blacker(2), SS Bassett(3)
(1) Dept. of Epi, Univ. of AL, Birmingham, USA (2) Dept. of Epi, Mass. Gen. Hosp., USA (3) Dept. of Med., Johns Hopkins Univ., USA

We performed a linkage analysis for age at onset (AAO) in the Alzheimer's disease (AD) NIMH sample. The total 437 families were subset as late-onset (LO) N=320, AAO≥65 and early/mixed (EM) N=117, ≥1 member with AAO < 65. Treating AAO as a censored quantitative trait, we obtained the gender and APOE4

adjusted residuals in a parametric logistic model and then analyzed the residuals as the quantitative trait in variance component linkage analysis in SOLAR. For comparison AAO/Age at Exam (AAOE) was analyzed as a quantitative trait in SOLAR in a similar manner with affection status, gender, and APOE4 status as covariates. Heritability for residual and AAOE outcomes were 66.3% and 74.0% respectively for the total sample, 56.0% and 57.0% in the LO sample, 33.0% and 33.0% in the EM sample. Among the total sample, peaks for the residual outcome with LOD>1 were identified on chromosome (Chr) 1 at 50cM (LOD=1.31) and 189cM (LOD 1.97), Chr 2 at 64cM (LOD 1.02), and Chr 6 at 87cM (LOD 1.03); for the AAOE model peaks were found on Chr 1 at 190cM (LOD 1.41), Chr 3 at 133cM (LOD 1.03), and Chr 6 at 87cM (LOD 1.64). Among LO families peaks with LOD>1 were identified on Chr(s) 1, 6, and 10 for both models. Among EM families peaks with LOD>1 were identified on Chr(s) 1, 3, 10, and 12 for both models and on Chr(s) 2, and 8 for the residual model and Chr 9 for the AAOE model. Similarities and consistency of the models are discussed. Results suggest the genetics of AAO in AD is complex with many potential modifying genetic regions.

70

Heritability of Mammographic Breast Density in a Sample of Amish Women

J.A. Douglas(1), M-H Roy-Gagnon(1), C.V. Van Hout(1), A.M. Levin(1), D.S. McConnell(2), H.P. Chan(3), M.A. Helvie(3), B.D. Mitchell(4), A.R. Shuldiner(4)
(1) Depts. Of Human Genetics, (2) Epidemiology, and (3) Radiology, Univ. of Michigan, USA; (4) Dept. of Medicine, Univ. of Maryland, USA

Increased breast density (defined as the absolute or percentage of breast area that is occupied by dense tissue on a mammogram) is one of the strongest known but perhaps least understood breast cancer risk factors. We recently conducted a pilot study of the genetics of breast density in 287 related women from the Old Order Amish population of Lancaster County, PA, with the goal of estimating the heritability of breast density in this reproductively unique genetic isolate. Similar to values reported in other highly parous populations, mean absolute and percent breast density were 15 cm^2 (SD=9 cm^2) and 16% (SD=11%), respectively. After adjusting for significant covariates (age, menopausal status, and number of live births), we estimated the heritability of absolute breast density to be 0.46 (SE=0.14). Of the total variation in absolute density, covariates explained 21%, additive genetic factors accounted for 36%, and 43% remained unexplained. Similarly, after adjusting for significant covariates (age, menopausal status, number of live births, and waist circumference), the heritability of percent breast density was 0.43 (SE=0.15). Of the total variation in percent density, covariates accounted for 50%, additive genetic effects accounted for 22%, and 28% remained unexplained. These findings justify further study of this population to identify loci underlying variation in breast density.

71

New multipoint linkage statistic applied to Alzheimer disease data

E. Drigalenko
Dept. of Neuropsychiatry and Behavioral Sciences, Texas Tech University Health Sciences Center, Lubbock, USA

Liang et al. (2001) proposed an identity-by-descent (IBD) based method to estimate the location of an unobserved susceptibility gene framed by multiple markers within a chromosomal region. Based on this study, I found a specific IBD pattern on a chromosome that can be used to reveal candidate chromosomal regions for genes associated with the disease and proposed a simple multipoint regression-based statistic for genome-wide linkage scanning (Genetic Epidemiology 2005; 29: 244). Here I report the results of the genome scan by the new method.437 Alzheimer disease families were collected by the National Institute of Mental Health Genetics Initiative (Blacker et al. 1997). This data have been analyzed by Blacker et al. (2003) using FASTLINK and GENEHUNTER-PLUS.I applied the new method to the same affected sib pairs data as Blacker et al. (2003). Many of known genes are confirmed, including AGT, APOD, APOE, BCHE, OLR1, PNMT, SNCA. I found several candidate regions with $P < 0.05$, where no candidate genes have been reported: 3q26.1, 4p26, 4q31-q32, 5q14, 5q34-qter, 5q34-qter, 11q25-qter, 13q12, 18q12, 19q11-q12, and 21pter-21q21. These chromosomal regions are good targets to search new candidate genes for Alzheimer disease.

72

Application of robust score statistics to quantitative trait linkage analysis in the Framingham Heart Study

J. Dupuis(1), A. Manning(1), K. Cho(1), L.A. Cupples(1), J.B. Meigs(2,3), D. Karasik(3,4), D.P. Kiel(3,4), D. Siegmund(5)
(1) Biostat. Dept., Boston Univ. SPH., USA, (2) Mass. Gen. Hosp., USA, (3) Harvard Med. School, USA, (4) Hebrew SeniorLife Inst. for Aging Research, USA, (5) Dept. of Stat., Stanford Univ., USA

Linkage analysis of quantitative traits is often performed using a likelihood ratio (LR) statistic from a variance component model because of its flexibility. LR statistics can be applied to large pedigrees and extended to accommodate multivariate or longitudinal traits as well as gene x covariate interaction. A disadvantage of LR statistics is that they rely heavily on an assumption of multivariate normality, which can result in false positive results when violated. Use of robust score statistics based on nonparametric variance estimates has been proposed to circumvent this problem. We compared the LR and robust score statistics for continuous traits in the largest 330 Framingham Heart Study families. For approximately normally distributed traits, such as bone mineral density, both approaches gave similar results. However, for a highly non-normal trait like fasting plasma glucose, the LR statistic gave inflated evidence for linkage compared to the robust score statistic, even after log transformation of the trait. We found that incorporation of gene x covariate

interaction can enhance linkage signals, with an example of a continuous modifying covariate (body mass index effect on fasting plasma insulin) and of a dichotomous covariate (sex effect on bone mineral density). We also illustrate how the robust score test can be applied to ordinal phenotypes.

73

Evidence of interaction between DTNBP1 and IL3 in schizophrenia

Todd L Edwards(1), Xu Wang(2), Brandon Wormly(2), Brien Riley(2), F Anthony O'Neill(3), Dermot Walsh(4), Kenneth S Kendler(2), Marylyn D. Ritchie(1), Xiangning Chen(2)
(1) Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232. (2) Virginia Institute for Psychiatric and Behavioral Genetics and Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298. (3) The Department of Psychiatry, The Queens University, Belfast, Northern Ireland, UK (4) The Health Research Board, Dublin, Ireland

Schizophrenia is a psychotic mental disorder with a lifetime prevalence of approximately 1%. Heritability estimates of approximately 80% indicate this disease is largely genetic. Reports of risk loci and an epistatic model in the literature suggest a complex genetic etiology. In this study, we explored gene-gene interactions in population-based (657 cases: 414 controls) and family-based (269 families; 654 cases: 737 controls) datasets of English or Irish ancestry. The Multifactor Dimensionality Reduction Pedigree Disequilibrium Test (MDR-PDT) was used to explore epistasis in families. A highly significant 3-locus interactive model was identified. The 3-locus family interactive model was between genes IL-3, RGS4, and dystrobrevin-binding protein 1 (DTNBP1) (rs2069803, rs2661319 and rs2619539 respectively), p=0.006. To verify these findings, we used MDR on the case-control dataset containing the same markers typed in the 3 genes. While we could not replicate the 3-locus interaction, we saw evidence of a joint effect of IL3 and DTNBP1 with different markers (rs31400 in IL3 and rs760761 in DTNBP1), $p < 0.001$, odds ratio=1.25, 95% CI *0.95, 1.65*.

74

Power of Multifactor Dimensionality Reduction to detect epistatic interactions in simulated genetic data with up to ten-thousand SNPs

Todd Edwards, Scott Dudek, Marylyn Ritchie
Center for Human Genetics Research, Vanderbilt University

In genetic studies a common consideration for the investigator is how to detect joint effects at several variables in a modest sample size. The search space to find such associations is very large due to the combinatorial explosion when exploring gene-gene interaction effects. The Multifactor Dimensionality Reduction (MDR) algorithm searches these large spaces with an exhaustive approach and has been shown to have good power to detect interactive effects in modest size datasets on the order of tens of SNPs. An unknown aspect of the performance of MDR is the power to detect joint effects given large numbers of SNPs, many of which can be considered noise variables. To answer this question, a simulation study was conducted. Twenty purely epistatic two-locus and three-locus genetic models with a minor allele frequency of 0.2 or 0.4, and broad sense heritability ranging from 0.05 to 0.25 were simulated. All datasets consisted of 500 cases and 500 controls. For each genetic model, we simulated one hundred datasets for each of the following total numbers of SNPs: 100, 500, 1000, 5000, or 10,000 SNPs for each two-locus model, and 100 or 500 SNPs for each three-locus model. The results of this study demonstrate that MDR has excellent power to detect these interactive effects in datasets that exceed the largest candidate gene studies. The power ranged from 94%–100% for two-locus models and from 62% - 100% for three-locus models. Thus, MDR is a powerful analytical tool for large scale SNP studies of common, complex disease.

75

The impact of pedigree structure on heritability estimates

C. T. Ekstrom
Dept. of Natural Sciences, KVL, Denmark

Heritability measures the familial aggregation of a disease or trait and a non-zero heritability suggests that a genetic component may be present. Reliable heritability estimates are necessary in the planning phase of a linkage or genetic association study but often these estimates are obtained from other studies where the composition of pedigrees may be different from the study that is prepared. The impact of pedigree structure on precision and accuracy of narrow and broad sense heritability estimates of quantitative traits is examined for data both with and without a dominant genetic effect. We find that all but the simplest pedigree structures provide the same information about the narrow sense heritability of a quantitative trait but that the pedigree structure and sample size has a substantial influence on the bias and precision of broad sense heritability estimates.

76

Lipid and dietary factors confound detection of a QTL on 2p24-p25 for adiposity: the NHLBI-Family Heart Study

M.Feitosa(1), K. North(2), R.H.Myers(3), J.S. Pankow(4), I.B.Borecki(1)
(1) Washington U., St. Louis, MO; (2) U.N.C., Chapel Hill, NC; (3) Boston U., Boston, MA; (4) U. Minnesota, Minneapolis, MN

We sought to identify loci influencing adiposity traits by linkage analysis using phenotypes with different covariate adjustment strategies. 988 sibs (243 autosomal STR markers) and 2,666 pedigree members (402 markers) were genotyped. Suggestive evidence of a QTL on 2p24-p25 influencing age-sex-adjusted body mass index (BMI: LOD=1.5) and waist circumference (WAIST: LOD=1.2)

was found. When Willets dietary factors and lipid were added to the age-sex covariate adjustment, the evidence of linkage was significantly enhanced (BMI: LOD=3.5, FDR-p=0.0003, WAIST: LOD=4.2, FDR-p=0.000086). To verify whether these increases were attributable to covariate effects per se or due to sample size changes, we repeated these analyses for the subset of subjects with no missing covariate values. In this restricted sample, the linkage signals for both age-sex adjusted BMI and WAIST were lower than in the fully adjusted models (LODs of 2.5 and 2.1, respectively). Thus, the LODs for BMI and WAIST were approximately 1 LOD units higher when adjusting for dietary and lipid related factors. Our results suggest: (i) evidence of a QTL on 2p24-p25 at marker D2S2952 influencing BMI and WAIST; (ii) dietary factors (carbohydrates, fats and vitamins) and lipids may confound the identification of QTLs; and (ii) adjusting BMI and WAIST for these covariate effects may aid in the identification of genes influencing obesity. Prominent candidate genes reside on 2p24-p25, including POMC and LPIN1.

77

Genetic determinants of pulse pressure in American Indians: the strong heart study

N Franceschini, SA Cole, S Laston, KM Rose, S Rutherford, HHH Goring, V Diego, JW MacCluer, ET Lee, LG Best, BV Howard, RR Fabsitz, MJ Roman, KE North
UNC, Chapel Hill, NC, SFBR, San Antonio, TX, U of OK, Oklahoma City, OK, MBTL, SD, MedStar, Washington, DC, NHLBI, Bethesda, MD, Weill Medical, New York, NY

Pulse pressure, a measure of central arterial stiffness and a predictor of cardiovascular mortality, has been demonstrated to have a genetic basis. We conducted a genome-wide scan of pulse pressure in 1892 participants of the Strong Heart Family Study (SHFS), an observational study of American Indian tribes recruited in Arizona, North and South Dakota, and Oklahoma. Blood pressure was measured three times and the average of the last two measures was used for analyses. Pulse pressure, the difference between systolic and diastolic blood pressures, was log-transformed to reduce kurtosis and adjusted for the effects of age within study center and sex. Variance component linkage analysis (implemented in SOLAR) was performed using marker allele frequencies derived from all individuals and multipoint IBDs calculated in Loki. We identified a quantitative trait locus (QTL) influencing pulse pressure on chromosome 7 at 37 cM (marker D7S493, LOD=3.3) and suggestive evidence of linkage on chromosome 19 at 92 cM (LOD=1.8). Adjusting for diabetes, body mass index, log urine albumin-to-creatinine, smoking status and hypertension treatment reduced the chromosome 7 QTL LOD score to 2.6. This signal on 7p15.3 overlaps with positive findings for pulse pressure among Utah population samples, suggesting that this region may harbor gene variants for blood pressure.

78

Reproducibility of Genotype Data Using the Affymetrix GeneChip® 100K Human Mapping Array Set

B.L. Fridley(1), S.T. Turner(2), K.R. Bailey(1)
(1) Division of Biostatistics, Mayo Clinic, USA, (2) Nephrology and Hypertension, Mayo Clinic, USA

In genomic studies, one is often faced with the problem of genotyping errors which go undetected. To address issues of genotyping errors, many studies include blind duplicates to be genotyped twice as a means to assess the error rate in genotype calls. In addition to the planned duplicates, chips may be re-run due to low call rates. This, then presents an additional set of duplicate data that can be use to assess and investigate mismatch in genotype calls between the two runs or replicates. Assessment of genotype errors is an even bigger issue when one is dealing with a large number of SNPs measured for a genome wide association study. Results from a genome-wide association study using Affymetrix GeneChip® 100K Human Mapping Array Set using the Dynamic Modeling calling algorithm for a pharmacogenomics study investigating genetic determinants for response to antihypertensive drug for two sets of duplicate data (blind duplicates, and non-random duplicates) will be presented along with result for modeling the probability of a mismatch in genotype calls as a function of the confidence of the calls.

79

A QTL mapping method in large pedigrees

GAO, G
University of Alabama at Birmingham

A QTL mapping method in large pedigrees G. Gao University of Alabama at Birmingham, AL 35294, USA. A large pedigree dataset always consists of two parts: upper ancestral generations with no genotype and phenotype data (part A) and recent generations with marker and phenotype data (part B). When the pedigree size is large enough, calculating identity by descent (IBD) probabilities for the entire pedigrees or for part B is time consuming or infeasible. In quantitative trait loci (QTL) mapping on this large pedigree, a traditional method is to split the pedigree into small subunits. This method can cause the loss of power in detecting QTL. We propose a QTL mapping method for a large pedigree based on calculating IBD matrices at putative QTL for part B by using multiple marker haplotype configurations and calculating the (additive) relationship matrix for polygenic effects by using the entire pedigree structure (parts A and B). We use the variance component linkage analysis which can be implemented in software ASR eml and Solar. Simulation studies show the proposed method has higher power in QTL detection than that of the method splitting the pedigree into small units.

80

Testing For Allelic Association Based On Population-based Quantitative Trait Data: ANOVA Versus Model-free Alternatives

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

The paradigm of allelic association in the context of quantitative traits is based on the intuitive concept of differences in allele frequencies between individuals having high values of the quantitative trait and those with low values of the trait. While some novel family-based association methodologies for quantitative traits have been developed, population-based quantitative data have usually been analyzed using classical analysis of variance (ANOVA) methods. However, ANOVA is valid in a strict statistical sense only under the assumption of equality of variances in each underlying group. On the other hand, the assumption of equality of variances of the quantitative traits at the different QTL genotypes is genetically unrealistic. Using Monte-Carlo simulations, we find that the rates of both false positives and false negatives can be adversely affected even if the significance levels are evaluated empirically. Thus, it is of interest to explore for model-free alternatives that would circumvent this problem. We propose two methods: (i) a quantile-based regression and (ii) a goodness of fit chi-square to test for allelic association. The basic paradigm of both the methods is that a marker allele in linkage disequibrium with an allele at the QTL will have a non-uniform frequency distribution across the range of quantitative trait values. We perform extensive simulations for different genetic parameters to assess the power of the proposed methods. We find that the power of the quantile-based regression method is higher than that based on the chi-square test.

81

Genome-wide linkage screen for Waldenström macroglobulinemia susceptibility loci in high-risk families

L.R. Goldin(1), M.L. McMaster(1), Y. Bai(1), M.Ter-Minassian(1), S. Boehringer(2), M.A. Tucker(1)
(1) Genetic Epidemiology Branch (2) Biostatistics Branch, DCEG, NCI, Bethesda, MD

Waldenström macroglobulinemia (WM) is a distinctive subtype of non-Hodgkin lymphoma featuring overproduction of immunoglobulin M (IgM) and has a clear familial component. No susceptibility genes have been identified. We performed a genome-wide linkage analysis in eleven high-risk WM families having 122 individuals with DNA samples, including 34 WM cases and 10 cases of IgM monoclonal gammopathy of undetermined significance (IgM MGUS) which is a potential precursor to WM. We genotyped 1058 microsatellite markers (average spacing=3.5 cM), performed both non-parametric and parametric linkage analysis. The strongest evidence for linkage was found on chromosomes 1q and 4q when both WM and IgM MGUS cases were considered affected; non-parametric linkage scores were 2.5 (p=0.0089) and 3.1 (p=0.004), respectively.Other locations suggestive of linkage were found on chromosomes 3 and 6. We used the program, Genehunter 2L to perform 2-locus non-parametric linkage analysis in suggestive regions. The two-locus results were consistent with independent effects at each locus. The regions identified on chromosomes 3 and 4 were also identified in a similar genome scan of families with Hodgkin lymphoma, suggesting that common susceptibility genes affect distinct B-cell malignancies. The findings from this first linkage analysis of high-risk WM families represent important progress toward identifying gene(s) that modulate susceptibility to WM and toward understanding its complex etiology.

82

MAFs and LD within the ADH Gene Cluster: Comparison of CEPH Samples and a Control Study Population

EL Goode, TA Sellers, J Schildkraut, C Phelan, E Iverson, LC Hartmann, Y Huang, L Kelemen, JE Olson, JM Cunningham, M Liebow, RA Vierkant, DN Rider, ZS Fredericksen, B Calingaert, L Fan, VS Pankrat
Mayo Clinic; H Lee Moffitt Cancer Center; Duke University

SNP selection assumes similarity of minor allele frequencies (MAFs) and linkage disequilibrium (LD) between resesequenced samples and the study population. We selected intra-genic tagSNPs within alcohol dehydrogenase (ADH) genes on 4q21-23 (ADH1A, 1B, 1C, 4, 6) using data from 22 unrelated Utah residents with European ancestry (CEPH, NIEHS SNPs) and the algorithm of LDSelect (r^2>0.8, MAF>0.05, AJHG 74:106–20, 2004). We genotyped 41 SNPs in 451 European-American women serving as controls in an ongoing ovarian cancer study. Control MAFs were similar to CEPH MAFs (mean difference -0.005; range -0.119 to 0.070); only 1 comparison yielded p-value <0.05. There was slightly less LD among controls (HAPLOVIEW, r^2 difference mean -0.021; range -0.519 to 0.386); comparison of 820 pairwise r^2s resulted in 21 p-values <0.05 (not more than expected by chance). Intra-genic LD (157 SNP pairs) was expected to be $r^2 < 0.8$ in controls; this was true for all but two pairs (0.773 v. 1.000; 0.780 v. 0.955). LD between 663 inter-genic SNP pairs was also examined. Of 654 pairs with CEPH $r^2 < 0.8$, 8 pairs (1%) had control $r^2 > 0.8$, suggesting under-binning and loss of efficiency. Of 9 pairs with CEPH $r^2 > 0.8$, 2 pairs (22%) had control $r^2 < 0.8$ (0.869 v. 0.759; 0.861 v. 0.565), suggesting over-binning and loss of information. These data suggest that critical assessment of the appropriateness of samples used for SNP selection is needed.

83

Map-Misspecification and an Unknown Genetic Model in Multipoint Linkage Analysis: An Evaluation of the Sex-Specific Multipoint PPL, HMOD and MMLS

M. Govil(1), M.W. Logue(2), V.J. Vieland(3)
(1) U of Pittsburgh, PA, (2) U of Iowa, IA, (3) CCRI, Columbus, Ohio

Multipoint linkage analyses normally utilize a sex-averaged genetic map (SA) although the true sex-specific maps (SS) are of different lengths. For complex traits, where the true underlying trait parameters are never known a priori with any certainty, there are several parametric approaches which allow for an unknown genetic model. These approaches include maximization (e.g., a LOD score maximized over a dominant and a recessive model (MMLS), a heterogeneity LOD score fully maximized with respect to all of the elements of the genetic model

(HMOD)) and integration over the parameter space, as with the posterior probability of linkage or PPL. However, the impact of map misspecification, due to the assumption of SA, on multipoint analyses with such approaches is yet to be studied. In a series of simulations, 50 replicates each of nuclear pedigrees and of sibpairs were generated with 3 markers at 0.05:10 Kosambi cM inter-marker distance on the male:female map and the trait locus at 0 recombination to the middle marker on both maps. Comparing the linkage signal under SA and SS showed that assumption of SA has little or no impact on magnitude, distribution, or localization information from any of the three statistics. These results suggest that the assumption of SA may be made without loss for analyses even when the true underlying map is sex-specific. They also agree with findings for 2-point analyses that allowing for sex-specific recombination rates does not improve 2-point PPL performance (Logue & Vieland, 2005).

84

Which alternative to the biased allelic test in case-control association studies

M. Guedj(1,2), E. Della-Chiesa(2), K. Forner(1), J. Wojcik(1) and G. Nuel(2)
(1) Serono, (2) Statistic and Genome Laboratory

Association studies are traditionally performed in the case-control framework. As a first step in the analysis process, comparing allele frequencies using the Pearson's chi-square statistic is often invoked. However such an approach assumes the independence of alleles under the hypothesis of no association, which may not always be the case. Consequently this method introduces a bias that deviates the expected type-I error-rate. In this article we first propose an unbiased and exact test as an alternative to the biased allelic test. Available data require to perform thousands of such tests so we focused on its fast execution. Since the biased allelic test is still widely used in the community, we illustrate its pitfalls in the context of genome-wide association studies and particularly in the case of low-level tests. Finally, we compare the unbiased and exact test with the Cochran-Armitage test for trend and show it perfoms similarly in terms of power. The fast, unbiased and exact allelic test is available in R, C++ and Perl at: http://stat.genopole.cnrs.fr/software/fueatest.

85

Variance Components Analysis of the Electrocardiogram phenotype QTc interval in nuclear families from the general population

C Hajat(1,3), PR Burton(1,3), A Ng(2), J Gracey(2), T Smith(2), PW Macfarlane(4), NJ Samani(2), MD Tobin(1,3)
(1) Department of Health Sciences, (2) Department of Cardiovascular Sciences, (3) Department of Genetics, University of Leicester, UK, (4) Medical Sciences, University of Glasgow, UK

The heart-rate corrected QT interval (QTc) is associated with cardiovascular morbidity and mortality. Improved understanding of the genetic and environmental determinants of the QTc interval will ultimately contribute to the reduction of sudden cardiac death and related cardiovascular disorders. Our aim is to investigate the influence of these determinants in a UK population based-sample of 1491 subjects from 386 representative white European families. The mean QTc interval, determined from a resting 12 lead electrocardiogram using an automated technique, was 404.7ms (95% CI 402.8ms-406.7ms) and was higher in females (410.1ms) compared with males (399.3ms) and in parents (409.7ms) compared with their offspring (399.6ms) ($p < 0.001$). Using Gibbs sampling-based variance component models implemented in WinBUGS 1.4, we estimated that the proportion of the variance in the QTc interval attributable to additive polygenic effects (narrow sense heritability or h2N) was 41.8% (95% CI 22.7%–59.8%). Our findings, some of which we believe to be unique, would suggest that there is considerable utility in continuing to search for the genetic variants that underlie this clinically important trait. Our ongoing work includes studying the association of variants in candidate regions for the QTc interval.

86

BRCA penetrance for breast and ovarian cancers: a heterogeneity study

S.Hassid(1), C.Noguès(2), J.Carayol(3), F.Alarcon(3), A.Mohamdi(3), M.Labbé(1), A.Rezvani(1), D.Stoppa-Lyonnet(4), P.Berthet(5), JP.Fricker(6), N.Andrieu(1), C.Bonaïti-Pellié(3)
(1) InsermU794-Dept Biostat, Institut Curie,Fr, (2) Ctre R Huguenin,Fr, (3) InsermU535,Fr, (4) Institut Curie,Fr, (5) Ctre F Baclesse,Fr, (6) Ctre P Strauss,Fr

Heterogeneity in mutations and/or familial factors might cause a variation in BRCA penetrance estimates. We estimated BRCA mutations penetrances for either breast (BC) or ovarian cancer (OC) in a French sample of 650 families fulfilling the criteria for genetic testing. We studied heterogeneity by mutation type and region. The statistical method used was developed by Carayol et al. (Genetic Epid, 27,2004,109) and based on a retrospective likelihood approach correcting for ascertainment of families on multiple affected relatives and a mutated proband. Penetrances of BRCA1 and BRCA2 were estimated by age 70 at 42% and 50% for BC , and 7% and 1% for OC. BC penetrance among BRCA1 carriers was higher when associated with truncating than with missense mutations (52% vs 16%), and, among truncating, higher when associated with those not subjected to nonsense-mediated mRNA decay (NMD) than with those subjected to NMD (70% vs 48%). The OC cluster region previously described on BRCA2 (Gayther et al, Nature Genetic, 15,1997,103) was not confirmed in this sample, indeed the OC cluster was found outside this region. Variation in penetrances were also studied by family characteristics. To conclude, we found penetrance estimates in the lower bounds of those published so far and showed variation factors which, if confirmed, might be used to modulate medical survey of carrier women.

87
Independence of Sib-pair IBD allele sharing between sib-pairs and its application
Q He
Pacific Health Research Institute, Honolulu, Hawaii, USA

The mean test, a non-parametric linkage method, uses the IBD sharing to detect the linkage between disease and genes. Originally, the estimated variance based on the observed IBD allele sharing was used in the calculation of the variance of IBD sharing and student t-test was used in the analysis. Recently, McQueen et al suggested using a boot-strap method to calculate the variance of IBD sib-pair sharing. In 2000, I proved that the IBD sharing between sib-pairs are ndependent, which can be used to deduct a simple way to calculate the variance of IBD sib-pair sharing statistic. Here I will show that, after adjust for the non-informative sib-pair and parents genotype combination, this simple method can be used to calculate the statistic used in the mean test. Also, I will discuss the possible extension of this result to other type of relative pairs and the possibility to combine IBD sharing of different relative pairs in one test statistic.

88
Genetic architecture of the APM1 gene and its influence on adiponectin plasma levels and parameters of the metabolic syndrome in 1727 healthy Caucasians
I.M. Heid(1,2)*, S.A. Wagner(1,3)*, H. Gohlke(1), B. Iglseder(3), J.C. Mueller(4), P. Cip(3), G. Ladurner(3), R. Reiter(3), A. Stadlmayr(3), V. Mackevics(3), T. Illig(1) , B. Paulweber(3), F. Kronenberg(5)* contributed equally
(1) GSF, Neuherberg, Germany, (2) LMU, Munich, Germany (3) University of Salzburg, Austria (4) Technical University, Munich, Germany (5) Innsbruck Medical University, Austria

The associations of the adiponectin (APM1) gene with parameters of the metabolic syndrome are inconsistent. We performed a systematic investigation based on fine-mapped SNPs in a particularly healthy population of 1727 Caucasians avoiding secondary effects from disease processes. Genotyping 53 SNPs (average spacing of 0.7kb) in the APM1 gene region in 81 Caucasians revealed a two-block LD structure and enabled comprehensive tagSNP selection. We found particularly strong associations with adiponectin levels for 11 of the 15 tagSNPs in the 1727 subjects (five p-values < 0.0001). Haplotype analysis provided a thorough differentiation of adiponectin levels with 9 of 17 haplotypes showing significant associations (three p-values < 0.0001). No significant association was found for any SNP with the parameters of the metabolic syndrome. We observed a two-block LD structure of APM1 pointing towards at least two independent association signals, one including the promoter SNPs and a second spanning the relevant exons. Our data on a large number of healthy subjects suggests a clear modulation of adiponectin concentrations by variants of APM1, which are not merely a concomitant effect in the course of type 2 diabetes.

89
The MC4R 103I Allele is Associated with Features of the Metabolic Syndrome in the Population-Based KORA Study
IM Heid(1), C Vollmert(1), F Kronenberg(2), C Huth(1), D Ankerst(3), A Luchner(4), A Hinney(5), G Brönner(5), H Löwel(1), HE Wichmann(1), T Illig(1), A Döring(1), J Hebebrand(5)
(1) GSF, Neuherberg, Germany (2) GenEpi,Innsbruck Medical Univ. (3) LMU,Munich (4) Univ. klinikum,Regensburg (5) Univ.Duisburg-Essen

The melanocortin-4-receptor (MC4R), part of the melanocortinergic pathway, controls energy homeostasis. The V103I (rs2229616) polymorphism was shown to be associated with decreased body mass. We analyzed this polymorphism in 7888 adults of the MONICA/KORA cohort with features of the metabolic syndrome (metS), along with exploratory mediator analyses incorporating pulse rate and life style factors. Subjects with the rs2229616 G/A genotype (frequency 3.7%) exhibited significantly decreased waist circumference (1.46 cm, p=0.021), decreased HbA1c (-0.09%, p=0.040), and increased HDL-cholesterol (+1.76 mg/dl, p=0.056), but no change in blood pressure. The odds of having 3 of these four components of the metS was substantially reduced among G/A subjects (OR=0.46, p=0.003). The findings were consistent for both surveys and both genders. There was no role of pulse rate, physical activity, smoking, alcohol, or fat intake frequency as a mediator for the association between the variant and obesity or metS. Only high frequency of carbohydrate intake showed borderline association with the polymorphism (OR=1.26, p=0.06), but did not qualify as mediator. Our study extends the role of the MC4R 103I variant from the previously described association with body mass to an association with features of the metS in a large representative sample, which is in line with the described functional mechanism.

90
Functional analysis of candidate genes associated with hypertension
C. Hicks
Department of Preventive Medicine and Epidemiology, Loyola University Medical Center. 2160 S. First Ave, Maywood, IL 60153. USA

Human hypertension is a classic example of a complex, multifactorial, polygenic disease with a major impact on public health worldwide. Several strategies have been developed over the last decade to dissect genetic determinants of hypertension. Of these, the most successful have been studies that identified rare Mendelian syndromes in which a single gene mutation causes high blood pressure. However, rare syndromes with Mendelian inheritance account for a very small fraction of the pathological human blood pressure variation. Since the completion of the Human Genome Project, the search for genes associated with hypertension has been accelerated, and remarkable progress has been achieved. Over 200 hypertension candidate genes have been discovered. However,

the functions of many of the identified genes are not known. In order to understand the genetic basis of hypertension, it is imperative that we identify genes and classify them according to their molecular functions and biological processes in which they are involved. The objective of this study was to elucidate the molecular functions of hypertension candidate genes using computational approaches. We classified 220 hypertension candidate genes according to their molecular functions and biological process in which they are involved. The functional classes identified included, apolipoproteins, adhesion molecules, endothelins, angiotensin steroids, transporters, lipid and glucose regulators, pituitary axis, intracellular messengers, Nitriuretic peptides, members of the kallikrein pathway, thromboxins, prostaglandins and genes involved in the sympathetic nervous system. Further analysis based on multiple sequence alignment reveled that most of the identified genes are evolutionary conserved between species.

91

Powerful, conservative and robust family-based association test: PSEUDOMARKER

T. Hiekkalinna(1,2), J.D. Terwilliger(2,3,4,5,6)
(1) Dept of Mol Med, National Public Health Inst, Finland, (2) Finnish Genome Center, Univ of Helsinki, Finland, (3) Dept. of Genetics and Devel, Columbia Univ, NY, USA, (4) Dept of Psychiatry, Columbia Univ, NY, USA, (5) Columbia Genome Center, Columbia Univ, NY, USA, (6) Division of Medical Genetics, NY State Psychiatric Inst, NY, USA

PSEUDOMARKER is joint Linkage and/or LinkageDisequilibrium software for Qualitative traits in the presence of errors and unknown model parameters. PSEUDOMARKER can utilize different pedigree sizes jointly such as cases and controls, trios, sib pairs, sib ships and multi generational families,which leads more robust and powerful tests. PSEUDOMARKER uses 'direct search' in ILINK program from the FASTLINK 4.1P package to maximize likelihoods over allele frequencies under null hypothesis of Linkage (HO), presence of Linkage (H_1), haplotype frequencies conditional on disease under null hypothesis of no Linkage and LD (H_2) and presence of Linkage and LD (H3). Then multiple statistics can be calculated as log-likelihood ratio tests: Linkage=log10 (H_1/H_0), LD given Linkage=log10 (H_1/H_3), LD given no Linkage=log10 (H_2/H_0), linkage given LD=log10 (H_3/H_2) and joint test of Linkage and LD. PSEUDOMARKER statistical approach provides a consistently more conservative, robust, and powerful way to conduct joint linkage and association analysis on heterogeneous data structures than the other statistical approaches in wider use at the present time, especially when there is linkage but no LD, for which many other methods have enormously high type I error rates, and typically lower power than PSEUDOMARKER over a wide variety of models considered.

92

Classification and Regression Tree Analysis for Polygenic Risk Stratification

B.D. Horne(1,2), J.L. Anderson(1,3), N.J. Camp(2)
(1) CV Dept, LDS Hospital, (2) Genet Epidemiol Div, Univ of Utah, (3) Cardiology Div, Univ of Utah. Salt Lake City, UT

The study of genetic association for common, complex diseases such as myocardial infarction (MI) may be plagued by allelic and locus heterogeneity and by epistasis. Methods that reduce dimensionality of genetic analysis from evaluation of all single nucleotide polymorphism (SNP) interactions to a few multi-SNP variables may improve association analysis. Classification and regression tree (C&RT) analysis may be such a method. C&RT tree-validation methods were studied with 8 biallelic SNPs in genes for matrix metalloprote inase (MMP)-1, MMP-2, MMP-3, MMP-9 (2 SNPs) and MMP tissue inhibitor (TIMP)-1, TIMP-2, and TIMP-3. Patients (N=5169) with MI (n=1709) were compared to MI-free patients. Tree growing strategies included: unvalidated trees, cross-validated trees with 2, 5, 10, or 25 folds, and split-sample trees with random patient assignment to a training set (60%) or test set (40%). The unvalidated and all cross-validated trees were the same, but the split-sample method (repeated 10 times) failed to replicate itself or the unvalidated tree (in the 10 trees, depth ranged from 4–6 nodes, terminal nodes from 6–18, and mean terminal node depth from 3.9–5.4). All trees resulted in significant ($p < 0.01$) global stratification of MI risk. Sex-specific results were similar. This study suggests that C&RT may be a powerful method for extraction of multi-SNP genotype groupings. Automated methods of tree validation, however, may not be useful. Operator-based tree pruning and extra-population tree validation deserve further evaluation.

93

A score statistic for linkage analysis of censored outcomes applied to a candidate region at chromosome 4 for human longevity

J.J. Houwing-Duistermaat(1), A. Callegaro(1), M.Beekman(1), R.G.J. Westendorp(2), E. Slagboom(1), J.C. van Houwelingen(1)
(1) Dept. of Med Stat and Bioinfo, (2) Dept. of Int Med, LUMC, Leiden, Netherlands

Tan et al used the non-parametric linkage (NPL) score for affected sib pairs to study the power to detect linkage for longevity in sib pairs. By using this statistic the information available in the current ages of the sibs is not used. We derived a score statistic to test for linkage based on the gamma frailty model. It appeared to be a weighted NPL statistic with weights equal to the product of the cumulative hazards of the siblings. Simulations show that this new statistic has more power than the unweighted NPL statistic. For longevity Puca et al identified a region at chromosome 4 using sibling pairs with one sibling above 98 years (lod score of 3.5). Using the Kong and Cox statistic, Beekman et al analysed this region in Dutch long-lived sibling pairs both older than 90 years. No evidence for linkage was found (lod score of 0.02). We reanalysed these data using the new score statistic. Further to take into account reduced marker informativeness, we used simulations to compute the variance of the statistic using our program (http://www.msbi.nl). Again no evidence for linkage was found (lod score of 0.1). Our conclusion is that the score statistic based on the gamma frailty model is a

powerful tool for genome wide linkage analysis of longevity aiming to identify new regions. Beekman et al. J Gerontol A Biol Sci Med Sci. 61 (2006):355–62 Puca et al Proc Natl Acad Sci USA. 98 (2001):10505–8 Tan et al Gen Epi. 26 (2004) 245–53.

94

A quantitative trait loci-specific gene-by-sex interaction for age of diabetes onset: The San Antonio Family Heart Study

KJ Hunt[1], HHH Göring[2], DM Lehman[3], L Almasy[2], BD Mitchell[4], J Sung[2], S Cole[2], T Dyer[2], J Blangero[2], JW MacCluer[2], MP Stern[3]
[1]MUSC, Charleston SC; [2]SFBR and [3]UTHSCSA, San Antonio TX; [4]UM, Baltimore MD

While the prevalence of diabetes is similar in men and women, there are sex differences in risk factors for diabetes. Therefore, we performed a genome-wide linkage scan to localize diabetes susceptibility genes, and tested the possibility of a gene-by-sex interaction in 46 randomly ascertained, extended, Mexican American pedigrees with 859 women and 591 men. Participants were examined at baseline and ~5.1 years later. Information from a participant's most recent exam was used. Prevalence of diabetes was 16.9% in women and 15.5% in men. For all analyses we used a 10 cM genetic map and a variance decomposition method implemented in SOLAR. Using the quantitative Martingale residual obtained by modeling age of diabetes diagnosis with Cox proportional hazard models in SAS, we performed variance component linkage analysis in the full sample ignoring and accounting for gene-by-sex interaction. Controlling for birth decade and BMI, the strongest evidence for linkage was on chromosome 1q at marker D1S1609, with a MLOD score of 2.34. This signal substantially improved when accounting for gene-by-sex interaction, with a MLOD score of 4.38 on chromosome 1q at marker D1S1609. In exploratory sex-stratified analyses, the MLOD on chromosome 1q at marker D1S1609 was 0.68 in women and 3.76 in men. We have identified a locus on chromosome 1q harboring an allele(s) that appears to affect diabetes risk with differential effects between men and women.

95

Association Mapping by Generalized Linear Regression with Density-based Haplotype Clustering

R.P. Igo, Jr.(1), J. Li(2) and K.A.B. Goddard
Depts. of (1) Epi. and Biostat. and (2) Electrical Eng. and Computer Sci., Case Western Reserve Univ., Cleveland OH USA

Haplotypes of closely linked single-nucleotide polymorphisms (SNPs) contain more information than individual SNPs about the underlying genetic structure, and thus may provide greater power to detect associations between genetic variants and disease phenotypes. We present a novel method for association mapping in which the dimensionality of a score test for association is reduced through density-based clustering of haplotypes. We implemented the haplotype clustering criteria of Li and Jiang (2005; *Bioinformatics* **21,** 4384) to reduce the number of coefficients in the generalized linear model (GLM) approach of Schaid et al. (2002; *Am. J. Hum. Genet.* **70,** 425). A flexible haplotype similarity score, a generalization of previously used measures, forms the basis for grouping haplotypes of probable recent common ancestry. All haplotypes within a cluster are assigned the same regression coefficient within the GLM, and evidence for association is assessed with a score statistic. Results of simulation studies demonstrated that clustering improved the power of the score test to detect association, over a variety of conditions, while preserving valid type I error. The increase in power was the most dramatic in the presence of high haplotype diversity, although a slight improvement was often observed at low diversity. In addition, we assessed the sensitivity of the test to variation in user-defined input parameters defining the haplotype similarity score, to develop guidelines for selecting these parameters.

96

Multivariate Combined Linkage and Association Mapping of Quantitative Trait Loci

Jeesun Jung (1), Ruzong Fan (2), and Lian Liu (2)
(1) Department of Human Genetics, University of Pittsburgh, Graduate School of Public Health, A300 Crabtree Hall, 130 DeSoto Street, Pittsburgh, PA 15261; (2) Department of Statistics, Texas A&M University, 3143 TAMUS, College Station, TX 77843

In this report, variance component models are proposed for high resolution multivariate combined linkage and association mapping of quantitative trait loci (QTL), based on both pedigree andpopulation data. Suppose that a quantitative trait locus is located in a chromosome region which exerts pleiotropic effects onmultiple quantitative traits. In the region, multiple markers such as single nucleotide polymorphisms (SNPs) are typed. Tworegression models, ''genotype effect model'' and ''additive effect model'', are proposed to model the association between the markers and the trait locus. The linkage information, i.e., recombination fractions between the QTL and the markers, is modeled in thevariance and covariance matrix. By analytical formulae, we showthat the ''genotype effect model'' can be used to model the additive and dominant effects simultaneously; the ''additiveeffect model'' only takes care of additive effect. Based on the two models, F-test statistics are proposed to test association between the QTL and markers. The non-centrality parameter approximations of F-test statistics are derived to make power calculation and comparison.

97

Interacting Genetic Factors Influencing Fasting LDL and HDL Response to Fenofibrate

RJ Kelly(1), DK Arnett(2), JA Smith(1), JM Ordovas(3), YV Sun(1), PN Hopkins(4), JE Hixson(5), JM Peacock(6), SLR Kardia(1)
(1) Dept of Epid, Univ of MI, (2) Dept of Epid, Univ of AL-Birmingham, (3) Nutrition and Genomics Lab, Tufts Univ,

(4) Cardiovas Genetics Res, Univ of UT, (5) Human Genetics Ctr, Univ of TX, (6) Dept of Epid, Univ of MN

There is a limited understanding of how gene-gene (epistatic) and gene-environment interactions affect low density ipoprotein (LDL) and high density lipoprotein (HDL) metabolism and their response to fenofibrate (FF), a triglyceride lowering drug. FF has also been shown to raise HDL levels and lower LDL levels in some patients. As part of the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study, 792 subjects from 168 families were assessed for their fasting LDL and HDL response to FF treatment. We genotyped 91 SNPs in 25 candidate genes and examined their association with response, both alone and in combination with clinical covariates. We applied a four-fold cross-validation scheme, repeated ten times, to each significant association for verification. We examined the patterns of fasting FF response via a multi-dimensional visualization system, paying particular attention to epistatic and gene-environment interactions and found little overlap between the interactions that affect LDL and HDL. LDL is primarily affected by interactions involving SNPs in APOA4, APOB, NOS3, and LRP1, while HDL is primarily affected by SNPs in APOA1, APOA4, APOC3, and ABCG8. These patterns give insight into this metabolic complexity and allow identification of patients likely to benefit from FF treatment.

98

KGraph: Visualizing and Evaluating Complex Genetic Associations

RJ Kelly, DM Jacobsen, YV Sun, JA Smith, and SLR Kardia
Dept of Epid, School of Public Health, Univ of Michigan, Ann Arbor, MI

The KGraph is a data visualization system that displays the complex relationships between the univariate and bivariate associations among an outcome of interest, a set of covariates, and a set of single nucleotide polymorphisms (SNPs). Eight graphical regions each display the results from a set of statistical tests and are composed of cells showing the results of individual tests. The KGraph allows for the easy viewing and interpretation of SNP-covariate associations, covariate-covariate correlations, SNP-SNP linkage disequilibrium, covariate associations with an outcome of interest, SNP associations with an outcome, covariate-covariate interactions predicting an outcome, SNP-covariate interactions predicting an outcome, and SNP-SNP interactions predicting an outcome. It also displays information about the cross-validation and replication of these associations, both important techniques for distinguishing true and false positive associations. The KGraph allows the user to more easily investigate multicollinearity and confounding through visualization of the correlation structure underlying genetic associations. It emphasizes gene-environment and gene-gene interaction, both important components of any genetic association frameworks that are often overlooked. The KGraph is implemented in Java and offers graphic features to help users to highlight and summarize the significant genetic effects with their background information. It is available at: http://www.epidkardia.sph.umich.edu/software/kgraph.

99

Bayesian Selection of Optimal Multivariate Polygenic Models for Linkage Analysis: Application to Correlated Behavioral Measures

J.W. Kent Jr.(1), J.C. Lopez-Alvarenga(1), T.D. Dyer(1), S.P. Fowler(2), H.P. Hazuda(2), R. Arya(2), M.P. Stern(2), J. Blangero(1), R. Duggirala(1)
(1) Dept. of Genetics, Southwest Foundation for Biomedical Research, USA, (2) Univ. of Texas Health Science Center-San Antonio, USA

Behavioral phenotypes derived from interview or observation are not expected to map one-to-one with the biochemical pathways that influence them, which may reduce the power to detect genetic effects when such measures are analyzed individually. However, these measures are likely to be inter-correlated, and derived combinatorial phenotypes may offer more power for detection of genetic effects. We are automating Bayesian model selection for combinations of univariate, bivariate, and higher n-variate models that best represent the genetic and environmental correlations within a dataset. Such models can be used as the basis for linkage or linkage/association analysis. Using a set of 8 psychosocial scales measured in the San Antonio Family Diabetes/Gallbladder Study (SAFADGS), we compare our model selection strategy to principal components factor analysis based on phenotypic correlations.

100

Symptom Dimensions in Psychotic Bipolar Disorder An Attempt to Address the Clinical Heterogeneity

B. Kerner
Center for Neurobehavioral Genetics, Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, UCLA, CA

Bipolar disorder (BPD) is a complex genetic disease with a heterogeneous phenotype. Psychotic symptoms have recently attracted attention, as they may point to an important subgroup of patients within the BPD spectrum with a more homogeneous genetic predisposition. They aggregate in families and occur both in mood disorder, as well as in schizophrenia (SZ). Therefore, this phenotype may lead to the identification of genes that carry common risk factors for both disorders. The goal of our study was to describe an exploratory model for the dimensionality of acute psychotic symptoms in BPD. This model may help to formulate hypotheses for an alternative definition of the phenotype in linkage and association studies. Methods: Factor analysis on acute psychotic symptoms in 412 bipolar patients data from the National Institute of Mental Health Bipolar Genetics Initiative (BPGI) was performed using the analysis program Mplus. The results suggest a three dimensional model for psychotic symptoms in bipolar disorder similar to that proposed in schizophrenia. A linkage scan performed in the entire data set of the BPGI using the presence or absence of symptoms describing the

first factor as phenotype resulted in a significant linkage signal on chromosome 5q32-33. Conclusion: These results suggest that symptom dimensions could serve as a quantitative or qualitative trait in linkage analysis. Symptom dimensions may be closer to the genetic risk factors than BPD as a whole.

101

Assessment of cancer risk in first-degree relatives of glioma cases

KF Kerstann, AM Goldstein, RT Falk, and DM Parry
Division of Cancer Epidemiology and Genetics, NCI, NIH, DHHS, Bethesda, MD, USA

The aim of this study was to evaluate the risk of cancer in first-degree relatives of 365 adults with glioma. Glioma, a malignant primary brain tumor, arises from glial cells and includes glioblastoma, astrocytoma, oligodendroglioma, ependymoma and mixed glioma. Telephone interviews were conducted with 1,509 first-degree relatives to obtain complete family and personal medical histories. We attempted to confirm reports of cancer in first-degree relatives with medical records or death certificates. From 562 reports of cancer 298 (53.2%) were confirmed, although not all cancers were as originally reported, 94 (16.7%) reports were confirmed as false and 170 (30.4%) reports were unconfirmed due to a lack of medical records. Among the 365 families, 310 had no relative other than the index case with a primary brain tumor (simplex) and 55 families had one or more additional relatives with a primary brain tumor (multiplex). We used standardized incidence ratios (SIR) to estimate the risk of cancer in the first-degree relatives. Overall, the relatives did not have substantially increased risks for cancer. However, some families showed clusters of specific cancers. The patterns of cancers within families were compared to those of known heritable syndromes associated with increased risk of glioma. Our results will be compared to previously published reports based on unconfirmed cancers in first-degree relatives of glioma families. Knowledge of the patterns of cancers associated with glioma-prone families could provide clues to the etiology of primary brain tumors.

102

Familial aggregation of metabolic phenotypes in a Korean community population

I.-K. Kim, T.-H. Kim, S.-I. Cho, D. Paek
Seoul National University School of Public Health

Metabolic syndrome (MetS) is a complex trait resulting from insulin resistance, but presents with a cluster of phenotypes, such as abdominal obesity, hyperlipidemia, hypertension, and diabetes. Each of these metabolic phenotypes is known to have familial aggregation, but those phenotypes may be heterogenous in the origins, one of the pathways being MetS as a disease entity. There are limited number of studies that compared the heritability of those phenotypes in a single population. In a familiy study among a community population in Korea, we assessed the familial correlations and heritabilities of the phenotypes related to metabolic syndrome. A total of 298 families were recruited in the study, including 218 parent-offspring pairs and 69 sibships. FCOR and ASSOC programs in SAGE software were used for the analyses. Age and sex were adjusted in all analysis. Among the phenotypes related to MetS, diastolic blood pressure had the greatest heritability (0.74), followed by systolic blood pressure (0.63), serum triglyceride levels (0.63), waist circumference (WC) (0.53), and blood glucose levels (0.39). Among the other indices of obesity, percent body fat (BF%) and body mass index (BMI) showed even higher heritability (0.79 and 0.67 respectively) than most of the MetS phenotypes. BF% also showed higher correlations than BMI or WC in most of the family pairs. We conclude that percent body fat may be the most relevant phenotype reflecting the underlying genetic pathways to MetS.

103

Concordance of Overweight in Korean Twins

T.-H. Kim(1), J. Sung(2), J.-S. Choi(3), Y.-M. Song(3), K. Lee(4), E.-Y. Choi(5), M. Ha(5), H. Kim(1), J.-H. Kim(3), K.-S. Hong(3), Y. Kim(1), E.-K. Shin(1), Y. Kim(1), S-I. Cho(1)
(1) Seoul National University, Korea, (2) Kangwon National University, Korea, (3) Sungkyunkwan University Medical School, Korea, (4) Inje University Medical School, Korea, (5) Dankook University Medical School, Korea

Obesity is a complex trait known to have a strong genetic component. We assessed the concordance of overweight and correlation of body mass index (BMI) among twins in The Korea Twin Family Cohort, which includes 26,000 twin pairs. Of those, 13,064 twin pairs had measured their BMI in the same year at least once between 1994 and 2004, at the medical exam provided by the Korean National Health Insurance Corporation. Overweight was defined by BMI>25. The study population included 1,535 male-female pairs, 7,593 male-male pairs and 3,936 female-female pairs. Of those, 8.4% were overweight concordant, and 70.4% were normal concordant, with pair-wise concordance of 79%, a significant increase from the expected concordance of 35% in a random population. BMI correlations among male-male, male-female, and female-female pairs were 0.49, 0.03, and 0.48 respectively. The correlations decreased significantly over time, suggesting an increase in the environmental effects due to changing lifestyles in Korea.

104

Investigating association of mitochondrial SNP with type 2 diabetes using machine learning methods

Y. Kim(1)(3), Y. Seo(2), J. E. Bailey-Wilson(3), H. Kim(1)
(1) Dept. of Biostatistics & Epidemiology, Seoul National Univ. School of Public Health, Seoul, Korea, (2) Univ. of Texas, San Antonio, TX, U.S.A (3) NIH/NHGRI/IDRB, Baltimore, MD, U.S.A

Association studies of diabetes and mitochondrial DNA variants have been widely performed since mitochondria have a central role in ATP production, which is related to insulin production and release. Most such studies were conducted using chi-square tests or logistic regression

models in case-control designs. We applied these methods, and then also used machine learning methods for classification, clustering, and regression in order to validate the results from the general chi-square test and find the mtSNPs that are associated with case-control status. 132 mitochondrial SNP sites from 130 type 2 diabetes and 65 non-diabetes controls in Korea were used in the analyses. Six mtSNPs; 3173, C4883T, A4985G, C8271, A8281, and G10310 are selected as associated with susceptibility toward diabetes using the chi-square test. We found A4985G, 3173, G10310, and C4883T as important mtSNPs according to the Gini index from Random Forests. The results using the dissimilarity matrix calculated from the random forest method will be presented. In CART (Classification and Regression Trees), G5231A and C3290 were found as determinants to classify cases and controls. Synthesizing and summarizing the results from the machine learning methods could provide empirical evidence of association. The unique aspects faced in the analysis of mitochondrial DNA variants will be discussed.

105

Comparison of Model free Linkage Methods for Mapping Quantitative Trait Loci by Monte-Carlo Simulations

A. Kleensang, I. R. König, A
ZieglerInstitute of Medical Biometry and Statistics University Hospital Schleswig-Holstein - Campus Lübeck

Over the last two decades complex traits with quantitative markedness gained considerable interest in genetic epidemiology which was accompanied by an explosive development of methods to map quantitative traits. Unfortunately, the robustness and power of these methods depends on the distribution of the phenotypic data, and the analysis of selected data and/or non-normally distributed traits is not a rare situation. Some recommendations are available in the literature to decide which method may be appropriate in a given situation; however, a comprehensive comparison under different conditions in sense of robustness and power is still lacking. To fill this gap, we present the results from a Monte-Carlo simulation study comparing seven commonly applied methods (Haseman-Elston, maximum-likelihood-binomial, maximum-likelihood variance component models, Merlin-QTL, Merlin-Regress, new-Haseman-Elston, Wilcoxon signed rank test) under 36 scenarios. In more detail, we used the Falconer and Mackay additive model with a major gene effect, an environmental effect simulated as a family effect and an error term. We considered three genetic models (dominant, additive, recessive), three selection schemes (random sib-pair-, single proband sib-pair-, extreme sib-pair-design), two family structures (nuclear families with two offspring and with two to five offspring) and two distributions for the error term (normal, log-normal).

106

Correlation analysis of ocular biometric measurements in the Beaver Dam Eye Study

AP Klein(1), P Duggal(2), K. Lee(3), JE Bailey-Wilson(2), B Klein(3), BEK Klein(3)
(1) Dept. of Oncology and Epidemiology, Johns Hopkins Medical Institutions Baltimore MD, (2) Statistical Genetics Section, NIH/NHGRI/IDRB, Baltimore, MD, (3) Dept. of Ophthalmology and Visual Sciences, Univesity of Wisconsin Medical School, Madison WI

Ocular refraction measures the power of an external lens needed to focus images on the retina and is influenced by the underlying ocular biometry. The biometry is described, in part, by corneal curvature, axial length and anterior chamber depth. While many studies have demonstrated a high heritability for quantitative refraction, no large-scale study has examined the heritability of these measures of ocular biometry. To examine the correlation for these measures we conducted familial correlation analysis using FCOR, S.A.G.E v5.2. Heritability estimates were obtained using SOLAR. Analyses were based on 552 individuals in 155 extended pedigrees who participated in the Beaver Dam Eye Study. Within an individual, measures of axial length, corneal curvature and anterior chamber depth were each correlated with refraction and each measure was highly correlated among sibling pairs (pair-wise correlations ranged from 0.31-0.47). Measures of axial length, corneal curvature and anterior chamber depth demonstrated stronger familial correlation than measures of quantitative refraction. Heritability estimates were 58% for refraction, 77% for anterior chamber depth, 62% for axial length and 87% for corneal curvature. Our analyses have implications for genetic studies of quantitative refraction. Examination of the multivariate traits is ongoing and will be presented.

107

Index Selection as a Multivariate Technique for Quantitative Personalized Medicine

AT Kraja(1), M Crosswhite(1), S Marsh(1), RS Lin(1), HL McLeod(2), MA Province(1)
(1) Division of Statistical Genomics, Washington University School of Medicine, (2) UNC Chapel-Hill

The dream of personalized medicine is to use the genetic profile to individualize therapy. Index Selection (IS) is an established multivariate statistical technique developed in plant and animal breeding research designed to find signal sets of genes which operate together. We applied IS and compared it to more traditional methods to the problem of discovering the genetic basis of cytotoxicity response in cancer therapy. We treated 30 trios of immortalized CEPH cell lines from HapMap subjects with multiple doses of two chemotherapy drugs Docetaxel and 5-Fluorouracil to establish individual dose-response curves. 3,523,637 SNPs were available from the HapMap. A group of filters (MAF>10%; tagSNPs; and permutation tests of association) were performed to select the most promising SNPs. Only the additive effects of these SNPs are included in the IS model IS=b(z)'z + b(m)'m, where z represents a vector of quantitative traits, m is a vector of the selected SNPs, and b is the corresponding additive effects. In a preliminary analysis, 68 significant SNPs located within genes, out of 228,084 SNPs on chromosome 5 were included in an IS, which previously demonstrated strong

linkage to cytotoxicity response. The correlation between the IS and the observed viability of CEPH cell lines was 0.814. This method holds promise to quantify the predicted individual response to chemotherapy.

108

The Relationships Between Polymorphisms in Hypoxia Inducible Factor 1Ü and Hypertension in Families with Sleep Apnea

EK Larkin(1), S Redline(1), NR Prabhakar(1), SR Patel(1), GL Semenza(2), RC Elston(1), C Gray-McGuire(1)
(1) Case Western Reserve University, (2) Johns Hopkins University

HIF1 plays an important role in oxygen metabolism by controlling cellular responses to oxygen deficits. HIF1 mediates the blood pressure response to intermittent hypoxemia, an integral feature of sleep apnea (SA). It is unknown whether the susceptibility to hypertension (HTN) is related to polymorphisms in HIF1α and whether SA would alter the relationship. Methods: 7 single nucleotide polymorphisms (SNPs) evenly spaced across HIF1ά were genotyped in 249 adults from 78 African American (AA) families and 188 adults from 50 European American (EA) families. Minor allele frequencies of the SNPs differed by race. Each SNP was treated as independent. Using an association test that accounts for familial relationships, we assessed the relationship between the presence of HTN and each SNP, adjusting for age, sex and SA. An interaction term between SA and each SNP was also included. Results: In AAs, one SNP (RS10144011) was associated with HTN (p=.02). Another SNP (RS1957754) by SA interaction provided some evidence that the relationship between HTN and a SNP was mediated by the presence of SA (p=0.08). Using a more stringent definition of SA produced slightly stronger results (p=.05). In EAs no SNP was associated with HTN and no interaction was significant. Conclusion: A suggestive association between HTN and HIF1α and a weak interaction between a single intronic SNP and SA were detected in AAs. Further follow-up may elucidate the role of HIF1α as a mediator of HTN in individuals with SA.

109

Subgroup Weighted Association Testing

Michael LeBlan
Fred Hutchinson Cancer Research Center, USA

We investigate a class of test statistics for large-scale association studies based on weighting environmental subgroup information. Given many analyses involve testing univariate or marginal SNP associations with subject outcome, there is a concern that more complex relationships such as multiple gene combinations or gene-environment (or gene-treatment) interactions could reduce the power to detect the marginal associations. However, directly testing interactions is often more difficult due to limited power and leads to dramatic increases in the number to tests conducted. In situations where the association is of a greater magnitude within a subgroup of subjects with specific environmental/clinical attributes (i.e. among smokers), we propose a computationally simple panel of weighted marginal tests that can exploit such subgroup associations if they exist. The strategy modestly increases the search by focusing (or enriching) association tests within subgroups. A Monte Carlo method based on simulating from the approximate large sample distribution the panel of statistics is used to control Type 1 error. Results from simulation studies confirm improved power of the proposed approach.

110

How to Incorporate Population Covariate Effects into Linkage Analysis for Binary Traits

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

In linkage analysis of continuous traits, where random samples of families are often available, the marginal effect of important covariates such as sex and age is usually incorporated into the mean structure of the variance components model. For binary traits, families are often ascertained (e.g. in affected sib pair designs) and estimates of the effect of covariates at the population level cannot be obtained. In such designs, incorporation of covariates is usually based on models where the identity-by-descent (IBD) probabilities are allowed to depend on those covariates. External information on covariates obtained via population-based studies is therefore ignored altogether. We propose to use a Generalized Linear Mixed Model (GLMM) as a natural extension of the variance components model in order to incorporate marginal covariate effects. Based on a score test of a quasi-likelihood of this GLMM, we obtain a tractable expression for the linkage test. The effect of covariate adjustment then becomes apparent and we are able to identify situations where the efficiency can be substantially increased compared to approaches where covariates are either ignored or conventionally adjusted for using the linkage data only. Our test applies to general pedigrees with both affected and unaffected individuals and accommodates person-specific covariates. The GLMM framework allows a simple generalization of the approach to other types of traits such as count and censored data.

111

Familial aggregation of lung function in Korean families

H.-J. Lee(1), S.-I. Cho(1), S. Park(1), H.-J. Jhun(2), D. Paek(1)
(1) Seoul National University School of Public Health, Korea (2) Hallym University Medical School, Korea

Lung function has a definite genetic component, but also can be affected by environmental pollutants often shared within the family. The complex pathways of gene-environment interaction associated with lung function is not well known, and may vary among different populations. We explored the familial correlations and heritability of lung function in Korean families. A total of 245 families

were included in the study, with 352 parent-offspring pairs and 66 sibling pairs. FCOR and ASSOC programs of SAGE software were used for data analyses. Age, sex, height, and asthma status were adjusted for in the analyses. Heritability was greatest for forced vital capacity (FVC), 0.55; followed by forced expiratory volume in 1 second (FEV1), 0.39; forced expiratory flow (FEF25-75), 0.35; and FEV1/FVC ratio, 0.25. For FVC, father's correlations were 0.33 with sons and 0.21 with daughters, suggesting greater effects compared to mothers, whose correlations were 0.07 with sons and 0.10 with daughters. Familial correlations for FEV1/FVC ratio were the lowest of the 4 phenotypes in most of the relationships. We conclude that FVC has the strongest genetic component, particularly from fathers, and FEV1/FVC appears to be most sensitive to environmental effects, among the phenotypes of lung function.

112

Heritability of refractive error in an urban population

H.-J. Lee, S. Park, S.-I. Cho, D. Paek
Seoul National University School of Public Health

The etiology of refractive error is complex, but the major form of refractive error results from an interplay between genetic factors and environmental influences. Although there is evidence that parental refractive error influences the position of their offspring within that new population distribution, few studies have estimated the heritability of refractive error phenotype. In the present study, we examined the genetic regulation of refractive error in individuals from nuclear and extended families. Our study population includes 257 men and 316 women aged 4–82 years (mean+/-SD: 32.8+/-18.4) who belong to 298 pedigrees. Three measures of refractive error were collected from both the right and left eyes using an autorefractor (R-10; Canon, Japan). We used a variance components based maximum likelihood method to estimate the heritability of refractive error while simultaneously adjusting for age and sex effects. The mean refractive error (spherical equivalent) were 1.47+/-2.01D for the right eye and -1.42+/-2.09D for the left eye. Heritability of refractive error (h^2=+/-SE) were 0.45+/-0.11 for the right eye and 0.51+/-0.11 for the left eye. Variance components among siblings, which mostly represent shared environment, were 0.42 for the right eye, and 0.36 for the left eye. Among the familial correlations, coefficients between sisters were highest (0.4), followed by that between mother-daughter (0.2). Our results suggest that genetics play a significant role in determining refractive error, particularly among females. The genetic influence was greater in the left eye than in the right eye.

113

Incident Age-Related Macular Degeneration Risk from Smoking and the CFH gene

KE Lee(1), MD Knudtson(1), SK Iyengar(2), R Klein(1), EL Moore(1), BEK Klein(1)
(1) Department of Ophthalmology and Visual Science, University of Wisconsin Medical School, Madison. (2) Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Risk of developing early signs of age-related macular degeneration (AMD) from smoking and the Y402H polymorphism in the Complement Factor H (CFH) gene was analyzed from data obtained through the population-based Beaver Dam Eye Study. Over half of this population had a family member also in the study. The CFH gene polymorphism Y402H (1277T–>C) was genotyped in 1971 participants of this study, who were 43 to 86 years of age. The T to C change in this SNP causes a change from tyrosine (Y) to histidine (H). Smoking history and photographs, from which early AMD was identified, were obtained during a baseline examination from 1988–90. Three follow-up examinations, with photographs, have occurred at 5 year intervals. Persons free of early AMD in both eyes at baseline and with follow-up evaluation of early AMD were eligible for analyses (N=1125). The 15-year cumulative incidence of early AMD was 13.4% in the 447 persons YY, 16.3% in the 523 persons HY and 24.6% in the 137 persons HH for the Y402H polymorphism. Among never smokers, the rates were 15.5%, 16.9% and 15.1%, for YY, HY, and HH polymorphisms, respectively. Among past smokers the rates were 11.3%, 19.3% and 24.9%, respectively, while among current smokers the rates were 12.9%, 9.2% and 49.9%. The test for interaction was significant (p=0.02). These data suggest that the risk of early AMD from smoking is modified by the CFH polymorphism.

114

A Random Forest Approach to Identify Important Interacting Markers in Quantitative Trait Linkage Analysis

S.F. Lee(1, 2), L. Sun(1, 3), S.B. Bull(1, 2)
(1) Department of Public Health Sciences, University of Toronto, Toronto, Canada, (2) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada, (3) Hospital for Sick Children, Toronto, Canada

Regression random forest is an ensemble of tree predictors with high predictive accuracy and a feature that measures the importance of covariates from a complex data structure. We developed a random forest for multi-locus quantitative trait linkage analysis that accounts for ambiguity in marker data by incorporating the posterior identical-by-descent (IBD) probabilities from EM Haseman Elston (HE) regression as weights on each sibpair in the tree predictors. We also propose a new hybrid variable importance measure, which considers the correlation structure inherent in IBD linkage data, and compared its ability to detect quantitative trait loci with the HE LOD score by simulation under several genetic and epistatic models. The new random forest and hybrid variable importance measure show promising results in identifying important markers influencing a quantitative trait while addressing ambiguity in marker data and exploring complex interaction among all markers simultaneously.

115

Tag SNP Selection Using Genetic Structure and Prior Information

W. Lee, D. Li, D.V. Conti
Dept. of Preventive Medicine, USC, Los Angeles, CA. USA

Current SNP genotyping technology has made the characterization of many genes in large populations feasible. However, the genotyping and analysis of all possible SNPs within each gene can be costly and inefficient. To reduce the number of SNPs, we have developed a SNP selection program that combines linkage disequilibium (LD) with additional information for each SNP. Designed to integrate with the Illumina platform, we use genotype information from HapMap or user-defined populations and group SNPs with pairwise LD above a pre-specified threshold into bins. Unique to this process is the ability to determine bins conditional on LD structure and SNP frequency within across ethnic groups. Final selection of tagSNPs to characterize bin structure is predicated on the minor allele frequency, the potential for genotyping success, location within a gene (intron, exon), and known functionality. Additional SNPs may be forced in to insure even spacing of SNPs throughout the gene region and/or based on prior functional study. To address Illumina requirements, we implement an extensive error checking procedure. SNPs that cannot be genotyped successfully are excluded via the Illumina genotyping score and SNP combinations that may lead to unstable allele calls via close physical proximity are prioritized and replaced using initial selection conditions. Our goal is to capture genetic information across a gene region with a minimal set of tagSNPs without losing SNPs with substantial prior information. Thus, we hope to characterize the regional association and increase our potential to identify specific disease-causing variants.

116

The Relationship of the G-250A Polymorphism of the Human Hepatic Lipase Gene Promoter with Metabolic Syndrome

Y.H. Lee(1), M. C. Kim(1), J.W. Kim(2), J.Y. Kwak(2), Y.S. Hong(2)
Dept. of Preventive Medicine College of Medicine, Kosin University

The -250G to A polymorphism in the promoter of the hepatic lipase (LIPC) gene has been associated with lowered hepatic lipase activity, dyslipidemia, and diabetes. The aims of this study were to elucidate the relationship of the G-250A polymorphism of LIPC with metabolic syndrome. A total of 943 healthy examinees who were examined in Kosin University Gospel Hospital from December 2004 to February 2006 were enrolled in this study. Height, weight, body mass index, body fat mass, visceral fat mass, waist circumference, and systolic and diastolic blood pressure of the subjects were examined. Fasting blood glucose, total cholesterol, HDL cholesterol, LDL cholesterol, and triglyceride were also measured in venous serum. The genotype at position -250 of the hepatic lipase promoter was determined by single base extension and electrophoresis. The observed frequencies of the -250G to A polymorphism in the promoter of LIPC were 49.8% for the metabolic syndrome, 51.4% for the control group and 50.9% in total subjects. The frequency of the A allele was 36% in total subjects. Concentration of triglyceride in venous serum was significantly higher in subjects with GA and AA genotypes than in GG type in women.. Frequency of the -250A allele in Korean was 36%, and the -250G to A polymorphism in the promoter of LIPC might influence on triglyceride level in venous serum in women.

117

SUP: an extension to SLINK to allow a larger number of autosomal or sex-linked markers to be simulated in linkage equilibrium or disequilibrium with a trait locus and conditional on trait values

M. Lemire(1,2) N.M. Roslin(2)
(1) McGill University and Genome Quebec Innovation Centre (2) Research Institute of the McGill University Health Centre

SLINK is a flexible simulation tool that has been widely used to generate the segregation and recombination processes of markers linked to, and possibly associated with, a trait locus, conditional on trait values. In practice, its most serious limitation is the small number of loci that can be simulated. I describe the implementation of a two-step algorithm to be used in conjunction with SLINK to enable the simulation of a large number of marker loci linked to a trait locus and conditional on trait values in families, with the possibility for the loci to be in linkage disequilibrium. SLINK is used in the first step to simulate genotypes at the trait locus conditional on the observed trait values, and also to generate an indicator of the descent path of the simulated alleles. In the second step, marker alleles or haplotypes are generated in the founders, conditional on the trait locus genotypes simulated in the first step. Then the recombination process between the marker loci takes place conditionally on the descent path and on the trait locus genotypes. The second step of the algorithm has been implemented in the program SUP (SLINK utility program), that now supports the X chromosome.

118

Detecting Associations in the Presence of Extreme Allelic Heterogeneity

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

Linkage disequilibrium (LD) is being utilized for common disease gene mapping. However a critical assumption of this indirect approach is that the trait is due to a common variant and there is no allelic heterogeneity, which will greatly reduce the power of LD mapping. Currently large scale candidate gene sequencing is underway to discover multiple rare alleles associated with diseases. In this study extensive case-control simulations were performed to

evaluate the performance of various statistical methods for detection of main effects in the presence of allelic heterogeneity. The focus is on the power and robustness in the presence of missing causative variants and misclassification (inclusion of non-causative polymorphisms). The methods evaluated include single marker Fisher exact test, multi-marker tests (Fisher product, Hotelling's T2 and logistic regression), and collapsing genotype/haplotype method based on the presence of rare mutations on chromosomes. The results show that collapsing method is most powerful in the absence of misclassification but its power drops quickly with non-causative markers. All multi-marker tests are more powerful than the single marker test and multi-marker tests are more robust against misclassification but more sensitive to missing causative variants. Applying the tests on non-causative polymorphisms demonstrated that the false positive rate is well controlled. Therefore in the presence of allelic heterogeneity, multi-marker tests such as logistic regression are recommended and less stringent criteria should be used when pre-filtering non-causative markers.

119

A logic-regression based approach for the marker-locus association studies

D Li, D.V. Conti
Dept. of Preventive Medicine, University of Southern California, Los Angeles, CA, USA.

Association studies have become the workhorse of genetic epidemiology. Recent developments in association studies indicate that single-locus tests can be helpful for the construction of powerful inference methods for multi-locus analysis. Sasieni has pointed out that a commonly used allele frequency based test is appropriate only when Hardy-Weinberg equilibrium holds and suggested to use the Cochran-Armitage trend test instead. This method requires assigning a set of scores to the genotypes, based on priori heredity models which is usually unknown in reality. Several methods have been proposed to address this problem, such as the ''Max'' method by Freidlin et al. and the ''max and min scores'' approach by Zheng. Model-free approaches, such as the Constrained-Likelihood Approach and U-test, have also been proposed. Here we propose a logic-regression approach to determine the underlying heredity model based on the residual sum squares of each model and then a subsequent likelihood ratio test of association. The power of our method and these methods are compared across a variety of scenarios. We demonstrate that the logic-regression approach is generally more powerful when the true heredity model and the marker allele associated with the causal alleles are unknown, which is true for most association studies. Another advantage of the logic-regression approach is that it can directly detect the heredity model and estimate the parameter, which is not applicable in the other four approaches. We further discuss the implications and expansion of our approach in the construction of the multi-locus models.

120

Association between Phosphatidylinositol 3-Kinase p85á Regulatory Subunit Met326Ile Genetic Polymorphism and Colon Cancer Risk

L. Li(1,2), S. Plummer(3), C.L. Thompson(1,2), T.C. Tucker(4), G. Casey(3)
(1,2) Departments of Epidemiology and Biostatistics and Family Medicine, Case Western Reserve University; (3) Department of Cancer Biology, Cleveland Clinic Foundation; (4) Markey Cancer Center, University of Kentucky

The phosphoinositol 3-kinase (PI3K) signaling pathway critically regulates cell growth and cell survival. Aberrant activation of PI3K and somatic mutations in the gene (PIK3R1) coding for the p85á regulatory subunit have been reported in primary human tumors, including colon cancer. We sought to assess the association of a putatively functional SNP (Met326Ile) in PIK3R1 with colon cancer in an ongoing population-based case-control study. We included 366 incident cases and 442 controls in the analysis. Odds ratios (ORs) were estimated in logistic regressions controlling for age, gender, race, BMI, family history, and non-steroidal anti-inflammatory drug use (NSAIDs). Compared to those homozygous for Met/Met, those with one or two copies of the variant (Ile/Met or Ile/Ile) had a significantly increased risk (OR=1.67, 95% CI=1.20-2.32, p=0.002). Stratified analyses revealed that the risk is more pronounced among older patients: the estimated ORs were 2.01 (1.25-3.22, p=0.004) for cases ≥ 64 years at diagnosis, and 1.38 (0.95-2.01, p=0.09) for cases < 64 years, respectively. To our knowledge, this is the first report of a direct association between the p85α gene polymorphism and colon cancer risk, providing evidence supporting PIK3R1 as a susceptibility gene for colon tumorigenesis.

121

Bioinformatics Approach for Candidate Gene Selection in Genetic Association Study

R Li(1), J Snoddy(2), G Liu(1) and B Zhang(2)
(1) Center for Genomics and Bioinformatics, University of Tennessee Health Science Center, (2) Dept. of Biomedical Informatics, Vanderbilt University

With the advantage of high-throughput technology, an overwhelming amount of information about genes and gene products has been accumulated daily from basic science. We identify a set of candidate genes underlying pathogenesis pathways of cardiovascular disease for genetic association study using bioinformatics tools. First, we pooled all currently available genes identified through the Programs for Genomic Applications (PGAs), National Heart, Lung and Blood Institute (NHLBI), NIH, forthe selection of cardiovascular disease candidate genes. Second, we selected the genes related to coronary heart disease or cerebrovascular disease, using the locally developed bioinformatics tool, MeSH Disease (MD) Tree Machine. Disease categories in the MD Tree Machine are based on the medical subject heading (MeSH) and organized in a hierarchical tree structure. The gene-publication relationship is defined in Entrez Gene, while

the publication-MeSH relationship is defined in PubMed. Third, we further filtered the gene lists by their over-expression in the tissues of interest using the tissue expression profiling tool in WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt). Finally, in order to understand the underlying biological organization of the genes, we mapped the genes to the pathway maps using the pathway analyzing tool in WebGestalt. There are 545 genes currently included in the NHLBI PGAs databases. Among them, 162 genes are linked to coronary heart disease and/or cerebrovascular disease categories by MD Tree Machine. The expression profile of these genes in 10 human tissues, which are thought to be related to cardiovascular disease development, showed that 84 genes are over-represented ($p < 0.001$) in one or multiple of the 10 tissues. The results from functional analysis indicate that, of the 84 genes, 57 genes are significantly ($p < 0.001$) over-represented in pathogenesis pathways towards the cardiovascular disease development, such as MAPK signaling and Toll-like receptor signaling pathways. There are a maximum of 13 genes in one pathway (e.g., Focal adhesion pathway) and a gene (e.g., MAPK14) in a maximum of 13 pathways. The bioinformatics approach may overcome the limitatations of single gene and of ''fishing expedition'' by whole genome scans for common disease association studies.

122

Selection of cases in case-control studies using identity by descent sharing from families with multiple affected siblings

C Liu(1), L. A. Cupples(2), Q. Yang(2), J. Dupuis(2)
(1) Dept. of Neurology, Boston Univ. School of Medicine, USA, (2) Dept. of Biostatistics, Boston Univ. School of Public Health, USA

A case-control study often follows an initial linkage analysis. To reduce the cost and increase the power of detecting disease-marker associations, different strategies have been exploited previously to identify genetically loaded individuals. Fingerlin, Boehnke and Abecasis (Hum Genet 2004, 74:432–443) reported that improved power results from selecting cases based on identity by descent (IBD)-sharing estimated at a disease locus.In this study, we propose to pre-select one case from families with multiple affected siblings using linkage LOD scores, multipoint IBD sharing from microsatellite markers, and using information on disease severity. This proposed extension allows one 1) to select cases solely based on the genetic markers used in the linkage analysis, before typing any of the single nucleotide polymorphisms (SNPs) and 2) to select the same samples across a region/gene to be tested for association. This could represent significant genotyping cost savings over the original Fingerlin et al. proposal, where the case selection required that the SNP genotypes of all affected siblings be available. By simulations we demonstrate that the case selection strategy based on microsatellite markers has equivalent power to the selection procedure requiring genotypes of all affected individuals. When families have only 2 affected sibs, both sibs have identical IBD sharing; thus, selection of cases based on a severity index is the most powerful method.

123

Haplotype-Based Regression Analysis of Case-Control Studies with Unphased Genotypes and Measurement Errors in Environmental Exposures

Iryna Lobach(1), Raymond Carroll(1), Christie Spinka(2), Mitch Gail(3) and Nilanjan Chatterjee(3)
Texas A&M University

It is widely believed that risks of many complex diseases are determined by genetic susceptibilities, environmental exposures, and their interaction. Chatterjee and Carroll (2005) developed an efficient retrospective maximum-likelihood method for analysis of case-control studies that exploits an assumption of gene-environment independence and leaves the distribution of the environmental covariates to be completely nonparametric. We extend this approach to situations of parameter estimates when the variance of the measurement error is known and when it is estimated using replications. The performance of the proposed method is illustrated using simulation studies emphasizing the case when genetic information is in the form of haplotypes and missing data arises from haplotype-phase ambiguity. An application of our method is illustrated using a population-based case-control study of the association between calcium intake with the risk of colorectal adenoma.

124

Identifying putatively linked pedigrees for molecular follow-up based on principal component analysis of the multidimensional linkage likelihood surface

M.W. Logue
Center for Statistical Genetics Research, Univ. of Iowa, USA

A strategy is proposed for identifying subsets of pedigrees linked to each of several loci at which linkage (for the entire set) has already been established. This is done by treating the linkage likelihood surface for each pedigree, computed over a grid of genetic model values for each linked marker, as a multivariate observation. Principal component analysis is used to define a new set of predicted factors which, in essence, summarize the by-pedigree variation of the likelihood surfaces in a small number of dimensions. These factors can then be used to display this variability, and give an indication of which pedigrees are linked to a particular locus. This represents a more elegant solution than simply looking for pedigrees which yield large LOD scores at one locus or the other, and incorporates information from multiple genetic models, without the need to specify the genetic model a-priori. This method is demonstrated in a series of test cases, simulated under admixture, to explore ways that genetic model information can be used to determine linkage group membership. Possible extensions of this method include incorporation of additional likelihood parameters (e.g. linkage disequilibrium), and inclusion of pedigree specific covariate information (e.g. average age of onset).

125

Extensions to the Weighted-Average Statistic for Fine-Mapping in Case-Control Studies

D. Londono, R.C. Elston, and K.A.B. Goddard
Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, OH USA

Song and Elston (Statistics in Medicine 25:105–126, 2006) presented a method for fine-mapping of a disease-susceptibility locus in the context of case-control designs termed the Weighted-Average (WA) statistic. This method combines two measures of association: the Cochran-Armitage trend test and a novel trend test based on the difference between Hardy-Weinberg disequilibrium (HWD) for cases and controls. This latter test is called the HWD trend test. The expected advantage of the WA statistic lies in the fact that combining two independent tests of association highlights the strong features of each individual test while lessening their respective weaknesses. We propose two variations of the WA statistic. The first one is an extension of the WA statistic to quantitative trait loci. The second variation is a modification of the HWD trend test part of the WA statistic. We compute the test using the population allele frequency rather than frequencies estimated separately for cases and controls. This modification reduces the test to a comparison of two genetic frequencies, which eliminates the need for estimating population allele frequencies altogether. For a recessive model, the improvement in power compared to the original formulation is modest, whereas for a dominant, additive and multiplicative model the improvement in power can be very large depending on the underlying model. These two variations of the WA statistic are demonstrated with a simulation study using a case-control design.

126

Relationship between Genotype Variants in Coronary Artery Disease (CAD) and HDL level

X. Lou, L. Wang, W. E. Kraus, E. R. Hauser, S. Shah
Department of Medicine, Duke University Medical Center, USA

Multiple lines of evidence lead to the presence of a CAD risk gene on chromosome 3q13 which may be related to risk through HDL levels. We have shown linkage evidence to early onset CAD (EOCAD) on chromsome 3q13 (LOD=3.5). HDL cholesterol loci were mapped on chromosome 3q. Bowden et al. revealed evidence for coincident mapping of type 2 diabetes, metabolic syndrome, and measures of cardiovascular diseases to 70–160 cM on chromosome 3. OSA analysis on EOCAD families showed that linkage evidence concentrated in subsets with favorable lipid profiles. We examined evidence for association between CAD and gene variants in this region in an independent dataset with two case groups (old (OA) and young (YA) CAD subjects) and a control group (ON). In order to study the possible relationship between HDL levels, CAD and genetic risk in this region, 119 SNPs which are significantly associated were genotyped in 167 OA, 556 YA, 256 ON. Measured genotype analysis was performed using mean HDLC (mg/dL) by genotype separately for each group, using linear models adjusted for sex and race. We found that HDLC was lowest in YA (39.8), then OA (44.0) and ON (51.7). 10 SNPs showed significant (p=0.05) differences in HDLC by genotype in YA, including: RS17310144 (mean HDLC(SD) by genotype (11,12,22): 36.8(9.86), 40.2(12.7), 42.1(11.3), p=0.035); RS2335052 (38.6(11.04), 42.1(13.3), 47.2(9.12), p=0.02); RS2811529(38.7(10.5), 38.6(12.4), 44.85(12.5), p=0.009). These results suggest that genetic variants in chromosome 3q13 may be associated with HDLC in early onset CAD.

127

Prospective and Retrospective Analyses of Type-2 Diabetes and Controls in Ashkenazi and UK Populations

J.A. Luan(1), J.H. Zhao(1), Q.H. Tan(2)
(1) MRC Epidemiology Unit, Cambridge, UK, (2) Odense University Hospital and University of Southern Denmark, Denmark

Case-control design has been one of the most established study designs in biostatistics and epidemiology, and is now widely used in genetic association studies. Among the many statistical issues involved, the prospective versus retrospective approach of analysis is particularly interesting. In contrast to its prospective counterpart where phenotype is customarily treated as outcome, the retrospective approach treats phenotype as predictor, which seems to be counter-intuitive to many genetic epidemiologists. Although it was examined a long time ago, together they remain a topic of active discussions. Another issue in case-control genetic association studies is related to the staged design and currently it enjoys enormous popularity. Since joint analysis of data from all stages is shown to be more powerful and preferred, it is as yet to be fully integrated into routine data analysis. Here we use both prospective and retrospective approaches to analyse data from a large study of type-2 diabetes involving 5,034 Ashkenazi and four UK populations using two-staged design and 4,540 SNPs across a 10Mb region of chromosome 20q. A recent proposed framework is used to account for heterogeneity between the UK populations. Various aspects in relation to these approaches are discussed. With a large number of whole-genome association studies of common diseases being currently carried out, our work has important implications in analysis of data generated from such studies.

128

Large scale SNP epistasis detection of complex traits using pairwise epistasis tests

L. Ma(1), D. Dvorkin(2), Y. Da(1)
(1) Dept of Anim Sci, Univ. of Minnesota, USA, (2) Dept of Biostat, Univ. of Minnesota, USA

The large number of SNP markers available provides opportunities for large scale genome-wide SNP candidate gene studies to identify DNA variations or genes affecting complex traits through direct or indirect SNP effects. Epistasis or gene interaction effects have been increasingly recognized as important genetic factors underlying

complex traits. A flexible statistical method based on an extended Kempthorne model to allow Hardy-Weinberg and linkage disequilibria was developed for pairwise epistasis testing of potentially large numbers of SNPs. A computer package named epiSNP was developed to implement the pairwise epistasis testing method using a two-step regression analysis. The main program in the package, EPISNP, offers pairwise tests of the two-locus interaction and four epistasis effects, additive x additive, additive x dominance, dominance x additive, and dominance x dominance, and offers tests of three single locus effects of each SNP. The EPISNPPLOT program produces graphical views of single-locus significant results and sample sizes for each chromosome. The CPUHD program estimates the CPU time and disk space required to execute the EPISNP program. Test runs of the epiSNP package on PCs showed that computer speed is the main limiting factor and parallel computing is needed for large scale SNP epistasis analysis.

129

Development of Ascertainment Correction for Monte Carlo Markov Chain Segregation and Linkage Analysis

J.Z. Ma, E.W. Daw, and C.I. Amos
Department of Epidemiology, U.T. M.D. Anderson Cancer Center

Even though families are often selected through probands with extreme levels of a quantitative trait, Monte Carlo Markov Chain methods have not been able to perform ascertainment corrections. The goals of this research have been 1) to evaluate the bias associated with failing to correct for ascertainment under varying degrees of stringent in the selection process, 2) development and implementation of an ascertainment correction procedure into the LOKI software, 3) evaluation of the bias of parameter estimates following ascertainment correction. With respect to aim 1, we found only limited bias in estimates of location in linkage analyses when ascertainment correction was not applied to the selected samples. However, the genotypic means and allele frequencies showed increasing bias as the selection stringency increased. The ascertainment correction that we developed conditions on a selection threshold and required modifying the likelihood and the acceptance ratio used by the MCMC sampler of LOKI. This ascertainment correction can accomodate one or more loci, but the current implementation requires the user to specify the number of loci that are included in the model. We found that ascertainment correction greatly improves the estimates of allele frequency and genotypic means particularly for high valued genotypes, when selection is through individuals with high values. Misspecifying the number of loci led to accurate estimation of the trait locus positions, but when two loci were simulated but only one modeled, the genotypic means were inaccurately estimated.

130

Homocysteine-related polymorphisms, serum homocysteine levels and frailty in older women: the Women's Health and Aging Study

A.T. Mahoney(1), J.D. Walston(2), W.H. Kao(1), K. Bandeen-Roche(3), L.P. Fried(2), S.P. Stabler(4), M.D. Fallin(1)
(1) Department of Epidemiology, (2) Medicine, and (3) Biostatistics, Johns Hopkins University, USA (4) Department of Medicine, University of Colorado, USA

Frailty is described as a syndrome of physiologic decline across multiple biologic systems leading to poor outcomes in older adults. Elevation in total homocysteine (tHcy) has been implicated in age-related health issues, such as cardiovascular disease and osteoporosis. Prior analyses have demonstrated a statistically significant association between elevated tHcy and baseline frailty in community-dwelling older women. It is hypothesized that genetic polymorphisms within homocysteine-related genes disrupt the homocysteine metabolic cycle, contributing to chronic tHcy elevation and to the prevalence of frailty syndrome. Using Illumina bead-array technology, 49 SNPs in four candidate genes (cystathione-beta synthase, methylenetetrahydrofolate reductase, methionine synthase reductase, methionine synthase) were genotyped in 770 participants of the Women's Health and Aging Study I and II cohorts. Continuous and binary tHcy outcomes and binary frailty outcome were modeled in single SNP and haplotype regression analyses. Haplotypes were built for each candidate gene using Haploview software and used for haplotype analysis in the Haplo.stats package. Vitamin deficiencies, smoking status and alcohol consumption were tested for confounding and effect modification. Identification of genetic risk factors for frailty could establish a specific risk population for geriatricians to target.

131

A non-parametric approach to detect gene-gene and gene-time interaction for longitudinal data in cohort studies

D. Malzahn(1), A. Neumann(1), M. Müller(2), H.-E. Wichmann(2), H. Bickeböller(1)
(1) Department of Genetic Epidemiology, University of Goettingen, Germany, (2) Institute of Epidemiology, GSF, Germany

Longitudinal data show the time dependent course of phenotypic traits. In this contribution, we consider longitudinal cohort studies and investigate the association between 2 loci and a dependent quantitative longitudinal phenotype. The set-up defines a factorial design which allows us to test simultaneously for the overall gene effect of the two loci as well as for possible gene-gene and gene-time interaction. The latter would induce genetically based time-profile differences in the longitudinal phenotype. We adopt a non-parametric statistical test to genetic epidemiological cohort studies and investigate its performance by simulation studies. The statistical test was originally developed for longitudinal clinical studies (Brunner, Munzel, Puri, 1999 J Multivariate Anal 70:286–317). It is non-parametric in the sense that no assumptions are made about the underlying distribution of the quantitative phenotype. Longitudinal observations belonging to the same individual can be arbitrarily dependent on one another for the different time points whereas trait observations of different individuals are independent.

The two loci are assumed to be statistically independent. They are modeled by two factors with arbitrary numbers of factor levels, respectively. The nonparametric null hypothesis of no interaction is formulated in terms of vanishing contrasts of the respective marginal distribution functions. It is tested with an ANOVA-type rank sum statistic.

132

Clinical factors in prostate cancer affected men with early age at onset: a comparison between African-American and Caucasian cases in Louisiana

D.M. Mandal(1), S.L. Halton(2), J.E. Bailey-Wilson(3), W. Rayford(4)
(1) Dept. of Genetics, LSU Health Sciences Center, New Orleans, LA, (2) Baylor Clinic, Houston, TX, (3) NHGRI/NIH, Bethesda, MD, (4) Dept. of Preventive Medicine, University of Tennessee, Memphis, TN

Incidence rate of prostate cancer in African-American (AA) men is twice as high as in Caucasian men. Earlier studies in the literature show that AA men are often diagnosed with advanced stages of prostate cancer and their age-specific PSA is higher than their Caucasian counterparts. We have identified 125 (60 AA and 65 Caucasian) males in our prostate cancer study with early age at onset (Ãœ 65 years) from Southern Louisiana. Family history was verified for all of them and pathological reports were reviewed. Data were analyzed to observe the distribution of age, PSA and Gleason Scores in two races. A significant difference ($p=0.0008$) in the age at onset values was observed between AA (range 38–65 years with median age of 57.5 years) and Caucasian (range 44–65 years with median age of 61 years) males. No statistical significance was observed in PSA and Gleason Score values in the two races. In further analyses, data were stratified with respect to sporadic cases and cases with family history of prostate cancer. AA and Caucasian cases with family history produced a significant difference of $p<0.01$ in the age at onset values. This result enables us to observe clinical characteristics of prostate cancer in different races in Louisiana males, which may contribute significantly in prostate cancer screening strategies.

133

The impact of genotyping errors on a Mantel statistic based haplotype sharing analysis

Marquard V(1), Beckmann L(1), Fischer C(2), Hein R(1), Rohde K(3), Chang-Claude J(1)
(1) German Cancer Research Center DKFZ, Heidelberg, Germany (2) University of Heidelberg, Heidelberg, Germany (3) Max-Dellbrück Center, Berlin-Buch, Germany

We analyzed the type I error rate and the power of a Mantel statistic based on haplotype sharing for simulated case-control data in the presence of genotyping errors. Haplotypes were estimated with an EM- algorithm and we focused on the results obtained via a Markov Chain Monte Carlo approach (MCMC) by sampling over all possible individual haplotype pairs according to their a posterior probability. Case-control data sets were simulated using haplotype distributions of *BRCA1* derived from the Hapmap project. Marker loci with different allele frequencies and different distances to the disease locus were chosen to have genotyping errors. Genotyping errors were simulated under different models (Leal, Genet Epidemiol 29, 204–214, 2005) and with differential and non-differential effects between cases and controls. With non-differential error rates, we observed generally a correct type I error rate, except for some scenarios with slightly conservative error rates. In the case of differential genotype errors, an inflated type I error was observed only for markers with a minor allele frequency of equal or less than 0.1 and very high genotype error rates, about 0.6 in cases and 0.1 in controls. This is in contrast to generally inflated type I error rates when using a haplotype based generalized linear model for analysis of association. In scenarios of low error rates, equal or less than 0.15, no influence on power could be detected.

134

SNPs and haplotypes in the 15-LOX gene are associated with intermediate phenotypes but not risk of coronary heart disease

P.A. McCaskie(1), C.M.L. Chapman(2), J.P. Beilby(2,3), J. Hung(4,5), B.M. McQuillan(4,5), K.W. Carter(1), P.L. Thompson(4), L.J. Palmer(1,5)
(1) Lab for Gen. Epi., WA Inst. for Medical Research, Uni of WA, Australia (2) Clinical Biochemistry, PathWest, Perth, Australia (3) School of Surgery and Pathology, Uni of WA, Australia (4) Heart Research Institute of WA, Perth, Australia (5) School of Medicine and Pharmacology, Uni of WA, Australia

Coronary heart disease (CHD) is a major health and economic burden, accounting for approximately 29.3% of deaths worldwide. The major underlying cause of CHD is atherosclerosis and oxidised LDL is thought to play an important role in its development. The specific mechanisms underlying CHD are poorly understood and research into the genetic determinants can provide important information to help understand its complexity. We analysed three SNPs and haplotypes in the promoter region of the 15-Lipoxygenase (15-LOX) gene in a community based study of 1,111 individuals and a sample of 556 CHD patients. Recessive associations of the GGG haplotype were seen with increased risk of carotid plaque in cases (OR=7.26, 95%CI=2.00-26.34, P=0.0004) and in healthy subjects under 53yrs (OR=4.00, 95%CI=1.16-13.88, P=0.03), an increased risk of infarction in cases (OR=2.89, 95%CI=1.05-7.90, P=0.025) as well as increased LDL in cases (P=0.02). No association was observed between 15-LOX haplotypes and risk of CHD. These findings may suggest a role of 15-LOX in disease progression, and suggest that genetic mechanisms of CHD are complex and intermediate phenotypes may act as poor surrogates for disease to study these mechanisms.

135

Too correlated to be true? Complete separation and the problem of "bouncing betas"

P.A. McCaskie, L.J. Palmer
Lab. for Gen Epi, WAIMR, Uni of WA, Australia

A common problem with modeling binary response variables arises when one or more covariates completely predicts the outcome of interest ('complete separation'). This leads to inflated beta coefficients ('bouncing betas') and SEs in a GLM framework. The most widely adopted solutions are to omit offending variables from the analyses or to modify the data in order to eliminate the separation. The former solution is unrealistic when a SNP or haplotype is the primary explanatory variable, and the latter has been demonstrated to perform poorly. Several other solutions have been suggested outside of genetics, the most promising of which accommodates multivatiate analysis by using maximum penalized-likelihood estimates (MPLEs) in place of the more commonly used maximum likelihood estimates (MLEs). These issues have not been addressed at all in a genetic setting. We simulated haplotypic associations with case-control status under a range of genetic models and sample sizes. As expected, the rate of perfect separation decreases as both sample size and haplotype frequency increase, however even under models powered to detect relative risks of>=2.0, massively inflated beta coefficients were observed in up to 11% of simulated data sets. Although we observed reduced inflation in the beta estimates when using MPLEs, the 'bouncing betas' remained a substantial problem. The effectiveness of MPLEs and other methods to reduce inflated estimates will be discussed. We conclude that perfect separation is an important problem in the context of genetic analysis that to date has been largely ignored and further work is needed.

136

Epistatic Interactions Associated with Left Ventricular Mass in African-Americans

KJ Meyers(1), TH Mosley(2), E Fox(2), E Boerwinkle(3), SLR Kardia(1)
(1) Dept of Epid, Univ of Michigan (2) Dept of Medicine, Univ of Mississippi Med Center (3) Dept of Human Genetics, Univ of Texas Health Sciences

Increased left ventricular mass (LVM) is a powerful predictor of cardiac mortality and morbidity. LVM is a complex, quantitative trait with well-established environmental and physiological risk factors. In spite of this, up to 75% of the inter-individual variation of LVM is unexplained and likely the consequence of complex genetic and environmental interactions. The Genetic Epidemiology Network of Arteriopathy study screened 1427 African-Americans from Jackson, Mississippi using echocardiography and genotyped 405 SNPs in 81 genes. To detect epistatic contributions to LVM variability, an over-parameterized linear model of SNP-SNP interactions was used to identify significant interactions ($p < 0.10$). These epistatic factors where then assessed for their predictive ability using 10 iterations of 4-fold cross validation. There were 7,421 SNP-SNP interactions (of 82,215 possible interactions) with p-values < 0.10, only 413 cross-validated. Cross validation is an additional discriminator between true and false positives because despite the strong linear relationship between p-values (-log p) and cross-validated predictive ability, p-values only explained 5.9% of its variability. There were five genes (TGFB3, AGT, HTR2B, ATP6V11, and MMP9) that were strongly associated with epistatic interactions underlying LVM. These results underscore the importance of robust statistical procedures for identifying combinations of genetic factors that act jointly to increase disease risk.

137

Combined Individual- and Family-level Association Analyses of Quantitative Traits

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

Genetic association studies may be performed using either a population- or a family-based design. However, the data available can include both families and unrelated individuals. To analyze quantitative traits us ing both population- and family-level data, we investigate a linear model that allows for between- and within-family orthogonal components. This model is an extension of the QTDT (Abecasis et al. 2000, Am J Hum Genet 66:279–292) that also includes individuals with no available relatives. Individuals with informative parental genotypes contribute to both the within- and between-family components. Singleton individuals and those with non-informative relatives contribute only to the between-family component. Each component provides a separate test for association and both components may be simultaneously examined in a joint test. For a quantitative trait, we simulated a range of scenarios including those with population stratification, to assess and compare the size and power of the between- and within-family tests and the joint test for association. In some cases, the association parameter estimate from the between-family component suffers from bias due to population structure and only the within-family component is valid.We examined a test for differences in the parameters provided by the between- and within-family components and suggest guidelines for when it is appropriate and useful to perform joint analyses.

138

Detecting Epistatic Needles in Genome-Wide Haystacks

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

The detection of epistasis is an important priority in the genetic analysis of common human diseases. The most challenging epistatic effects to model are those that do not exhibit significant marginal effects. Detecting epistasis in the context of genome-wide association studies is considered a *needle in a haystack* problem. Thus, it is unrealistic

to expect that stochastic search algorithms will do any better than a random search. Our goal was to develop and evaluate a stochastic search algorithm that is capable of finding epistatic needles in genome-wide haystacks with the assistance of expert knowledge. We first developed a genetic programming (GP) approach to picking SNPs for epistasis evaluation using multifactor dimensionality reduction (MDR). This GP-MDR approach was evaluated using simulated epistatic interactions of varying heritability (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4) and minor allele frequency (0.2, 0.4) that were embedded in genome-wide datasets with varying numbers of SNPs (1000, 10000, 100000) and varying sample sizes (200, 400, 800, 1600, 3200, 6400). We found no evidence to suggest that GP-MDR performed better than random search on these simulated genome-wide datasets. Next, we modified GP-MDR to select SNP combinations based on prior information about the quality of each SNP as assessed during a pre-processing analysis by the ReliefF filter algorithm. We found that the expert knowledge provided by ReliefF significantly improved the ability of GP-MDR to identify epistatic SNPs. An important advantage of this approach is that any form of expert knowledge could be used to guide the stochastic search algorithm.

139

Symbolic Modeling of Epistasis

J.H. Moore(1), N. Barney(1), C.-T. Tsai(2), F.-T. Chiang(2), B.C. White(1)
(1) Dept. of Genetics, Dartmouth Medical School, USA, (2) National Taiwan University Hospital, Taiwan

The workhorse of modern genetic analysis is the linear model. An important limitation of the linear model is that it makes assumptions about the nature of the data being modeled. This assumption may not be realistic for complex biological systems such as disease susceptibility where nonlinearities in the genotype to phenotype mapping relationship such as epistasis are the norm rather than the exception. We have previously developed a flexible data mining approach called symbolic discriminant analysis (SDA) that makes no assumptions about the patterns in the data. Rather, SDA lets the data dictate the size, shape, and complexity of a discriminant function that could include any set of mathematical functions from a list of candidates provided by the user (e.g. +, -, *, /, log, sqrt,>,<, max, min, IF, AND, OR, etc.). Here, we outline a new five step process for symbolic model discovery that uses genetic programming (GP) for coarse-grained stochastic searching, experimental design for parameter optimization, graphical modeling for generating expert knowledge, and estimation of distribution algorithms for fine-grained stochastic searching. Finally, we introduce function mapping as a new method for the visual interpretation of symbolic discriminant functions in the form of expression trees. We illustrate the ability of this approach to detect and interpret both linear (i.e. main effects) and nonlinear (i.e. epistatic effects) patterns in a genetic study of atrial fibrillation (n=500).

140

Increased detection of genetic association with disease by modelling gene-environment interaction

A.P. Morris
Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

The biological processes underlying complex disease phenotypes incorporate intricate interplay between multiple genes and exposure to environmental risk factors. In the presence of modifying non-genetic risk factors, we expect to increase our power to detect association with disease by modelling the gene-environment interaction over simpler tests that focus only on the marginal genetic effects. However, gene-environment interactions are often overlooked in association studies because of the fear that a less parsimonious model will lose power unless the effects are large. One solution to overcome this problem is to make use of Bayesian model averaging techniques, where the true underlying model of association between disease, environmental risk factors and a given single nucleotide polymorphism (SNP) is treated as unknown. The main effect of each environmental risk factor and their interaction with the SNP is assigned a prior probability of inclusion in the model of disease association. Markov chain Monte Carlo techniques are then used to sample from the posterior distribution of models consistent with the observed phenotype and genotype data, and hence to calculate a Bayes factor in favour of association of the SNP with disease.We present results of a detailed simulation study to demonstrate: (i) minimal loss of information by modelling gene-environment interaction when these effects do not exist; and (ii) substantial increases in power over simple tests of SNP association with disease when the genetic effect is modified by exposure to environmental risk factors.

141

A Re-examination of the Power of Multifactor Dimensionality Reduction in the Presence of Genetic Heterogeneity

A.A. Motsinger, T.J. Fanelli, and M.D. Ritchie
Vanderbilt University, Medical Center, Nashville, TN 37232

Genetic heterogeneity is a challenge for any analytical method in association studies of common, complex disease. A commonly cited article in Genetic Epidemiology: *Ritchie et al, Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003 Feb; 24(2):150–7* reported that the power of the Multifactor Dimensionality Reduction (MDR) method to detect gene-gene interactions in the presence of genetic heterogeneity was extremely low. Simulated data was used to examine the impact of 50% genetic heterogeneity (where 50% of cases were simulated as a result of one genetic model and 50% were due to another) on the performance of MDR in a range of gene-gene interaction models. The results presented in that article were based on an extremely conservative definition of power, where the

MDR method was required to find all loci from both genetic models to be considered successful. Recently, these results have been re-evaluated with a less stringent definition of power which is more in line with everyday practices. Here, we examine the power of MDR to detect any of the correct loci, not necessarily all of them. This examination reveals that the performance of MDR in the presence of genetic heterogeneity is significantly improved over what was previously reported. In many disease models, MDR is useful for the detection of gene-gene interactions, even in the presence of genetic heterogeneity.

142

Power of Grammatical Evolution Neural Networks to Detect Gene-Gene Interactions in the Presence of Error Common to Genetic Epidemiological Studies

A.A. Motsinger, T.J. Fanelli, and M.D. Ritchie
Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232

With the advent of increasingly efficient means to obtain genetic information, a great insurgence of data has resulted, leading to a need for useful and expeditious methods for analyzing this data beyond that of traditional parametric statistical approaches. Recently, we introduced a Grammatical Evolution Neural Network (GENN), a machine-learning approach to detect gene-gene or gene-environment interactions, also known as epistasis, in high dimensional genetic epidemiological data. GENN has been shown to be highly successful in a range of simulated data, but the impact of error common to real data is unknown. In the current study, we examine the power of GENN to detect interesting interactions in the presence of noise due to genotyping error, missing data, phenocopy, and genetic heterogeneity. We find that the GENN method is relatively robust to all error types including genetic heterogeneity.

143

A new test for hardy-weinberg disequilibrium using genotypes

A. Murphy(1), C. Lange(2), and S.T. Weiss(1)
(1) Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, (2) Dept of Biostatistics, Harvard School of Public Health, Boston, MA, USA

We introduce two new test statistics for Hardy-Weinberg disequilibrium testing that use only genotype frequencies in affected subjects, for two-sided and one-sided alternative hypotheses. These test statistics do not require specification of the marginal allele frequency of the disease susceptibility allele, which may be over- or underestimated in a completely ascertainedsample. We investigate the behavior of the test statistics under the alternative hypothesis. Additionally, the power of these test statistics is demonstrated via simulation studies, and we show that they have improved power in comparison to the standard testing methodology for detecting deviations from Hardy-Weinberg equilibriumin case-only designs. Additionally, we demonstrate the applicability of the new genotype-based test statistics in a candidate gene analysis for asthma.

144

Association Tests for Candidate Genes Based on Gibbs Random Fields Models

R.Z. Nickolov(1), V.B. Milanov(1), S. Zhang(2)
(1) Dept. of Mathematics and Computer Science, Fayetteville State University, Fayetteville, NC, USA, (2) Dept. of Mathematical Sciences, Michigan Technological University, USA

We propose novel statistical tests for candidate-gene association studies, which utilize genotypes and haplotypes of multiple tightly linked genetic markers under case-control design. These tests are generally applicable to markers with multiple alleles. However, the focus of this work is the application to biallelic markers: single nucleotide polymorphisms(SNPs). The proposed tests are powerful likelihood ratio tests, based on Gibbs random field models. The probability of observing a multi-marker genotype (haplotype) is modeled by the Gibbs distribution. This leads to models similar to those of Potts and Ising, both with and without nearest-neighbor interaction terms. Permutation testing is used to assess the statistical significance. The merit of our proposed tests is explored through simulation studies under a variety of scenarious. The type I error rate and power of the tests is compared with the type I error rate and power of commonly used tests for case-control data, such as Pearson's chi-square goodness-of-fit, two-sample Hotelling's T^2, etc. The simulation results show that the proposed tests have the correct type I error rate and that, in most cases, they are as powerful as or more powerful then the other methods considered.

145

Development and Evaluation of a Novel Method to Control for Population Stratification: Comparison of Genome Matching versus Structured Association Strategies

K.K. Nicodemus(1,2), D.R. Weinberger(2), Y. Yao(1)
(1) Epi, Johns Hopkins SPH, USA (2) GCAP/NIMH/NIH, USA

Several case-control studies have demonstrated undetected population stratification (PS) can lead to spurious association signals or mask or reverse true association. Structured association (SA) methods have been proposed to control for PS by assigning ancestry estimates in the statistical model. Genome Matching (GM), a simple way to control for PS by matching cases to controls using percentage of shared genotypes informative of population ancestry (unlinked to disease status), is proposed. We recently developed a novel SNP panel for detection of PS. Using allele frequencies from this panel and haplotype frequencies in a candidate gene in 2 populations

(European Americans (EA) and African Americans (AA)), we simulated 1000 replicates under the alternative hypothesis of both association and PS in 1000 cases and 1000 controls with EA and AA genomewide characteristics and applied GM and SA methods to control for PS. Two types of null hypotheses were simulated: H/01/: PS, no association and H/02/: no PS and no association. Under the joint null hypothesis H/02/, p-values are $\sim$ U(0,1). However, under H/01/, the distribution of p-values are inflated due to uncontrolled PS. Our main aims are: (1) to test for residual confounding due to PS after use of GM and SA methods under H/01/, (2) to evaluate power and type I error of GM and SA using H/02/ and replicates simulated under the alternative hypothesis, and (3) to determine the optimal number of SNPs and individuals for detection of PS with GM and SA approaches.

146

Association between telomere length and bone mineral density in the Amish

O.T. Njajou(1), B.D. Mitchell(2), R.M. Cawthon(3), S-H. Wu(1), C.M. Damcott(2), A.R. Shuldiner(2), E. Streeten(2), W-C. Hsueh(1)
(1) UCSF, SF, CA; (2) U Maryland, Baltimore, MD; (3) U Utah, Salt Lake City, UT, USA

Telomeres are DNA capping structures at the end of chromosomes and which shorten at each cell division in humans. Telomere length (TL) has been shown to be associated with aging and survival. However, it is not clear whether such an association is due to influences on susceptibility to specific diseases or aging more generally. In this study, we examined the relationship between TL and osteoarticular aging. We studied the association between TL and bone mineral density (BMD) using a large sample from a founder population. Our total sample included 816 Amish in a 5-generation pedigree (312 men and 504 women, aged 19 - 91 yrs, mean: 51 $\pm$ 16 yrs). TL in leukocytes was measured using a validated quantitative PCR (value range: 0.61 - 4.29 units, mean: 1.47 $\pm$ 0.50). BMD (in g/cm2) at 3 anatomic bone sites (lumbar spine, hip and forearm) was measured by DXA. All analyses were adjusted for BMI, sex-specific age, age2, conditioning on pedigree structures.We observed that TL was significantly associated with BMD at both forearm and hip, but not at the spine (n=426–784). Strongest associations were with BMD at forearm (ß=-0.02 $\pm$ 0.01, p=0.03 at one-third radius; ß=-0.02 $\pm$ 0.01, p=0.2 for ultradistal radius). Similarly, TL was also negatively associated with hip BMD, in particular at hip neck (ß=-0.02 $\pm$ 0.02, p=0.066). In summary, unexpectedly, we found that longer TL was associated with lower BMD. The mechanism(s) driving this association warrants further investigations

147

Polymorphisms in Polycyclic Aromatic Hydrocarbon (PAH) Metabolism and Conjugation Genes and PAH-DNA Adducts in Prostate Cancer

N.L. Nock(1), D. Tang(2), A. Rundle(3), C. Neslund-Dudas(4), B. Rybicki(4)
(1) Dept of Epi & Biostat, Case Western Reserve U, (2) Dept Env Health Sci and (3) Epi, Columbia U, (4) Dept of Biostat & Res Epi, Henry Ford Health System

We recently reported that PAH-DNA adducts, which may induce mutations that contribute to carcinogenesis, are present in human prostate cancer cells. In the present study, we investigated associations between functional polymorphisms in PAH metabolism (CYP1A1 Ile462Val, CYP1B1 Ala119Ser and Leu432Val, mEH Tyr113His and His139Arg) and conjugation (GSTP1 Ile105Val, GSTM1 and GSTT1 null deletions) genes and PAH-DNA adducts in prostate tumor and adjacent non-tumor cells of 410 men with prostate cancer. No associations were observed in the total sample, but stratifying by race revealed that, in Caucasians, carrying at least one copy of the high activity mEH 139Arg allele was positively associated with PAH-DNA adduct levels in tumor cells (p=0.008) and non-tumor cells (p=0.03) and, in African-Americans, having two copies of the high activity CYP1B1 432Val allele was positively associated with adduct levels in tumor cells (p=0.004). Having the genotype combination with the highest susceptibility to PAH DNA damage, fast metabolism of PAHs to reactive forms (CYP1B1 432 Val and mEH 139 Arg) with low capacity to conjugate these metabolites (GSTP1 105 Ile/Ile) was positively associated with PAH-DNA adduct levels in tumor cells of African-Americans (p=0.01) and non-tumor cells of Caucasians (p=0.02). This is the first report describing effects between genetic variants and PAH-DNA adducts in prostate cancer.

148

An alternative way of constructing ancestral graph by estimating marker allele ages from population linkage disequilibrium information

L. Park
Dept. of Neuroscience, Mayo Clinic Jacksonville, FL USA

To overcome limitations of the previous coalescence approaches, an alternative way of constructing ancestral graph is proposed here. Instead of constructing the genealogy by coalescence, the focus is to estimate the relative allele ages from linkage disequilibrium (LD) data in constructing the ancestral graph. The LD between two close variants decays over generation due to recombination. By assuming the fixed allele frequencies of variants and a fixed recombination rate, the current LD decay state indicates the past number of generations after the emergence of younger allele among two variants. To estimate relative allele ages of variants, the number of generations from the initial complete LD to current state is calculated for each two variants depending on the direction of LD decay between variants. The estimates can be the age of either variant among two. From a simple algorithmic procedure, the expected allele ages are assigned to each variant, and finally the ancestral graph is derived, which is similar to the basic coalescence with mutation rate. Applications to the simulated and real genotype data from public databases are available. The advantage of this method can be the flexible incorporation

of selection pressure and relatively direct interpretation of ancestral recombination events in the genealogical graph.

149

Familial aggregation of heart rate variability in Korean families

S. Park(1), H.-J. Lee(1), S.-I. Cho(1), H.-J. Jhun(2), D. Paek(1)
(1) Seoul National University School of Public Health, Korea (2) Hallym University Medical School, Korea

Heart rate variability (HRV) is an important risk factor for cardiovascular disease, and recently shown to have genetic component. However, there have been only limited number of family studies on HRV. We explored the familial correlations and heritability of heart rate variability in Korean families. A total of 286 families were included in the study, with 254 parent-offspring pairs and 44 sibling pairs. Heart rate variability was measured by 5 minute recordings from electronic heart rate monitoring device. Spectral analysis were performed to extract frequency domain parameters such as low frequency components (LF) reflecting sympathetic activity and high frequency components (HF) reflecting parasympathetic activity. FCOR and ASSOC programs of SAGE software were used for data analyses. Age, sex, and systolic and diastolic blood pressure were adjusted for in the analyses. Heritability was greatest (0.44) for low frequency/high frequency ratio, followed by LF (0.13) and HF (0.01). Correlations were assessed among the pairwise familial relationships. Sibling correlations tended to be greater than parent-offspring correlations in all measures. We conclude that there is a strong familial aggregagion of HRV, particularly LF/HF ratio that reflects sympathetic-parasympathetic balance in the autonomic nervous system.

150

Log-linear Modeling Approach for the identification of the gene-gene interactions in the presence of missing data

Taesung Park(1), Junghyun Namkung(2), Yujin Chung(3), and Seung Yeoun Lee(4)
(1) Deptartment of Statistics, Seoul National University, Seoul, Korea. (2) Bioinformatics program, Seoul National University, Seoul, Korea. (3) Department of Statistics, University of Wisconsin, USA. (4) Deptartment of Applied Mathematics, Sejong University, Seoul, Korea.

For common complex diseases such as diabetes and hypertensions, the effect of single genetic variation will be likely dependent on other genetic variations (gene-gene interaction) and environmental factors (gene-environment interaction). Thus, an identification and characterization of susceptibility genes is a challengeable task. To address this issue, the multifactor dimensionality reduction (MDR) has been proposed and implemented by Ritchie et al. (2001), Moore et al. (2002), Hahn et al. (2003) and Ritchie et al. (2003). Unfortunately, the original MDR approach cannot handle appropriately the incomplete data with missing observations, in which missing observations are treated as an additional genotype category. Furthermore, the original MDR approach may suffer from a sparseness problem due to the small number of sample size. The MDR approach is equivalent to the log-linear model approach using the saturated model. We propose using the non-saturated log-linear model to handle the incomplete data with missing observations. The proposed model reduces more effectively the dimension of multilocus genotype predictors from n to one, which improves the identification of polymorphism combinations associated with disease risk. The proposed log-linear model approach can also smooth the sparse cell frequencies. We compare the proposed method with the original MDR method through simulation studies and show that the proposed method has almost equivalent powers with less number of model parameters.

151

Genome-wide sparse canonical correlation analysis of relationships of genes with complex phenotypes

E. Parkhomenko(1), D. Tritchler(1,2), J. Beyene(1,3)
(1) Dept. of Public Health Sciences, Univ. of Toronto, Canada (2) Division of Epidemiology and Statistics, Ontario Cancer Institute, Canada (3) Hospital for Sick Children Research Institute, Canada

Genome-wide studies of the association of gene expression with multiple phenotypic measures may require the identification of complex multivariate relationships. Canonical correlation analysis (CCA) establishes relationships between linear combinations of all covariates and responses. However, in microarray data analysis and genome-wide linkage analysis the number of genes under consideration often exceeds tens of thousands. In these cases linear combinations of all features lack biological interpretability. In addition, insufficient sample size leads to computational problems, inaccurate estimates of parameters and non-generalizable results. Sparse canonical correlation analysis (SCCA) solves the problem of biological interpretability and provides results that are applicable outside the study. Sets of covariates and response variables with sparse loadings also comply with the belief that only a small proportion of genes are expressed under a certain set of conditions. We develop methodology and algorithms for SCCA. Simulation shows that SCCA has superior performance compared to conventional CCA under the assumption that a large proportion of genes are independent of measured phenotypes and other genes.

152

Integration of statistical genetics and bioinformatics tools in a QTL mapping workflow

J.M. Peralta(1,2), A. Buil(3), R. Souto(3), L. Almasy(2), J.M. Soria(3)
(1) CIBCM, UCR, San José, Costa Rica, (2) Genetics Dept., SFBR, San Antonio, TX, USA, (3) Institut de Recerca, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Data analysis usually requires that information flows through several disparate and specialized software tools. Integration of those tools in an analysis pipeline is not a trivial task, as it can be time consuming. Higher

interoperability between specialized software is needed in order to manage and analyze the vast amounts of highly diverse data produced by life sciences. As part of a wider initiative that aims to integrate statistical genetics and bioinformatics tools in a grid environment, we designed a workflow composed by processors of different natures to carry out a QTL mapping experiment. We created web service interfaces to SOLAR, a popular program used by statistical geneticists, and to R, a commonly used software package for statistical computing. Then we used them with the Taverna Workbench to perform linkage analysis of a quantitative trait, using families of the GAIT project. MIBD matrices, pedigrees, phenotypes and a SOLAR script (to specify details of the linkage analysis, such as covariables) were provided as inputs. As a result of the enactment of this single workflow we obtained a list of linkage signals from SOLAR, the pairs of markers surrounding each significant QTL from R, and lists of candidate loci from BioMart. This workflow allowed us to integrate in a single, portable and reusable module, a complex linkage analysis tool and common bioinformatics tools which are often used during the first follow up of linkage results.

153

A novel method to detect genetic heterogeneity in multifactorial diseases: the OTDT

H Perdry, MC Babron, V Marquegnies, M Bourgey, F Clerget-Darpoux
Hopital Paul Brousse Leriche, Villejuif , France

Taking into account clinical heterogeneity may help reduce genetic heterogeneity, thus increasing the power to detect the genetic factors involved in multifactorial diseases. The Ordered Subset Analysis (OSA) *Hauser et al. (2004) Genet Epidemiol 27:53–63* introduces a trait-related covariate in linkage analysis to identify more genetically homogeneous samples.We design the Ordered Transmission Disequilibrium Test (OTDT) to detect the role of a candidate gene in trio families. It combines the principle of both OSA and TDT *Spielman et al. (1993*) Am J Hum Genet 52:506–516. OTDT is a lod score that tests the following null hypothesis: no effect of the candidate gene or no correlation between the covariate and the candidate gene. The power of OTDT was evaluated by simulation under various genetic models and coveriate distributions. The method has a double advantage: it not only increases the power to detect the involvement of the candidate gene, but it also allows the identification of a clinical covariate reflecting genetic heterogeneity. This study was funded by ARSEP.

154

GOLDsurfer2: A comprehensive tool for the analysis and visualization of whole genome association studies

F. Pettersson(1), A.P. Morris(1), P.S. Derwent(2), M.R. Barnes(2), L.R. Cardon(1)
(1) Dept Bioinformatics, Wellcome Trust Centre, Oxford, United Kingdom; (2) GlaxoSmithKline Pharmaceuticals, Genetic Bioinformatics, Harlow, Essex, UK

With recent advances in the efficiency of high-throughput single nucleotide polymorphism (SNP) genotyping technology, genome-wide association studies are now routinely undertaken with the sample sizes necessary to detect the modest genetic effects we expect for complex diseases. There is now a clear demand for efficient analysis tools that allow data pre-treatment, together with evaluation, visualization and interpretation of results. To meet these demands, we have developed GOLDsurfer2. The program can be used for pre-filtering of genotype data, using user-defined quality control thresholds and for statistical association analysis. Interactivity in terms of both visualization and data management are key concepts in the user-friendly GUI. Basic statistical calculations, including single-locus and pairwise models of SNP association with disease, are built in. The architecture is written to accommodate methods for more complex analyses as external modules. GOLDsurfer2 can link to extract annotation from public databases such as the UCSC genome database and it also supports browsing of gene ontologies. A core feature is the 3D interactive visualisation of linkage disequilibrium (LD) which is implemented to function on a genome wide level. In order to find putative causal alleles we include a capability to functionally weight SNPs characterized by the International HapMap project. Distributions are available for Mac OSX, Linux and windows.

155

A Visualization Tool for Genetic Parameters in Complex Human Traits

Qin, S. Schmidt, M. A. Schmidt, E. Martin, E. R. Hauser
Center for Human Genetics, Duke University Medical Center, Durham, North Carolina

We have developed a graphical display tool called SIMLAPLOT for visualizing different ways in which continuous covariates may influence the risk of complex human diseases. The goal is to explore how genetic model parameters vary as a function of covariate values and to better understand the role of these parameters in simulation studies performed with our previously distributed software SIMLA. SIMLAPLOT uses a prospective logistic regression model as the penetrance function. It allows multivariate models including terms derived from genotypes at a known susceptibility locus or at arbitrary genotyped bi-allelic markers, covariate terms, and product terms modeling interaction between genotypes and covariates. Models include quantitative trait locus (QTL), gene-environment (GxE) interaction, and genetic main effects with heterogeneity. For a specified model, SIMLAPLOT calculates parameters by solving an equation including user-specified parameters: relative risk, mode of inheritance, and disease prevalence. SIMLAPLOT produces plots of conditional genotype probabilities as a function of covariate values, covariate distribution for each genotype in both affected and unaffected individuals, and genotype-specific penetrance values as a function of covariate values. SIMLAPLOT also may be used as a graphical tool for analyzing real datasets. Plots may suggest possible models for the data and allow

comparison to curves from specified models. These explorations may help identify model parameters for SIMLA and help to explore the results from statistical analysis.

156

Mantel-Haenszel approach to case-triad dataL.Qing, Q.Lu, R.Sinha, C.Xing, R.C.Elston

Dept. of Epi & Biostat, Case Western Reserve Univ., USA

In the past decade there has been a dramatic increase in the use of association studies for the genetic analysis of complex diseases. Family-based association studies, such as use the case-parents design, are one of the most popular forms of association studies. They have been widely used because the case-parents design has the advantage of robustness against unobserved genetic population structure. In addition, it is a l design that has power to detect parental imprinting. Current likelihood methods, which include the log-linear model *Weinberg et al 1998 Am.J.-Hum.Genet*, are the most appropriate methods for such designs. However, most of these methods are designed for detecting a major causative locus and are less suitable for complex diseases. Here we propose a Mantel-Haenszel pooled method. With a simple equation, one can quickly calculate the relative risk estimate by hand. We also extend the method to the complex disease situation where the disease may be caused by several loci and their interactions. For the one-locus model, we show in simulations with 250 case-triads that the Mantel-Haenszel pooled point estimate is comparable to the estimate obtained from the log-linear model, but is much easier and faster to compute.

157

Further exploration of: linkage statistics that model relationship uncertainty

Amrita Ray(1), Daniel E. Weeks(1,2)
(1) Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA (2) Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

We propose that statistically modeling relationship uncertainties by weighting over the possibilities maintains power in the presence of relationship errors in linkage analysis. Consider the situation where we have collected affected relative pair (ARP) data and carried out a genome-wide scan for linkage. We develop four relationship uncertainty linkage statistics (RULS). Relationship uncertainty is modeled via weights, the weights being the conditional probability of a true relationship type given the appa rent one and the genome-wide marker data. To assess the RULS and to compare them to MLS approach of Cordell et al. and exponential Sall LOD score, we performed a simulation study on two pedigree structures, small families (SP) and larger complex families (LP). For SP, we considered only full sibs as the apparent relationship type and for LP we considered several different apparent relationship types. To calculate MLS and Sall we construct true and discarded pedigree structures. We simulated data with 367 markers and a disease locus on chromosome 10 under several disease models. We compute the genome-wide empirical thresholds at 0.01 level of significance from 1,000 replicates. For both SP and LP, based on 400 replicates, the power of three of the RULS is between the power of the MLS and Sall LOD on true and discarded structures, implying that the RULS perform better than discarding individuals with erroneous relationships.

158

Blood Pressure Response to the Cold Pressor Test: Phenotype and Heritability Analysis

H Roy-Gagnon(1), MR Weir(2), JD Sorkin(2), W Post(3), K Ryan(2), BD Mitchell(2), AR Shuldiner(2), JA Douglas(1)
(1) Dept. of Human Genetics, Univ. of Michigan, USA (2) Dept. of Medicine, Univ. of Maryland, USA (3) Dept. of Medicine, Johns Hopkins Univ., USA

Blood pressure (BP) reactivity to the cold pressor test (CPT) has been shown to predict hypertension and cardiovascular events. Associations between BP recovery after the CPT and longitudinal changes in BP have also been found. Few studies have investigated the genetics of BP response to the CPT, and none have examined BP recovery. As part of the Heredity and Phenotype Intervention (HAPI) Heart study, we administered a 2.5-minute hand CPT to 511 participants from 130 Amish families. We used repeated BP measurements taken before, during and after the CPT to examine the heritability of traits measuring BP reactivity and recovery. We considered the following traits (among others): positive incremental area under the curve (iAUC), reactivity and recovery slopes, and maximum positive change from baseline. After transformation for normality and adjustment for baseline BP and relevant covariates, heritability estimates for systolic BP (SBP) reactivity ranged from 0.24 (SE=0.08) for the slope to 0.46 (SE=0.09) for the maximum change. Estimates were similar for diastolic BP (DBP) reactivity. Estimates for SBP recovery ranged from 0.17 (SE=0.08) for the iAUC to 0.35 (SE=0.09) for the slope, while estimates for DBP recovery ranged from 0.03 (SE=0.07) for the slope to 0.13 (SE=0.08) for the iAUC. We conclude that CPT-derived traits are moderately heritable and thus amenable to studies to identify genes that influence these traits.

159

Assessing SNP-SNP Interactions in the Presence of Missing Genotype Data

I. Ruczinski
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

The majority of SNP association studies are based on data with missing genotype information. The most common approach for dealing with those missing data is to omit the observations that have missing records in the model's covariates. This approach however can have severe shortcomings for the statistical inference, namely a potential bias in the parameter estimates and a loss of power. In particular, if SNP-SNP interactions are considered, the loss of power can become overwhelming. In this presentation,

we demonstrate some of these shortcomings using a population based case-control study, examining the association between polymorphisms in XPD and XRCC1 genes and breast cancer (Brewster et al., 2006). We compare several methods to address the missing data issue, and propose a novel tree-based imputation algorithm (Dai et al., 2006). We also demonstrate how this approach can be used to draw valid statistical inference in the assessment of SNP-SNP interactions, using the Logic regression methodology (Ruczinski et al, 2003). Brewster AM, Jorgensen TJ, Ruczinski I, ..., Helzlsouer KJ (2006). "Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to Breast Cancer." Breast Cancer Research and Treatment, 95(1): 73–80. Dai J, Ruczinski I, LeBlanc M, Kooperberg C (2006). "Imputation Methods to Improve Inference in SNP Association Studies". Genetic Epidemiology (to appear).Ruczinski I, Kooperberg C, LeBlanc M (2003). "Logic Regression". Journal of Computational and Graphical Statistics, 12(3): 475–511.

160

Simple Correction for Population Stratification in Case-Control Studies

G. A. Satten (1), M. P. Epstein (2), A. S. Allen (3)
(1) Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta GA (2) Dept. of Human Genetics, Emory Univ., Atlanta GA (3) Dept. of Biostats and Bioinformatics, Duke University, Durham NC

Population stratification remains an important issue in case-control studies of disease-gene association, even within populations thought to be homogeneous. Recently Campbell et al. (Nature Genetics 2005 37:868–72) showed a case-control study of European-Americans where stratification induced a spurious association between the LCT gene and tall/short status that was not resolved by either genomic control or structured association models. We propose a simple two-step procedure to control for population stratification. First we construct a model for the odds of disease given a set of marker loci (excluding loci of interest) that can provide information on population substructure. Then, we use the first-step model to assign subjects to strata based on their odds of disease given marker genotypes. These strata are then used to test disease/genotype association using logistic regression. The resulting test has proper size and good power, even in the presence of population stratification. Our approach is computationally simple, easy to implement, and substantially less model dependent than existing approaches for controlling stratification. When applied to the case-control data of Campbell et al. it finds no evidence of association between LCT and height after controlling for stratification. In simulation studies, the power of our approach is close to that of analyses where the latent subpopulation is known.

161

The first association study between G72/G30 and unipolar depression in a large sample of patients and controls of German descent

TG Schulze(1), L Beckmann (2),A Karpushova(3), S Hoefels(4), J Schumacher(5), MM NÃ¶then(3), S Cichon(3), M Rietschel(1)
(1) Div Genet Epid CIMH, Mannheim, Germany (2) DKFZ, Heidelberg, Germany (3) Life & Brain Ctr (4) Dpt of Psychiatry, (5) Inst Human Genet, Univ of Bonn, Bonn, Germany

G72/G30 is considered a strong susceptibility gene for both schizophrenia (SZ) and bipolar disorder (BD). We recently reported association between identical G72/G30-haplotypes and SZ, BD and panic disorder (PD). An association study on major depression has not yet been conducted. We studied a German sample of 500 MD patients and 1030 population-based controls. We were interested whether our previously identified risk haplotype of markers M22, M23, and M24 was also associated with MD. To further explore any relationship between G72/G30 and MD, we genotyped 9 additional SNPs highlighted in other studies. The haplotype C-C-T was significantly more frequent in MD patients than in controls (40.5 % vs. 36.1 %; p=0.027; OR=1.2). The exploratory analysis on 9 further G72/G30 SNPs of interests yielded significant associations for M12 (p=0.038), M14 (p=0.046), while M15 (p=0.090) and rs1935062 (p=0.055) did not reach significance. Permutation analysis adjusting for the 9 SNPs yielded a global p=0.176. This is the largest case-control study on G72/G30 and MD to date. We found an association between MD and the same risk haplotype that we previously found associated with SZ, BD, or PD. Given that depressive symptoms are present across these diagnostic groups, G72/G30 might predispose to core symptoms prevalent in all four disorders.

162

Genetic Relationships of plasma homocysteine level, IMT, and the ankle-brachial index in the Old Order Amish

H. Shen (1), L.F. Bielak (2), P.A. Peyser (2), C.M. Damcott (1), AR. Shuldiner (1), R. Horenstein (1), W.S. Post (3), M.R. Weir (1), B.D. Mitchell (1)
(1) Dept. of Medicine, Univ. of Maryland, USA, (2) Dept. of Epi, Univ. of Michigan, USA, (3) Dept. of Medicine, Johns Hopkins Univ., USA

Though homocysteine stimulates vascular smooth muscle cell growth *in vitro*, and elevated plasma total homocysteine levels are associated with increased carotid intimal-medial wall thickness (IMT) in asymptomatic subjects, little is known about the relation between homocysteine levels and peripheral arterial disease (PAD). To better understand the genetic contributions to homocysteine levels, IMT and ankle-brachial index (ABI), a measure of PAD, we used variance decomposition methods to assess the joint genetic contribution to these traits in 738 men and women from Amish families enrolled in Heredity and Phenotype Intervention (HAPI) Heart

Study. Homocysteine levels were significantly correlated with ABI (r=-0.12, p=0.001), but not IMT (r=0.06, p=0.15), after adjustment for age and gender. To assess genetic contributions, we used univariate and bivariate variance decomposition analyses implemented in the SOLAR program. There were significant genetic components for all three traits (ABI: h^2=0.34, $p<0.0001$; IMT: h^2=0.51, $p<0.0001$; and homocysteine: h^2=0.58, $p<0.0001$). Significant genetic correlations were observed for homocysteine level with both IMT (R_G=0.41, p=0.008) and ABI (R_G=-0.33, p=0.047). Taken together, these results indicate: 1) genes influence a moderate proportion of the variation in all three traits; and 2) some genes appear to in fluence jointly variation between homocysteine and each measure of subclinical atherosclerosis. Further research is needed to identify the common genes that constitute the genetic basis for variation in plasma homocysteine level, IMT, and development of P AD.

163

A NOVEL APPROACH TO DETECT PARENT-OF-ORIGIN EFFECTS FROM PEDIGREE DATA

Sanjay Shete(1), Robert C. Elston(2), and Yue Lu(1)
(1) Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, TX (2) Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, OH

The parent-of-origin phenomenon in humans is now well recognized, and the deregulation of imprinted genes has been implicated in a number of human diseases. Recently, several linkage analysis methods have been developed to allow for parent-of-origin effects in the analysis of pedigree data. However, in general, one does not know a priori if disease-causing loci are imprinted or not. Linkage methods that allow for imprinting can lose power if there is no imprinting. Conversely, linkage methods that do not allow for imprinting will lose power if there is imprinting because of penetrance values not being correctly specified. Therefore, it is important to know whether imprinting is a possible mode of disease inheritance before performing linkage analyses. In this paper, we describe a simple covariate-coding scheme to test for the presence of parent-of-origin effects and provide a formula for calculating parent-specific penetrance values prior to any linkage analysis. Our coding scheme was successful in detecting parent-of-origin effects and in leading to more accurately estimated penetrance values. The use of accurate penetrance values in a linkage analysis that allows for imprinting can provide higher power in the case of a disease locus that is imprinted.

164

Genetic determination of bone density at the forearm and heel in Korean population

E.-K. Shin, S.-I. Cho, T.-H. Kim, I.-K. Kim, D.-M. Paek
School of Public Health, Seoul National Univ., Korea

This study aimed to estimate the magnitude of genetic determination of site-specific bone mineral density (BMD) in Koreans. The study population includes 479 men and 477 women aged 5–79 years from 319 Korean families. We measured BMD at both sides of distal forearm and heel by a peripheral dual energy X-ray absorptiometry. The sides of forearm and heel were defined as dominant and non-dominant by the dominant hand. Using the S.A.G.E program ASSOC, we estimated the age, sex, weight and height-adjusted narrow sense heritability (h2) and the variance components of BMD. For BMD at the heel, the h2 (+/- SE) of dominant and non-dominant side were 0.63+/-0.06 and 0.56+/-0.07, respectively. Similarly, the h2 for distal forearm were 0.22+/-0.11 and 0.07+/-0.11, respectively. Both sides of the heel BMD demonstrated statistically significant h2 ($P<0.001$). However, for the distal forearm, only dominant-side showed marginal significance (P=0.05). For the heel, over half of the proportion of total variance explained by polygenic effect. Random environmental variances of heel were 24% at dominant side and 16% at non-dominant side. For the distal forearm, the sibling variances from total variance were 39% at dominant side and 56% at non-dominant-side. Similarly, the marital variances of distal forearm were 39% and 37%, respectively. We conclude that the polygenic effects play a major role and the random environmental effects significantly affect on the heel BMD. On the other hand, the household effect could be a significant factor on the forearm BMD in Koreans.

165

Evidence of pleiotropy for Blood Pressure and Sleep Apnea

M Sinha(1),EK Larkin(1,2),S Patel(3), RC Elston(1),S Redline(2)
(1) Department of Epidemiology and Biostatistics, (2) Department of Pediatrics (3) Department of Medicine, Case Western Reserve University

Genetic studies of hypertension (HTN) have provided conflicting results, perhaps indicative of etiological heterogeneity. The apnea hypopnea index (AHI), a quantitative measure of sleep apnea, is associated with HTN, and joint modeling of both traits may provide insight into their genetic pathways. As part of the Cleveland Family Study, blood pressure (BP) and AHI have been measured in 1400 individuals from families of both African American and Caucasian ancestries. After linear adjustment of systolic BP for age, sex and BMI, we used the residuals in a weighted Haseman-Elston regression with both full- and half-sib pairs. We obtained a nominal p-value of 0.00169 at a marker on chromosome 8. After additional pre-adjustment of systolic BP with AHI, the p-value increased to 0.0708. For a statistical test of this difference, we performed a novel two-sample t-test for correlated samples, estimating the correlation in an approximate fashion. Then a conservative permutation test, permuting the alleles shared IBD of only the full sibpairs within a sibship and across sibships of same size was performed to determine the statistical significance of this test. A simulation study is being performed to study the type I error rate and the power of this new test. The reduced evidence for linkage at chromosome 8 indicates that this

locus may harbor genes that influence the expression of both HTN and sleep apnea through common pathways. We also have some weaker evidence of pleiotropy at other regions on the genome.

166

Efficient Intermediate Fine Mapping: Confidence Set Inference with Likelihood Ratio Test Statistic

R. Sinha and Y. Luo
Dept. of Epi. and Biostat., Case Western Reserve Univ., USA

In positional cloning of disease causing genes, identification of a linked chromosomal region via linkage studies is often followed by fine mapping via association studies. Efficiency can be gained with an intermediate step where confidence regions for the locations of disease genes are constructed. The Confidence Set Inference (CSI) *Papachristou and Lin, 2006* achieves this goal by replacing the traditional null hypotheses of no linkage with a new set of null hypotheses where the chromosomal position under consideration is in tight linkage with a trait locus. This approach was shown to perform favorably compared to several competing methods. Using the duality of confidence sets and hypothesis testing, CSI was proposed for the Mean test statistics with affected sibpair data (CSI-Mean). We postulate that more efficient confidence sets will result if more efficient test statistics are used in the CSI framework. One promising candidate, the Maximum Lod Score (MLS) statistic, makes maximum use of available identity by descent information, in addition to handling markers with incomplete polymorphism naturally. We propose a procedure that tests the CSI null hypotheses using the MLS statistic (CSI-MLS). Compared to CSI-Mean, CSI-MLS provides tighter confidence regions over a range of single and multi-locus disease models. The MLS test is also shown to be more powerful than the Mean test in testing the CSI null over a wide range of disease models, the advantage being most pronounced for recessive models. In addition, CSI-MLS is computationally much more efficient than CSI-mean.

167

Interacting Genetic Factors Influencing Fasting VLDL Response to Fenofibrate

JA Smith(1), DK Arnett(2), RJ Kelly(1), JM Ordovas(3), YV Sun(1), PN Hopkins(4), JE Hixson(5), JM Peacock(6), SLR Kardia(1)
(1) Dept of Epid, Univ of MI, (2) Dept of Epid, Univ of AL-Birmingham, (3) Nutrition and Genomics Lab, Tufts Univ, (4) Cardiovascular Genetics Res., Univ of UT, (5) Human Genetics Ctr, Univ of TX, (6) Dept of Epid, Univ of MN

Metabolic response to the triglyceride (TG)-lowering drug, fenofibrate, is shaped by interactions between SNPs and environmental factors, yet knowledge regarding the genetic determinants of this response is primarily limited to single gene effects. Since very low density lipoprotein (VLDL) is the central carrier of fasting TG, identifying gene-gene (epistatic) and gene-environment interactions that affect VLDL response to fenofibrate is critical for predicting individual fenofibrate response. As part of the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study, 792 individuals from 168 families were genotyped for 91 SNPs in 25 candidate genes. We examined epistatic and gene-environment interactions between these SNPs and clinical covariates for association with VLDL response. A ten-iteration four-fold cross-validation scheme was used to verify significant associations. Interaction patterns that affect response are presented via a multi-dimensional visualization system. SNPs in APOA1, APOA5 and APOC3 each exhibited interactions with SNPs in other genes, including ABCA1, NOS3, and LIPG, as well as with clinical covariates such as waist-to-hip ratio. The patterns yield insight into the complex biology of metabolic response to fenofibrate, which can be used to target fenofibrate therapy to individuals.

168

Comparison of modeling strategies for mapping multivariate phenotypes

C. M. Stein, N. J. Morris, T. Wang, R. C. Elston
Department of Epidemiology and Biostatistics, Case Western Reserve University

Many complex traits are manifested through several related phenotypes. Conducting a linkage analysis for these traits can be difficult because the causal genetic loci are expected to induce correlations between these phenotypes. Typically, investigators perform univariate analyses, ignoring the underlying multivariate structure of the related phenotypes. However, fully multivariate approaches can prove difficult when appropriate distributional assumptions are unknown. Using the Genetic Analysis Workshop 12 data, we compared the performance of a variety of linkage analysis strategies for their ability to detect loci influencing correlated phenotypes. The traditional methods evaluated included univariate analysis and principal component analysis; we compared these to a linkage version of pedigree discriminant analysis, a technique in which we estimate the linear function of phenotypes that best discriminate between the genotypes at a locus, and to the use of a fully multivariate linkage analysis model recently developed by Wang and Elston (Ann Hum Genet, in press).

169

On Using Linkage Signals to Improve Genome-Wide Association Studies - Weighted or Stratified False Discovery Control?

L. Sun
(1) Dept. of Public Health Sciences, University of Toronto, Canada, (2) Program in Genetics and Genomic Biology, Hospital for Sick Children, Canada, (3) Dept. of Statistics, University of Toronto, Canada

The goal of this study is to investigate two different ways of utilizing prior linkage information in the context of Genome-Wide Association (GWA) studies. Roeder et al. (2006) proposed a weighted p-value approach in which the

linkage signals (e.g. z values) are used to weight the association p-values (i.e. p=p/w, where w is proportional to z). They showed that the proposed method improves power considerably when the linkage study is informative, and the loss of power is small when the linkage study is uninformative, where power is measured in the context of multiple hypothesis testing using the False Discovery Control (FDR) framework. Sun et al. (2006) proposed a stratified false discovery control method that incorporates any form of auxiliary information by defining strata indicators. Both their analytic results and application demonstrate the potential advantages of control or estimation of FDR by stratum. The stratified FDR method can be easily applied to the above linkage+association setting in which one could use the linkage signals linkage+ to define strata (e.g. stratum 1 contains SNPs with z>a pre-determined threshold, and stratum 2 contains the remaining SNPs.) I will compare the performance of the above two methods through simulation studies and attempt to make the connection between the two.

170

Predicting Coronary Artery Calcification using Machine Learning Algorithms

Y.V.Sun(1), L.F.Bielak(1), P.A.Peyser(1), S.T.Turner(2), P.F.Sheedy,II(3), E.Boerwinkle(4), S.L.R.Kardia(1)
(1) Dept. of Epidemiology, Univ. of Michigan (2) Div. of Nephrology and Hypertension, Mayo Clinic (3) Div. of Radiology, Mayo Clinic (4) Hum Genet Ctr, Univ. of Texas Health Sciences Center

As part of the Genetic Epidemiology Network of Arteriopathy(GENOA) study, 367 hypertensive non-Hispanic white sibships were screened using 470 SNPs to identify genes influencing coronary artery calcification(CAC) measured by electron beam computed tomography. Individuals with CAC scores>70th percentile for their age and sex were classified as having a high CAC burden and compared to individuals with CAC<70th percentile. Two sibs from each sibship were chosen at random and divided into two samples, each with 367 unrelated individuals. Within each sample, we applied two machine learning algorithms, Random Forests (RF) and RuleFit, to build prediction models and to identify the best predictors among 12 covariates and 470 SNPs. Using 4-fold cross validation to evaluate the receiver-operator characteristics of each algorithm, both methods had ~70% sensitivity and ~60% specificity. For RF, among the top 50 predictors, there were 4 SNPs and 10 covariates in common across the two samples. Among the top 50 predictors from RuleFit, there were 5 SNPs and 7 covariates in common across samples. Replicable effects of 3 SNPs(in genes AGER,GPC6 and IL1B) and 7 covariates(age,BMI,sex,serum glucose,HDL, systolic blood pressure and cholesterol/HDL ratio) were identified by both methods. This study illustrates how machine learning methods can be used in affected sibpair samples to identify replicable genetic models for complex diseases.

171

Genetic Abnormality, Tobacco Smoke, Lung Disease, and Lung Cancer Risk

Sun Z, de Andrade M, Krowka MJ, Aubry MC, Scanlon PD, Bamlet WR, Wampfler JA, Thibodeau SN, Katzmann JA, Allen MS, Midthun DE, Marks RS, Yang P
Mayo Clinic College of Medicine, MN, USA

Genetic susceptibility in lung cancer risk has long been recognized but remains ill-defined. We hypothesized a pathway from tobacco smoke, genetic predisposition, and presence of emphysema and or chronic bronchitis (COPD) to lung cancer. Using a dual case control design, we tested if Ü-1 antitrypsin deficiency (Ü1ATD, PI1 locus) carriers are predisposed to a higher risk of lung cancer by applying multiple regression models and also examining the effects of tobacco exposure history and COPD. We first modeled 1,585 unrelated case-control pairs without the PI1 variable and confirmed previously reported results: people with COPD have a 4-fold higher lung cancer risk. Smokers are at a 2-11 fold higher risk than never smokers. In never smokers, passive smoking increases risk by 30%. With the PI1 allele types included in the second model, we demonstrated that Ü1ATD carriers are at 80% higher risk to develop lung cancer than non-carriers. Applying the generalized estimating equations using 488 probands and 902 siblings, we found a 2-fold higher lung cancer risk in 1ATD carriers (95% C.I., 1.5-2.7), further supporting the role of PI1 status in lung cancer risk. Stratified analysis by histology showed varied specific lung cancer risks associated with Ü1ATD carriers. There was no increased risk for small cell lung cancer and a significant increase for adenocarcinoma, particularly bronchioloalveolar carcinoma. Our results suggest that Ü1ATD carriers may account 9-12% of lung cancer in this US Midwest population.

172

Power comparison of different strategies to detect gene-gene and gene-environmental interactions in candidate gene association studies

N. Tanaka
Dept. of Clinical Bioinformatics, Grad. Sch. of Medicine, Univ. of Tokyo, Japan

Recently, several methods have been proposed to identify gene-gene and gene-environmental interactions from high dimensional data from association studies for common diseases. Here, I compare four of these methods, traditional logistic regression analysis, mixed modeling approach, extension of a multifactor dimensionality reduction (MDR) method, and a new method which is composed of combined resampling methods and semi-parametric additive mixed modeling. These methods are applied to the simulated data based on the real data on 27genes with different patterns of LD and different numbers of SNPs from the Internet-based database of Japanese SNPs for geriatric research (JG-SNP). A measure of the efficiency of the different methods is explored under scenarios considered to be relevant in the context of complex diseases.

173

Method for Estimating and Testing Ancestral Population Specific Effects in Genetic Association Studies in Admixed Populations

B.O. Tayo(1,2), X. Zhu(1), Y. Liang(2), M. Trevisan(3), A. Luke(2), R. S. Cooper(1)
(1) Dept. of Preventive Medicine & Epidemiology, Loyola University Chicago Stritch School of Medicine, USA, (2) Dept. of Biostatistics, SUNY at Buffalo, USA, (3) Dept. of Social & Preventive Medicine, SUNY at Buffalo, USA

Many analytical methods have been proposed for testing and accounting for population stratification especially in population-based genetic association studies. However, not much attention has been given to estimating the ancestral population specific effects in association studies in admixed populations. Using admixture probabilities, we describe a method that enables estimation of ancestral population specific effects in both genome-wide and candidate gene association studies in admixed populations. Our method uses estimates of admixture probabilities both to correct for population stratification or admixture, and also to estimate and test the ancestral population specific effects. We present results from the application of this method to simulated data on candidate gene association study in admixed population.

174

Family based association analysis of a copy number polymorphism in asthma families

MD Teare(1), I Sabroe(1), J Goeke(2), M Whyte(1)
(1) University of Sheffield Medical School, UK (2) Freie Universitat Berlin

Genetic variation in the chemokine system may account for differences in response to infection and influence the course of autoimmune and inflammatory disease. Gonzalez et al *2005* have shown a relation ship between CCL3-L1 copy number and progression of disease in HIV. Thus, genetic variation in CC L3-L1 copy may affect susceptibility to or the progression of diseases in which this chemokine plays a role. We have analysed this candidate locus in a series of 201 asthma families ascertained due to the presence of at least two affected sib pairs. As the insertion/deletion sequence which is copied in this gene is large, the copy number has been assessed by replicate qPCR assays. Hence only the total copy number is estimated and alleles are not observed directly. We present a WinBUGs implementation for the analysis of these data which can be applied to general nuclear families. Gonzalez et al. 2005 Science, 307, 1435.

175

Genome-wide linkage analyses for asthma predisposition loci in extended Utah pedigrees

C.C. Teerlink, N.J. Camp, L.A. Cannon-Albright

Dept. of Biomedical Informatics, University of Utah, USA.
Asthma is a multi-factorial disease with undetermined genetic factors. We performed a genome-wide scan to identify predisposition loci for asthma, using 540 STR markers on autosomal chromosomes. For all analyses, the asthma phenotype used was derived from physician-confirmed presence or absence of asthma symptoms. We used 82 extended Utah pedigrees ranging from three to six generations with 746 affected individuals, ranging from two to 40 per pedigree. We performed parametric multipoint linkage analyses with dominant and recessive models. Our primary analysis revealed suggestive evidence of linkage to regions 5q (LOD=3.01) and 11q (LOD=1.70), both occurring with a recessive model. Marginal evidence of linkage was found on chromosome 19q (LOD=1.14), which included a single pedigree with a pedigree-specific LOD of 2.48. To follow up to this result, we added 179 genetic markers to the chromosome 19 region, which increased the overall LOD evidence for this region to 2.59. In an attempt to counter possible intrafamilial heterogeneity, we split pedigrees to a maximum of three, four, five and six generations, and performed secondary linkage analyses. Of these analyses, the three-generation pedigree structures provided suggestive evidence to chromosome 4p (LOD=2.65). All of the regions indicated in these analyses (4p, 5q, 11q and 19q) confirm previously reported evidence for linkage to asthma phenotypes. These results indicate that the Utah extended pedigrees are useful in confirmation of regions of interest and support further investigation of these regions.

176

In silico genome mismatch scanning. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays

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

We examine the utility of high density genotype assays for predisposition gene localization using large extended pedigrees.In particular we focus on robust methods for identifying regions shared identical by descent across sets of distantly related cases.Results for the distribution of the number and length of genomic segments shared identically by descent among relatives,previously derived in the context of genomic mismatch scanning, are revisited and used to assess significance. Identity by state is evaluated by simulation.Methods are illustrated by analysis of an extended prostate cancer pedigree previously reported to show linkage to chromosome 1p. 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 offerpowerful designs for gene detection based on these statistics.

177

No Association between UGT1A6 Genetic Polymorphisms or Interaction with NSAID use and Colon Cancer

C.L. Thompson(1,2), S. Plummer(3), T.C. Tucker(4), G. Casey(3), L. Li(1,2)
Depts of (1) Epi and Biostats and (2) Family Med, Case Western Reserve Univ, (3) Dept of Cancer Biology,

Cleveland Clinic Foundation, (4) Markey Cancer Center, Univ of KY

Two recent studies report an interesting interaction between UDP-glucuronosyltransferase 1 (UGT1A6), a rate-limiting enzyme in aspirin metabolism, gene polymorphisms and non-steroidal anti-inflammatory drugs (NSAIDs) and the development of colon polyps, a well established precursor of colon cancer. However, others did not observe this interaction with respect to colon cancer. We sought to further assess this interaction in a population-based case-control study of colon cancer. We included 341 incident colon cancer cases and 441 population controls in the analysis. Two single nucleotide polymorphisms (SNPs) in the UGT1A6 gene were genotyped. Prior aspirin or ibuprophen use and lifestyle information were collected via a self-administered questionnaire. Risk for colon cancer was estimated in a multivariate logistic regression controlling for age, BMI, family history, gender and race. Consistent with previous reports, regular NSAID use (at least one pill a day for at least six months) was found to significantly protect against colon cancer (p=0.008). However, neither the main effect of either SNP in UGT1A6 nor their interaction with regular NSAID use were significant at the 0.05 level. While we cannot exclude other variants within the UGT1A6 gene, these SNPs, which were previously shown to modify the effect of NSAIDs, do not appear to influence susceptibility to colon cancer or modify the effect of NSAIDs in this population.

178

Meta analysis of relative predispositional effects of HLA DR-DQ genes and type 1 diabetes

G. Thomson, A. M. Valdes
Dept. Int. Biol., Univ. Cal. Berkeley, USA The HLA DRB1 and DQB1 genes represent the major genetic susceptibility to type 1 diabetes (T1D)

The DR-DQ genes display a complex hierarchy of predisposing, intermediate, and protective effects at the genotype and haplotype levels. There are marked ethnic differences in DRB1-DQB1 frequencies and these correlate with disease prevalence. The question we have addressed is the following: Do the DR-DQ haplotypes, and genotypes, show the same relative predispositional effects (RPEs) across all populations and ethnic groups? If they do, then the RPEs can be used to identify the amino acids involved in T1D. But the RPEs may differ if there are additional disease predisposing genes in the HLA region, which may show haplotype specific, regional and ethnic effects.A total of 45 T1D data sets have been analyzed. One statistic used to determine relative ranks is the Patient/Control (P/C) ratio, which is the MLE of the relative penetrances of the haplotypes (and genotypes), given a direct role of DR-DQ in disease. Categories of consistent predisposing, neutral, and protective haplotypes are identified which correlate with disease prevalence. Specific effects are also identified. For example, for predisposing haplotypes there is a statistically significant and consistent hierarchy for DR4 DQB1*0302 effects. There is also significant and consistent heterogeneity of DR4 DQB1*0301 and DR8 haplotypes within populations. Some haplotypes show different risks in different ethnic groups; these haplotypes are candidates for study of additional disease modifying genes in the HLA region.

179

Body mass index (BMI) interacts with P-selectin haplotypes derived from the S290N and N562D sites to modulate the susceptibility to myocardial infarction (MI)

D-A. Tregouet, C. Proust, V. Nicaud, L. Tiret, F.
Cambien INSERM U525, Paris,France

P-selectin is a cellular adhesion molecule expressed by platelets and activated endothelial cells that participates in the recruitment of leukocytes on the vessel wall. Several studies have suggested that it was involved in the development of atherosclerosis and its complication. Consistent with this hypothesis, we had previously observed that the presence on the same haplotype of two asparagine amino acids at sites S290N and N562D was associated with a higher risk of MI (OR=1.57 *1.08–2.26* p=0.017) in a case-control study of 1261 subjects from the ECTIM Study (Tregouet et al. Hum Mol Genet 2002). We attempted to replicate this finding in an independent sample of 560 cases and 526 controls from the ECTIM study who were not part of the initial study. As initially observed, the N290/N562 haplotype was more frequent in cases than in controls (0.108 vs 0.078) and was associated with an increased risk of MI (OR=1.45 *1.03–2.05*,p=0.032). Finally, in the whole ECTIM sample (1164 cases and 1183 controls), the N290/N562 was associated with an OR of 1.51 *1.18–1.92* ($p < 10^{-3}$). When dividing the population according to the population-specific median of BMI, the N290/N562 haplotype was at increased risk of MI only in individuals with low BMI (OR=2.16 *1.48-3.16*,$p < 10^{-4}$) while no association was observed in subjects with high BMI (OR=1.05 *0.76-1.46* p=0.769), these 2 ORs being significantly different (p=0.005). These results suggest an interaction between BMI and P-selectin haplotypes on the risk of MI.

180

Power and Type I Error in Comparison of Linkage Analysis Methods for Complex Qualitative Traits with Rare Disease Alleles and Environmental 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; (2) NHGRI/NIH, Baltimore, MD; (3) LSUHSC, New Orleans, LA

Previously work has been presented suggesting that, when a good estimate of the heredity model is available, model-dependent lod-score methods (in LODLINK) provide the greatest power when analyzing a qualitative disease/trait where disease etiology involves covariates and relatively common disease alleles. However, with an uncertain heredity model, sib-pair analysis with the squared sib-pair trait difference as the dependent variable would be a good alternative. Use of 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, all four linkage analysis methods were compared to study the effect of the inclusion of covariates on the Type I error rates and power of linkage analysis of a qualitative trait with rare disease alleles. Type I error rates were not inflated 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 (approaching 0.003 at the 0.001 nominal p-value and 0.11 at the 0.05 nominal p-value). Power gains when including covariates were impressive with rare disease allele frequencies only when LODLINK was used.

181

Bayesian Logistic Regression: Model Selection for Genetic Association Studies

H.-W. Uh, B.J.A. Mertens
Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

Our interest is the identification of genes that influence the risk of common and complex disease. We propose using auxiliary variable approaches for inference in Bayesian logistic regression, including covariate set uncertainty. Instead of testing each variant independently we consider the analysis of multilocus genotype data. The aim is to select a linear combination of single-nucleotide polymorphisms (SNPs) that would best describe the relation between genetic variation and a trait value. Our method is a fully Bayesian logistic regression model with hierarchical specification on the regression parameter vector. Model estimation proceeds via birth-death Markov Chain Monte Carlo (MCMC), which automatically accounts for variable selection. The potential of this method is flexible. It can be applied to complex high-dimensionality data: both to a gene-based approach in which all common variation within a candidate gene is considered jointly, and to some stage of genome-wide association studies.

182

Disentangling HLA Associations: Multivariate Association Study using Bayesian Logistic Regression

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

Diseases are often associated with the human leukocyte antigen (HLA) region on chromosome 6. It is becoming evident that this region also plays an important role in drug response. The presence of high linkage disequilibrium (LD) means that the finding of causal elements within the HLA region has been difficult. When LD is strong, different loci can show a similar degree of association, making it difficult to identify the causal variant(s). Logistic regression is used in association studies to predict disease risk and identify associated markers. In the presence of strong LD, highly-correlated predictors can cause some problems. One way to address these problems involves stepwise procedures, in which terms are added or removed sequentially to find a good model. However, in the presence of large numbers of SNPs, the variable selection process may be unstable (small changes in the data lead to very different models). Further, the final model can retain sets of highly-correlated predictors even when they add little to its predictive power. We adopt a Bayesian approach, which can better deal with the problem of multiple, correlated predictors. Our Laplace prior for each regression coefficient has mode zero. Typically, all but one of a set of correlated predictors will have posterior mode equal to zero. Terms with non-zero posterior modes indicate marker-disease associations, and the posterior mode for the regression coefficient gives a measure of effect size. Alternatively, the corresponding terms can be refitted with non-significant terms excluded. The approach involves a tuning parameter, the prior variance, which we selected using a permutation procedure to assess type-1 error. We compared this approach with the stepwise logistic regression using data arising from an association study on rheumatoid arthritis, with 842 cases, 957 controls and 2,302 SNPs from the Illumina MHC panel.

183

Is There Genetic Regulation of Hepatic Inflammation in the Metabolic Syndrome?

T. Yonebayashi(1), X. Guo(1), L. Steiner(2), J.M. Lee(2), B. Fang(1), A.H. Xiang(3), S. Cheng(2), M.J. Quinones (4), K.D. Taylor(1), W.A. Hsueh(4), T.A. Buchanan(3), J.I. Rotter (1), L.J. Raffel(1)
(1) Cedars-Sinai Med Ctr, (2) Roche Mol Syst, (3) Univ So Cal, (4) UCLA, USA

Purpose: In previous studies in Mexican American (MA) hypertensive families, we found insulin resistance (IR) to correlate with liver function. Others have also found a relation between liver function and the metabolic syndrome. We observed associations between variations in inflammatory genes and IR and thus investigated genetic determinants of liver function. We here report our studies of association between liver enzymes and inflammatory genes in MA at risk for IR.Methods: 599 healthy, non-diabetic offspring from 153 Hispanic American families were ascertained through probands with hypertension. Plasma liver function tests (LFTs) were measured and 42 SNPs in 31 inflammatory genes were genotyped by multiplexed PCR and sequence-specific oligonucleotide probe arrays. Associations were examined using quantitative-trait transmission disequilibrium tests as implemented in the QTDT program. The method that considers the variance within families was utilized, with adjustments for age, gender, and body mass index. Results: Associations were observed between alkaline phosphatase levels and polymorphisms in *IL1B*, *IL4*, and *IL10* (p=0.0008, 0.007, 0.03, respectively). *IL1B* variation was also associated with lactate dehydrogenase levels (p=0.009) and aspartate transaminase (AST) levels (p=0.04). Though not statistically significant, there was a trend toward an *ICAM1* and AST association (p=0.06). Conclusions: Inflammatory

genes play a role in variation in liver enzyme levels in MA hypertensive families. Further investigation is needed to better understand the relationships among inflammatory processes, liver function, and the metabolic syndrome.

184

Sample Selection to Perform Association Studies for Quantitative-Trait Loci

T. Wang, R.C. Elston
Dept. of Epi & Biostat, Case Western Reserve Univ., USA

In a quantitative trait association study, appropriate sample selection is important to improve the power to detect association. Slatkin (1999) adapted a case-control study to quantitative traits by using truncated selection (TS), a technique in which a selected sample (e.g. for high values) and a population sample correspond to the cases and controls, respectively. Chen et al. (2005) extended the TS approach to a two-sided selection approach (t-TS) by using a sample selected for low values rather than a random sample as controls, to further improve power. We studied analytically the properties of various selection strategies, including random selection, TS, t-TS and a one-sided selection (o-TS) approach. We found o-TS is the most powerful strategy for a trait locus segregating a rare-frequency allele with large effect, and t-TS is the most powerful for a trait locus segregating common alleles with similar effects. Further simulation studies confirmed our results that no single selection approach is uniformly optimal, and therefore it is critical to choose selection criteria carefully for a specific study.

185

A Genome-Wide Scan for Liver Enzyme Levels in the Mexican-American Coronary Artery Disease Study

D. Wang(1), X. Guo(1), K.D. Taylor(1), H. Yang(1), M. Quiñones(2), W.A. Hsueh(2), J.I. Rotter(1)
(1) Cedars-Sinai, L.A., CA, USA, (2) UCLA, L.A., CA, USA

Alkaline phosphatase (ALP), aspartate transaminase (AST), and alanine transaminase (ALT) are commonly used biomarkers in detecting liver damage of all degrees. Clinical and epidemiological evidences suggest that several coronary artery disease (CAD) risk factors, such as obesity, diabetes, and elevated triglycerides, may be associated with nonalcoholic fatty liver, which manifests as elevated liver enzymes. To identify genetic loci that contribute to the variation in liver enzymes, we performed a genome-wide scan for ALP, AST, and ALT in 101 two-generation Mexican-American (MA) families ascertained via a parent diagnosed with CAD. Liver enzymes were measured in 303 apparently healthy adult offspring. 461 subjects, including adult offspring and their parents, were genotyped using Marshfield screen set 12 (408 micro-satellite markers at $\sim$10cM interval). Heritability estimates and multipoint linkage analysis was performed using the variance component method implemented in SOLAR. The estimated heritability was 0.75 ($p=2x10^{-11}$) for ALP, 0.64 ($p=1x10^{-7}$) for AST, and 0.44 (p=0.0002) for ALT. The strongest evidence for linkage was observed with ALP on chromosome (chr) 7 at 104cM (LOD=3.2). Additional evidence for linkage was found on chr 21 at 17cM (LOD=2.0) for ALP, on chr 11 at 17cM (LOD=2.8) and chr 22 at 22cM (LOD=2.2) for AST, and on chr 19 at 68cM for both AST and ALT (LOD=2.2 for AST and LOD=2.3 for ALT). Our results indicated a strong genetic effect for variation in liver enzymes in MA families at risk for CAD.

186

Predicting Germline p16 Mutational Status within Melanoma Families using MELAPRO

W Wang(1), K.B. Niendorf(3), D. Patel(3), F. Marroni(4), G. Parmigiani(1,2) and H. Tsao(3)
(1) Biostatistics, (2) Oncology and Pathology, Johns Hopkins University, USA (3) Cancer Center, Massachusetts General Hospital (MGH), USA (4) EURAC Research, Italy

Germline mutations of p16 exist in a subset of familial melanoma cases and cases with multiple melanomas. A tool to estimate p16 carrier probability based on clinical information is useful to direct people into cancer risk counseling and ongoing genetic studies, and potentially away from inappropriate p16 testing. We developed MELAPRO, a Mendelian carrier probability model, to provide individualized risk of carrying a P16 mutation, fully utilizing biological information about family pedigree, and adjusting for multiple primary melanomas. Mendelian models use prior estimates of prevalence and penetrance of deleterious mutations and apply Bayesian methods to derive carrier probabilities. We used: 1) high-risk family based penetrance estimates by GenoMEL (JNCI 94: 894), and 2) population-based penetrance estimates by GEMM (JNCI 97: 1507). We validated our model on 140 familial cutaneous melanoma patients consecutively enrolled at the MGH Pigmented Lesion Center between April 2001 and September 2004. MELAPRO exhibited good discrimination with areas under the ROC curve of 0.87 (95% CI: 0.73, 0.96) and 0.83 (95% CI: 0.66, 0.95) with the GenoMEL and GEMM datasets, respectively. With penetrance estimates by GenoMEL, MELAPRO also shows no significant difference in the observed and the predicted number of carriers. However, with the GEMM dataset, MELAPRO over-estimates the number of carriers.

187

Linkage Analysis of Affected Sib Pairs Allowing for Parent-of-Origin Effects: Multi-Locus Trait Models

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

Parent-of-origin effects, also known as genomic imprinting, differentiate a higher level of expression of genes inherited from one of the two parental chromosomes. Some genes that affect development and behavior in mammals are known to be imprinted. The statistical methods for testing linkage while allowing for parent-of-origin effects generally have greater power for imprinted loci than the usual statistical methods that ignore the parent-of-origin effects. We previously proposed model-free methods that account for genomic imprinting and linkage in the framework of affected sib pairs. In order to

investigate genetically complex traits in the presence of parent-of-origin effects, multi-locus models of inheritance need to be specified. Here, we present extensions of our previous methods to multi-locus models of inheritance. We propose two types of multi-locus model for incorporation of parent-of-origin effects into linkage analysis. The first extension, a multiplicative model, allows for investigation of epistasis (interaction) among loci. The second extension, an additive model, is characterized by no inter-locus interaction.

188

A Joint Model of Segregation, Linkage, and Association for Family Data

Xing and R.C. Elston
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH

Linkage approach is a useful tool for the initial exploration of complex traits; however, its ability to localize the loci potentially segregating for susceptibility is limited. Association approach directly tests the correlation between one trait and marker alleles; however, allelic association may be due to pleiotropy, linkage disequilibrium, meiotic drive, selection, or population stratification. With the ever increasing number of completed linkage scans and the availability of single nucleotide polymorphism screening sets, it is important that analysis methods be developed that can jointly model linkage and association for family data. The joint modeling of association and linkage allows for much more precise localization of regions housing disease genes. We propose a mixture generalized linear model to jointly modeling segregation, linkage and association for family data. The segregation is parameterized by a mixture model. By employing the linkage multipoint algorithm, this approach can take all markers across a chromosome into consideration. The allelic association is simply modeled by regress of the trait phenotype on the marker genotype. The generalized linear model handles both continuous and categorical trait data by different link function. The simplest version of this model is for monogenic traits; however, simulation studies show that it is also powerful for complex traits; Moreover, it can also be extended to multivariate models for multifactorial traits.

189

Effects of Population Structure on Haplotype Construction

H. Xu(1), S. Shete(2), V. George(1)
(1) Dept. of Biostatistics, Medical College of Georgia, Augusta, GA (2) Dept. of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX

Haplotype-based analysis is becoming increasingly popular in genetic epidemiological studies, especially in the studies of complex diseases. Haplotypes for human can be determined experimentally, which are however, costly and time-consuming. Statistical construction of haplotype from multi-locus genotype data has been shown to be an effective approach. Several such methods have been proposed, notably those based on Expectation-Maximization algorithm, Bayesian methods and Monte-Carlo Markov Chain sampling. Haplotype structure is jointly affected by demography and recombination. We examine the effects of population structure on several commonly used haplotype construction methods through computer simulation based on coalescent theory. Our results indicate that population structure could affect the performance of these methods. This could have important implications in haplotype analysis using admixed populations or samples from metropolitan populations where the extent of population structure is expected to be high.

190

Multiple affected relative pair linkage analysis: GEE and quasi-likelihood ratio statistic

W. Xu(1,4), S.B. Bull(2,4), C.M.T Greenwood(3,4)
(1) Dept. of Biostat, Princess Margaret Hospital, Canada, (2) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Canada, (3) Program in Genetics and Genomic Biology, The Hospital for Sick Children Research Institute, Canada, (4) Dept. of Public Health Sciences, University of Toronto, Canada

In large pedigrees with multiple affected members, affected-relative-pair (ARP) linkage analyses are often based on all possible pairs of affected members that can be formed. Although the allele IBD status of any two sib pairs from the same family has been shown to be pairwise independent, multiple pairs are not jointly independent. We investigated methods for ARP linkage analysis that take the dependence into account. A novel approach was developed applying generalized estimating equation (GEE) to multiple ARP data, and the relative risk parameters were estimated by solving the estimating equation. Simulation studies confirmed that the GEE estimates are robust under various working correlation assumptions. Quasi-likelihood ratio statistics were applied as test statistic for linkage, based on the estimates from the GEE linkage model, and compared to an analysis using a log-likelihood ratio statistic from an allele-sharing-based linkage model (Olson 1999) with significance thresholds set empirically. Simulation studies demonstrated that using all ARPs as if they were independent in the likelihood ratio linkage model controlled by empirical significance threshold is preferred for linkage test on multiple ARP data.

191

Generalized LD index of SNPs and the method to plot their pertinent components

R. Yamada, K. Hirosawa, V. Renault, M.Yamaguchi,
S. Chida, F. Matsuda
Center for Genomic Medicine, Kyoto University

A set of N SNPs has 2^n subsets including itself and empty set. Although linkage disequilibrium (LD) can be defined for each subset except for the empty set, it is not realistic to visually display the LD information of all the subsets. We propose a method to calculate multiple marker LD index for any SNP subset from haplotype frequency and to

display data for $(n-1)^2$ subsets among 2^n. As a method to calculate multiple markers LD index, we give dummy value, 1 or -1, for each of two alleles of individual SNPs and defined value of individual haplotype in a subset as a product of dummy value of composing SNP alleles. Multiple marker LD index is defined based on deviation products of corresponding multiple SNP alleles. Among 2^n SNP subsets, we select $(n-1)^2$ representative subsets and plot the records in a triangle which were separated into $(n-1)^2$ cells. The representative subsets are consisted of n(n-1)/2 SNP pairs and n-(k-1) subsets with k consecutive SNPs (k=3,4,...,n). This plot contains approximately as twice as many information than the conventional 2D plot of pair-wise LD index ($(n-1)^2/(n\ (n-2)/2)=2(n-1)/n$).

192

Identification and Replication of Locus Interactions Using Family Based Liquid Association (FLA) Testing in Familial Combined Hyperlipidemia (FCHL)

Tun-Hsiang Yang(1), Rita M. Cantor(2), Tjerk deBruin(3), Aldons J. Lusis(2) Ker-Chau Li(1)
Departments of (1) Statistics and (2) Human Genetics UCLA, (3) GlaxoSmith Kline, North Carolina

Identification of gene-gene interactions has become feasible with the recent advent of whole genome association studies, although correction for multiple testing remains a consideration. To provide a more focused approach, we developed a statistical method to search genome-wide for locus interactions based on Haseman-Elston (HE) regression in nuclear families. FLA uses the squared trait differences (X), IBD sharing at locus 1 (Y) and IBD sharing at locus 2 (Z) to assess the effect of conditioning on Z on the HE analysis of X at Y. The test analyzes how the correlation between X and Y varies by conditioning on Z. We interpret that any locus Z that increase the linkage signal between X and multiple Y, thus mediating linkage at many loci, as interactive. In addition, we have shown that each Y supports conditional linkage of Z with X. To investigate the application of this method, we used it to detect locus-locus interactions for the quantitative trait Apolipoprotein B in a sample of 102 Dutch nuclear families within pedigrees ascertained for FCHL and genotyped at 377 multiallelic markers. Ten Z loci were identified using apermutation test to assess significance. In an independent sample of 91 Dutch nuclear FCHL families, genomewide analysis revealed that two of these loci at 16q24 and 17p13 were replicated. These analyses support the FLA approach for the identification of loci likely to harbor genes that interact, thus providing targets for dense SNP studies to identify the specific genes that interact.

193

Differences in risk factors for breast cancer molecular subtypes in a population-based study

X.Yang(1), M.E.Sherman(1), D.L.Rimm(2), J.Lissowska(3), L.A.Brinton(1), B.Peplonska(4), S.M.Hewitt(1), W.F.Anderson(1), N.Szeszenia-Dabrowska(4), A.Bardin-Mikolajczak(3), W.Zatonski(3), R.Cartun(5), D.Mandich(5), M.Garcia-Closas(1)
(1) DCEG,NCI,NIH,USA, (2) Yale Univ., USA, (3) Cancer Center and M. Sklodowska-Curie Institute of Oncology, Poland, (4) Nofer Institute of Occ. Med., Poland.(5) Hartford Hospital, USA

The aim of this study is to evaluate whether pathologic features and etiologic associations differ among molecular subtypes identified by gene expression analysis. We evaluated 804 women with invasive breast cancers and 2,502 controls participating in the Polish Breast Cancer Study. Immunohistochemical stains for ER-á, PR, HER2, HER1, CK5 were used to classify cases into five tumor subtypes: luminal A, luminal B, HER2-expresing, basal-like and unclassified. Compared to the predominant(69%) luminal A tumors, other tumor subtypes, in particular, HER2-expressing(8%) and basal-like(12%) tumors, were associated with unfavorable clinical features at diagnosis. Increasing body mass index significantly reduced the risk of luminal A tumors among pre-menopausal women(OR *95%CI*=0.71 *0.57, 0.88*), while it did not offer protection against basal-like tumors(P=0.003). On the other hand, increasing age at menarche provided stronger protection for basal-like tumors(OR=0.78 *0.68, 0.89*; P=0.0009 for basal-like vs. luminal A). Although family history increased risk for all subtypes, the magnitude of the relative risk was highest for basal-like tumors. Our data suggests that breast cancer molecular subtypes may vary in etiologic associations in addition to clinical behavior.

194

Incorporating gene-environment interaction in the screening of marginal genetic effects

Yu-Chun Yen(1), Peter Kraft(1,2)
Harvard School of Public Health, Departments of Epidemiology (1) and Biostatistics (2)

Often researchers have strong suspicions that a particular environmental exposure interacts with genetic factors to produce complex disease. It is currently unclear how this gene-environment interaction can best be used to detect genetic risk factors. To test for association between a genetic variant and dichotomous disease, one popular method is testing the genetic effect marginally without incorporating environmental information. To test for gene-environment interaction, one can use the standard 1-d.f. test for statistical gene-environment interaction based on logistic regression, or the case-only test, which assumes gene-environment independence. We discuss two flexible alternative methods to detect genetic effect within any level of an environmental exposure: a 2-d.f. likelihood ratio test that allows for the genotype odds ratios differ between exposed and unexposed, and two independent likelihood ratio tests for exposed and unexposed. We compare the sample size requirements of these five methods. The results indicate that the flexible 2-d.f. joint test provides good power for detecting a gene across a wide range of underlying models, The joint test generally has better power than the marginal test when the genetic effect is restricted to exposed subjects and much better power than the tests of gene-environment interaction when the genetic effect is not restricted to a particular

exposure level. Thus, we recommend that the flexible 2-d.f. joint test could be used when we have little information a priori about the true gene-environment interaction model.

195

Genome-Wide Association Analysis using Sequential Haplotype Scan

Zhaoxia Yu(1) and Daniel J. Schaid(2)
(1) Division of Biostatistics, Mayo Clinic, Rochester, MN, USA

We provide a sequential haplotype scan method to look for sets of adjacent markers that are jointly associated with disease status. For each locus, we add markers close to it in a sequential manner based on the Mantel-Haenszel statistic. A new marker will be added if it provides extra information for detecting association. A haplotype-based chi-square statistic for the combined markers and a summary statistic are proposed to evaluate the region-wide significance of association by a permutation test. We applied our sequential haplotype scan algorithm to both simulated data and experimental data for CYP2D6. The results indicate that the sequential scan procedure can identify adjacent markers whose haplotypes might have strong genetic effects or be in linkage disequilibrium with disease predisposing variant. Therefore, sequential haplotype scan can achieve higher power than single-locus method.

196

A Two-stage Multi-marker Test using the Same Date Set in Genome-wide Association Study based on Family Data

Zhaogong Zhang(1), Qiuying Sha(2), Shuanglin Zhang(3)
Department of mathematical sciences Michigan Technological University, Houghton, MI 49931

Complex diseases are pre-assumed to be the results of many genes and environmental factors, with each gene only has a small effect to the disease. Thus, multi-marker methods that can use the information of markers from different genes are appropriate. There already have several multi-marker methods proposed for case-control studies. In this report, we propose a multi-marker method to analyze family date in genome-wide association study. Using the same data set, we also propose a genomic screening test that is independent of the family-based multi-marker method. Thus, based on the same family data set, we first use the genomic screen test that uses the parental phenotypes to select SNPs, and then use the family-based multi-marker method to test association on selected SNPs. We use simulation studies to evaluate the performance of the two-stage approach. The results show that the proposed two-stage approach and the single marker test (the PDT) has correct type I error, and robust to population stratification. In almost all the cases we considered, the two-stage approach is much more powerful than the single marker PDT.

197

Integrated Association Studies Using Family and Case-control Subjects

J. Zhang(1), X. Zhu(2)
Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, Illinois 60153

Genome-wide association studies are widely planned in the study of genetic variants of complex diseases nowadays. The population based case-control study and family based Transmission Disequilibrium test (TDT) are the most commonly used ones. However, they both have limitations, such as case-control analysis is vulnerable to spurious association from population stratification and parental controls are usually difficult to collect. As data from both approaches of the same disease-marker association can be expected highly in the near future, an integrated analysis of all the information is desired. We present a nonparametric approach for this situation, where genotype data from parents, siblings, unrelated cases and controls are all available. We divide all the affected as one group against the group of all the unaffected, and compute their allele frequencies and variances, and give a test statistic. Advantages and disadvantages compared with other combining methods in the literature such as likelihood based approach and odds ratio approach are carefully studied through simulations.

198

An Ensemble Learning Approach for Identifying a Set of Interacting Loci with Complex Traits

S. Zhang(1,2),Z. Zhang (1,2), Q. Sha (1), M.Y. Wong(3)
(1) Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931 (2) Heilongjiang University, Harbin 150080, China (3) Department of Mathematics, Hong Kong University of Sciences and Technology, Hong Kong, China

Complex diseases are presumed to be the result of the interaction of several genes and environmental factors, with each gene only having a small effect on the disease. Thus, methods that can account for gene-gene interactions to search for a set of marker loci in different genes and to analyze these loci jointly are critical. In this article, an Ensemble Learning Approach for Set-association (ELAS) was proposed to detect a set of interacting loci that predicts the complex trait. In the ELAS, we first search "base-learners" and then combine the effects of the base learners. The ELAS can jointly analyze single-marker effects and two-order interaction effects for many markers including genome-wide association studies. Simulation studies demonstrated that the ELAS is more powerful than single-marker test in all the scenarios we considered. The ELAS also outperformed the other three existing multi-locus methods in almost all cases. In an application to a large-scale case-control study for Type 2 diabetes, the ELAS identified 11 SNPs that have a significant multi-locus effect (p-value=0.01), while none of the SNPs showed

significant marginal effect and none of the two-locus combinations showed significant two-locus interaction effect.

199

Admixture Mapping Identifies Association of the VNN1 Gene with Hypertension

Zhu X., Cooper R.S.
Loyola University Medical Center

Migration patterns in modern societies have created the opportunity to use population admixture as a strategy to identify susceptibility genes. To implement this strategy, we genotyped a highly informative marker panel of 2270 single nucleotide polymorphisms in a random population sample of African Americans (N=1743), European Americans (N=1000). We then examined the evidence for over-transmission of specific loci to cases from one of the two ancestral populations. Hypertension cases and controls were defined based on standard clinical criteria. Both case-only and case-control analyses were performed among African Americans. With the genome-wide markers we replicated the findings identified in our previous admixture mapping study on chromosome 6 and 211. For case-control analysis we then genotyped 51 missense SNPs in 36 genes spaced across an 18.3 Mb region. Further analyses demonstrated that the missense SNP rs2272996 (or N131S) in the VNN1 gene was significantly associated with hypertension in African Americans and the association was replicated in Mexican Americans, but with a non-significant opposite association in European Americans. This SNP also accounted for the evidence in the admixture analysis. This study is the first to identify genetic variants influencing a complex trait through admixture mapping, although the further confirmation is required by independent studies.